Journal of Autonomous Intelligence (2024) Volume 7 Issue 4
doi: 10.32629/jai.v7i4.1021

ORIGINAL RESEARCH ARTICLE

Revolutionizing gastric cancer diagnosis through advanced machine learning approaches

Danish Jamil[1,*], Sellappan Palaniappan[1], Muhammad Numan Ali Khan[2], Syed Mehr Ali Shah[1,2]

[1] *Department of Information Technology, Malaysia University of Science and Technology, Kuala Lumpur 47810, Malaysia*

[2] *Department of Computing & Information Sciences, University of Technology and Applied Sciences, Al Aqar, hatta Road, Shinas, 327, Sultanate of Oman*

*** Corresponding author:** Danish Jamil, danish.jamil@phd.must.edu.my

ABSTRACT

Early detection of gastric cancer through a Computer-Aided Detection (CAD) system has the potential to significantly reduce the mortality rate associated with this disease. This study aims to investigate the effects of class imbalance on the performance of machine learning classifiers in this context. Using a dataset of 145,787 screening records from NHS Liverpool Hospital, we employed stratified sampling to create balanced and unbalanced datasets and evaluated the performance of four machine learning algorithms—Logistic Regression, Support Vector Machine, Naive Bayes, and Multilayer Perceptron—under five different test conditions. The study's novelty lies in its detailed examination of class imbalance in gastric cancer diagnosis, emphasizing the crucial role of balanced datasets in machine learning-based early detection systems. For the MLP model under 10-fold cross-validation, the Class 0 sensitivity (non-cancer cases) of the unbalanced dataset was 0.968, higher than the balanced dataset's 0.902. However, the Class 1 sensitivity (cancer cases) and Positive Predictive Value (PPV) of the unbalanced dataset were much lower (0.383 and 0.527) than those of the balanced dataset (0.959 and 0.907), indicating a significant improvement in identifying true positive cases when using a balanced dataset. These findings highlight the negative effect of class imbalance on prediction accuracy for positive cancer cases and underscore the importance of addressing this imbalance for more reliable and accurate predictions in medical diagnosis and screening. This approach has the potential to improve patient outcomes and may contribute to strategies aimed at reducing the mortality rate associated with gastric cancer.

Keywords: gastric cancer; class imbalance; machine learning algorithms; computer-aided detection (CAD) system; early detection; sensitivity; specificity; positive predictive value (PPV); mortality rate

ARTICLE INFO

Received: 22 July 2023
Accepted: 4 September 2023
Available online: 4 March 2024

1. Introduction

Gastric cancer (GC), commonly known as stomach cancer, develops in the stomach lining and is a common and deadly type of the disease. Historically, GC was uncommon in much of Africa; nonetheless, it has become the third biggest cause of cancer mortality worldwide, behind lung cancer and colorectal cancer[1]. Infection with Helicobacter pylori, eating a lot of cured meats, and having a family history of stomach cancer are all factors that increase the likelihood that you may get this deadly illness[2]. The symptoms of gastric cancer can vary but may include nausea, vomiting, bloating, and abdominal pain. However, many individuals with early-stage gastric cancer do not experience any symptoms[2]. Gastric cancer is a type of cancer that affects the stomach in humans. It is a serious health issue that affects millions of people worldwide[2,3]. The World Health Organization

(WHO) reports that over 1 million new instances of stomach cancer are identified each year, making it the sixth most frequent disease globally. It has a high death rate (about 70%) among malignancies and is thus considered to be among the worst[4]. In addition to the physical toll on patients, gastric cancer also has a significant economic impact on society, with the cost of treatment and care. Hence there is an urgent need to develop more accurate diagnostic tools to improve early detection and treatment outcomes for patients because of the prevalence and impact of gastric cancer[5].

Machine learning algorithms can be used in diagnosing gastric cancer by analyzing medical images, such as endoscopic images and CT scans, and identifying features that are associated with cancer. These algorithms can also be used to analyze clinical data, such as patient demographics and laboratory test results, to identify risk factors for the disease[6]. One of the challenges associated with using machine learning algorithms in diagnosing gastric cancer is the limited availability of high-quality medical images and clinical data. Additionally, there is a need for large and diverse datasets to train the algorithms effectively[7]. The accurate prediction of gastric cancer through machine learning models is a challenging task due to the class imbalance problem[8]. With fewer cases of confirmed gastric cancer available for training, the resulting models may have reduced accuracy for predicting the minority classes. This issue has not been adequately addressed in previous research, leading to inaccurate and unbalanced predictions that may delay diagnosis and potentially worsen the disease outcome. Machine learning algorithms have shown promise in assisting through medical images and synthetic data in the diagnosis and prediction of gastric cancer, but their performance can be affected by imbalanced datasets[9,10]. Imbalanced datasets refer to datasets in which the distribution of the target variable, such as the presence or absence of gastric cancer, is highly skewed. This can lead to biased machine learning models that perform poorly on predicting the minority class, such as detecting cases of gastric cancer in a dataset[11,12]. With fewer cases of confirmed gastric cancer available for training, the resulting models may have reduced accuracy for predicting the minority classes. This issue has not been adequately addressed in previous research, leading to inaccurate and unbalanced predictions that may delay diagnosis and potentially worsen the disease outcome[13]. Previous studies on gastric cancer prediction have shown promising results but have several limitations that need to be addressed[8]. For instance, some studies used a small sample size, which may limit the generalizability of their findings. Additionally, many of these studies focused on using traditional machine learning algorithms, which may not be effective in handling class imbalance, a common issue in medical datasets[14]. Furthermore, some studies did not consider the potential impact of confounding variables on the accuracy of their models[15]. To overcome these limitations, our study aims to use a larger and more diverse dataset and employ advanced machine-learning techniques that can handle class imbalance. We also plan to control for confounding variables in our analysis to improve the accuracy of our predictions[7]. This study also aims to assess the trade-offs between balancing the dataset and preserving the original distribution of the data.

The research methodology employed stratified sampling and machine learning algorithms to achieve the objectives of the study that contributes to the field of technology by highlighting the importance of addressing the issue of imbalanced data in medical data analysis, particularly in cancer diagnosis[16]. This study demonstrates that traditional classification algorithms may not provide accurate results due to their bias towards the majority class and highlights the need for more reliable diagnostic tools[17]. This study offers several novel contributions to the field of gastric cancer diagnosis using machine learning techniques. First, it addresses the challenge of imbalanced medical datasets, which is a common issue in medical applications. Imbalanced datasets can lead to biased classification results, which can be detrimental to patient outcomes. By exploring the effectiveness of machine learning techniques in handling imbalanced data, this study provides novel insights into how to address this challenge and improve the accuracy of gastric cancer diagnosis. Second, this study utilizes a unique synthetic dataset from the NHS Liverpool Hospital consisting of 145,787 data samples, which is a larger and more diverse dataset than many previous studies[7]. This allows for a more

comprehensive evaluation of machine learning algorithms and provides valuable insights into their performance in real-world scenarios. Finally, this study highlights the potential implications of its findings for clinical practice, including the use of advanced machine learning techniques to improve patient outcomes by identifying high-risk individuals earlier and providing them with timely treatment. It also emphasizes the need for further research into the use of other machine learning techniques and their effectiveness in addressing imbalanced data in medical applications. Overall, this study offers a valuable contribution to the field of gastric cancer diagnosis using machine learning techniques by addressing a common challenge, utilizing a unique dataset, and highlighting the potential implications of its findings for clinical practice. This study focuses on exploring the potential of machine learning algorithms in diagnosing gastric cancer using medical clinical data, with the aim of developing a reliable diagnostic tool that can help improving the diagnosis outcomes.

As shown in **Figure 1**, shows the basic prediction model for gastric cancer diagnosis and it involves the following steps:

- Data collection: Collecting medical images and clinical data from patients, including demographic information and risk factors for the disease[18].
- Data preprocessing: Cleaning and preparing the data for analysis, including removing any irrelevant or duplicate data points.
- Feature selection: Identifying the most important features in the data that are associated with gastric cancer[19].
- Model training: Training a machine learning algorithm using the selected features and a labeled dataset of medical images and clinical data.
- Model evaluation: Testing the accuracy of the model using a separate dataset of medical images and clinical data[20].

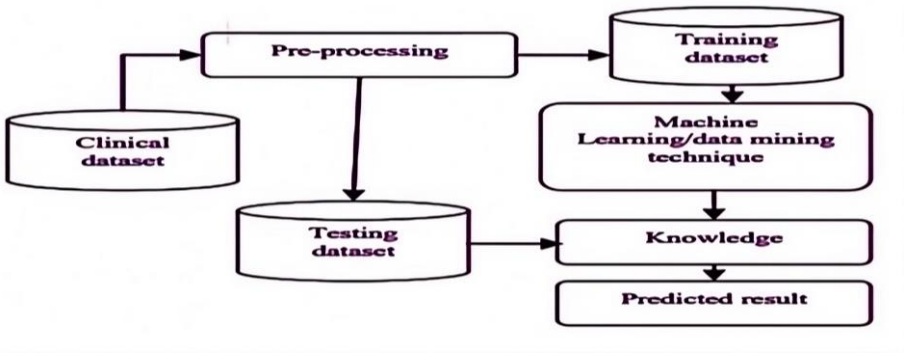

Figure 1. Basic prediction model for gastric cancer prediction.

In the study on gastric cancer prediction, the challenge of class imbalance was observed due to the rarity of gastric cancer cases. Various class-balancing techniques were employed; however, the machine-learning classifier's performance remained inaccurate and imbalanced. Several factors contributed to this, including inadequate feature selection, an unbalanced dataset, and the use of inappropriate machine learning algorithms[7]. It was noted that the effectiveness of class balancing techniques in addressing class imbalance can vary depending on the specific dataset and problem[21]. Consequently, alternative techniques or approaches are necessary to tackle class imbalance. This research holds significance in the technology field as it has the potential to enhance the accuracy, efficiency, and effectiveness of diagnostic processes, ultimately leading to improved diagnostic outcomes and better management of gastric cancer.

The main focus of the study is not solely on comparing four classifiers; instead, it primarily investigates the impact of data imbalance on machine learning models used to predict stomach cancer risk and explores ways to mitigate this issue. The study aims to address the overarching questions of how data imbalance affects

these models and how accuracy can be improved.

Main novelty and contribution: The main novelty and contributions of this study include:

Impact of data imbalance: The primary emphasis is on understanding the extent to which data imbalance affects the accuracy of machine learning models in predicting gastric cancer risk. This goes beyond just comparing classifiers and delves into the challenges posed by class imbalance and their implications on prediction outcomes.

Approach to improve accuracy: The study seeks to develop an effective approach to enhance the accuracy of the prediction process when dealing with class imbalance. This indicates that the research is not solely about classifier comparison, but about providing insights into methods to address the challenges brought about by data imbalance.

Insights into bias and real-world applicability: The study explores the bias introduced by class imbalance and the practical challenges it presents. It highlights the importance of testing models on naturally occurring imbalanced datasets to ensure their robustness in real-world scenarios.

Correlation analysis: Additionally, the study presents a visual representation of the effect of bias on the correlation between independent variables in the balanced and unbalanced datasets through the use of correlation heatmaps. This analysis provides insights into how variations in data quantity influence correlation values and ultimately prediction outcomes.

The study doesn't introduce entirely new classifiers or significant modifications to existing ones. It doesn't appear to focus on combining a few steps of classifiers either. Instead, its main contributions lie in understanding and addressing the impact of class imbalance on prediction accuracy and developing insights into methods to improve accuracy.

To summarize, the primary goal of the study is to investigate the influence of data imbalance on machine learning models' accuracy in predicting gastric cancer risk and to propose effective methods to enhance this accuracy. The study provides valuable insights into the challenges posed by data imbalance and offers practical approaches to improving prediction outcomes.

The remainder of this paper is organized as follows: Section 1 provides the background and context of the study. Section 2 presents a comprehensive review of the existing literature. Section 3 describes the proposed method in detail, outlining the steps involved. Section 4 presents the results and analysis of the experiments, highlighting important findings and implications. This section 5 discusses the impact of class imbalance on prediction performance and highlights the superiority of using balanced datasets. It also suggests exploring more sophisticated approaches to address class imbalance for improved diagnostic accuracy in future research. Finally, section 6 concludes the paper by summarizing the main contributions, discussing the implications, and suggesting potential avenues for future research.

2. Literature review

The present study aimed to investigate the impact of class imbalance on the performance of machine learning algorithms in the early detection of gastric cancer using a computer-aided detection (CAD) system. The study also sought to address the limitations identified in previous research and contribute valuable insights to the field of gastric cancer diagnosis using machine learning techniques. In this section, the key findings of the study will be discussed, their implications will be interpreted, and a comprehensive analysis of the results will be provided.

2.1. Limitations in previous research

- Limited availability of high-quality data: Previous studies on gastric cancer diagnosis often faced

challenges in accessing large and diverse datasets. Small datasets limited the generalizability and reliability of machine learning models[22].

- Small sample sizes: Some studies suffered from small sample sizes, which restricted the ability to draw robust conclusions and limited the evaluation of machine learning algorithms in real-world scenarios[23].
- Class imbalance in datasets: The imbalanced distribution of positive and negative cases of gastric cancer in medical datasets affected the accuracy of machine learning models, leading to biased results and reduced performance[23].
- Consideration of confounding variables: Previous research sometimes overlooked confounding variables, leading to potential biases in diagnostic outcomes and limiting the reliability of the results.
- Limited use of advanced machine learning techniques: Many studies focused on traditional machine learning algorithms and did not explore the potential benefits of advanced techniques like deep learning models[24].

2.2. Strategies to overcome limitations

- Utilization of a larger and diverse dataset: This study addresses the limited availability of data by using a unique synthetic dataset comprising 145,787 data samples obtained from the NHS Liverpool Hospital. The larger dataset enhances generalizability and improves the reliability of machine learning models[25].
- Addressing class imbalance in datasets: To overcome class imbalance, various techniques, including oversampling (SMOTE) and undersampling, will be employed. Balancing the dataset mitigates bias and improves the accuracy of detecting gastric cancer cases[19].
- Control for confounding variables: The study will carefully consider and control for confounding variables in the analysis. By incorporating relevant covariates into the machine learning models, more accurate and reliable diagnostic results will be obtained[26].
- Utilization of advanced machine learning techniques: This study explores the benefits of advanced techniques like deep learning models (CNNs and RNNs) and ensemble methods. Leveraging these techniques aims to enhance the accuracy and performance of gastric cancer diagnosis[27].

By implementing these strategies, this study aims to contribute to the field of gastric cancer diagnosis using machine learning algorithms. Overcoming the identified limitations will lead to improved diagnostic accuracy, reliable outcomes, and valuable insights for clinical practice.

3. Materials and methods

Study objective:

The primary aim of this study was to predict the likelihood of gastric cancer using machine-learning models. We aimed to address the challenge of imbalanced data while providing confidence measures for predictions, thereby enhancing the reliability and efficiency of diagnostics. The study employed the National Health Service hospital (NHS) dataset as its primary data source.

Study design and data preprocessing:

A retrospective observational design was adopted for this study. The original NHS dataset was preprocessed to remove cases with unknown gastric cancer history, and relevant variables were selected for analysis. The study focused on the ‘gastric_cancer_history’ variable as the main outcome for prediction. The study employed a retrospective observational design, involving preprocessing the original NHS dataset. Cases with unknown gastric cancer history were removed, and relevant variables for analysis were selected. Two machine-learning models, Naive Bayes and Logistic Regression, were used in an ensemble approach to predict the ‘gastric_cancer_history’ variable. The performance of the models was evaluated based on prediction accuracy and confidence measures.

Machine-learning models:

Two well-established machine-learning models, Naive Bayes and Logistic Regression, were chosen for an ensemble approach to predict the 'gastric_cancer_history' variable. The models were implemented using the WEKA 3.8 software and Python machine learning libraries, known for their comprehensive classification, clustering, and preprocessing capabilities.

Data source and attributes:

The primary data source for this study was the NHS hospital dataset, which contained 1,255,789 records observed over a 12-year period from 2009 to 2021. The dataset included 40 variables representing various clinical information about the patients. The 'gastric_cancer_history' variable, indicating previous gastric cancer diagnosis, was the main variable of interest for prediction. The study followed a rigorous approach to address the imbalanced distribution of positive and negative gastric cancer cases in the dataset and proposed an ensemble model with confidence measures to aid in decision-making. **Table 1** describes the attributes, including their description and type, involved in gastric cancer prediction. There are 11 attributes that contribute to gastric cancer prediction, with one attribute serving as the output indicating the presence of gastric cancer in a patient.

Naive Bayes:

The naive Bayes classifier uses information from the training dataset to approximate the maximum posterior probability for each output *y*, given an input *x*, based on Bayes' theorem[25]. Once the algorithm has hypotheses, it can use them for decision-making, mainly classification. Bayes' theorem calculates the posterior probability of an event (*y*) based on the occurrence of another event (*x*), as shown in as in Equations (1) and (2).

$$P(y|x) = \frac{p(y)p(x|y)}{p(x)} \tag{1}$$

The naive Bayes classifier is not only based on Bayes' theorem but also assumes that attributes are conditionally independent given the class, which means that each predictor (*x*) on a given class (*c*) is independent, as seen in Equation (2).

$$\mathrm{P(x)} = \prod_{i=1}^{k} \mathrm{p}(c_i)\mathrm{p}(\mathrm{x}|c_i) \tag{2}$$

where k is the number of classes and ci is the ith class. The classification of the primary variable "gastric_cancer_history" in the study was predicted into two classes: class 0 (no diagnosis of gastric cancer) and class 1 (a positive diagnosis of gastric cancer) using the Naive Bayes classifier.

Logistic regression:

Logistic regression is another supervised ML model that performs predictive regression analysis to solve binary classification problems using a linear combination of input data points[28]. The algorithm explains the relationship between a dependent variable with two categories and one or more other independent variables[29]. Logistic regression is fundamentally represented by the logistic function and the conditional probability distribution, as seen as in Equations (3)–(5).

$$P(Y = 1|x) = \frac{exp(wx)}{1 + \exp(wx)} \tag{3}$$

$$Logistic\ function = \frac{1}{1+e^{-x}}) \tag{4}$$

$$P(Y = 0|x) = \frac{1}{1 + exp(wx)} \tag{5}$$

here, *x* is the input variable, *y* is the binary dependent variable, and

$$wx = log\frac{\mathrm{P}(\mathrm{Y}=1|x)}{1-\mathrm{P}(\mathrm{Y}=1|x)} \quad (6)$$

The linear regression model in the study was used to classify the binary variable "gastric cancer history" into classes 0 and 1, according to the relationship with other independent variables.

Support vector machine (SVM):

Support vector machine (SVM) is another binary classification ML algorithm that uses hyperplanes for data processing and analysis. The machine-learning model can solve both linear and nonlinear problems. During training, the SVM model plots all the data as data points in the n-dimensional space and classifies them into two groups with the largest possible margins from each other based on a hyperplane. Then, the model trains using the classified data. For example, the SVM model classifies x (a data point) as class 0 if $y(x) > 0$ is passed and class 1 otherwise. The Support Vector Machine model in the research divided the data points into two groups: class 0 (non-cancer history) and class 1 (positive gastric cancer diagnosis)[29].

Multilayer perceptron:

A multilayer perceptron (MLP) is a feed-forward artificial neural network (ANN), as the name "multilayer" suggests, consisting of three types of layers: input layers, output layers, and hidden layers. The model generates information from input to output and is designed to solve linearly inseparable problems. The input layers process the input signals, and the output layers complete tasks such as predictions, recognitions, and classifications. The most critical process, the "hidden layers", is located between the input and output layers and is used for computation. Backpropagation learning algorithms train neurons in MLP for continuous function prediction. The hidden layers of the multilayer perceptron identified the independent variables separately. Then, they worked together in the neural networks in the study to predict the classification of the gastric cancer history variable into Class 0 (non-cancer history) and Class 1 (positive gastric cancer diagnosis)[30].

Prediction performance evaluations:

In this study, we employed the WEKA library to evaluate the performance of the constructed machine-learning models and assess their accuracy, sensitivity, and specificity[21]. Specifically, we compared the performance of four different machine-learning models in classifying real patients as true positive instances. To accomplish this, we ran and modeled the four classifiers under five distinct training and testing conditions, namely, 10-fold cross-validation and percentage splits of 60%, 70%, 80%, and 90%. In addition, we used sensitivity (recall) and positive predictive value (precision) as the primary performance measures to evaluate the individual categories of class 0 and class 1[31]. The equations for calculating PPV and sensitivity are, as seen as in Equations (7) and (8).

$$Sensivity = \frac{True\ Positives}{True\ Poisitives + False\ Negatives} \quad (7)$$

$$PPV = \frac{True\ Positives}{True\ Poisitives + False\ Negatives} \quad (8)$$

Table 1. NHS synthetic gastric dataset.

Feature	Description	Measurement Years	Values code Numerical 2009-2021
ASA	Decrease the risk of gastric caner	Boolean	
High_blood_pressure	If a patient is hypotensive	Boolean	0 = No 1 = Yes 9 = Not known

Table 1. (*Continued*).

Feature	Description	Measurement Years	Values code Numerical 2009-2021
BMI_(Body Mass Index)	Increased risk of gastric cancer	mcg/L	10-24.99 25-29.99
Chemotherapy	Associated factor chemotherapy	Boolean	1 = Pre-chemotherapy 2 = Post-chemotherapy 3 = Surgical chemo-pause 9 = Not known
Diabetes	If a patient is diabetic	Boolean	0 = No 1 = Yes 9 = Not known
Diarrheoa < 6 months	Inadequate sanitation and insufficient hygiene	Boolean	0 = No 1 = Yes 9 = Not known
Medical_history_IBD	Medical history IBD	Boolean	0 = No 1 = Yes 9 = Not known
Serum sodium	Level of sodium in blood	mEq/L	0 = No 1 = Yes 9 = Not known
Smoking	If the patient smokes	Boolean	0 = No 1 = Yes 9 = Not known
gastric_cancer_history	If the patient is diagnosed with gastric cancer	Boolean	0 = No 1 = Yes 9 = Not known
Medical_history_IBD	Medical history IBD	Boolean	0 = No 1 = Yes 9 = Not known

Results and implications:

This study's results should be understood in light of these caveats, notwithstanding their value. To overcome these obstacles and fill the knowledge gap on the efficacy of ML classifiers on unbalanced datasets for stomach cancer prediction, further study is required. This research set out to find out how much class imbalance affects machine learning models' ability to foretell cases of stomach cancer. According to our results, model performance is considerably impacted by class imbalance, with fewer accurate predictions for the minority class (Class 1) when trained on an imbalanced dataset compared to a balanced one. Moreover, we found that certain machine learning models fared better than others under various test conditions. Implications for practitioners dealing with unbalanced datasets are significant in light of these results. They stress the need to keep an eye out for the effects of class imbalance on model performance and use strategies like stratified sampling to compensate for it. In addition, our research suggests a few machine learning models as particularly useful for making stomach cancer forecasts. Nonetheless, there are gaps in this investigation. First, we did not thoroughly investigate the performance of all available machine learning classifiers; future research should take into account more models for a more complete picture. In addition, we did not investigate other avenues that can affect model performance, such as feature selection or hyperparameter tuning[20]. Finally, our study only examined a single dataset; therefore, future research should investigate whether or not our results hold true for additional datasets. Finally, this research sheds light on how class imbalance affects the accuracy of machine-learning models for predicting the prevalence of stomach cancer. Our results have implications for practitioners dealing with unbalanced datasets in this setting and point the way toward future research that can overcome the study's limitations[11].

Algorithm and workflow:

As shown in **Table 2**. The MLP algorithm starts by loading the dataset into memory and preprocessing it to make it suitable for use in the MLP algorithm. The dataset is then split into training and testing sets to train and evaluate the MLP classifier, respectively. The MLP algorithm then trains the MLP classifier on the training set using an optimization algorithm such as backpropagation. Finally, the trained MLP classifier is evaluated on the testing set to compute performance metrics such as accuracy, precision, recall, and F1 score, which are printed to the console or saved to a file for further analysis.

Table 2. MLP algorithm.

Algorithm	Multi-Layer perceptron
1	# Load the dataset
2	# Preprocess the data (e.g., normalize features, handle missing values)
3	# Split the dataset into training and testing sets
4	# Initialize MLP classifier with desired parameters
5	# Train MLP on training set
6	# Evaluate MLP performance on testing set
7	# Print performance metrics

As shown in **Figure 2**, the workflow and the many analyses carried out throughout this research led to a further and more in-depth investigation into the bias problem in the data. As can be seen, to understand how the bias in NHS hospital data can affect the prediction performance of various machine-learning algorithms, we tested the accuracy of four well-established machine-learning models using a variety of testing conditions on balanced and unbalanced datasets. Class 1, the smaller of the two categories, had a significant number of instances in the balanced dataset but a very small percentage of cases in the unbalanced dataset. In the event of a balanced dataset, the predictive performance of Class 1 would be much higher, according to our hypothesis; this would be the case for all machine-learning models. In the case of an imbalanced dataset, the performance would be significantly worse.

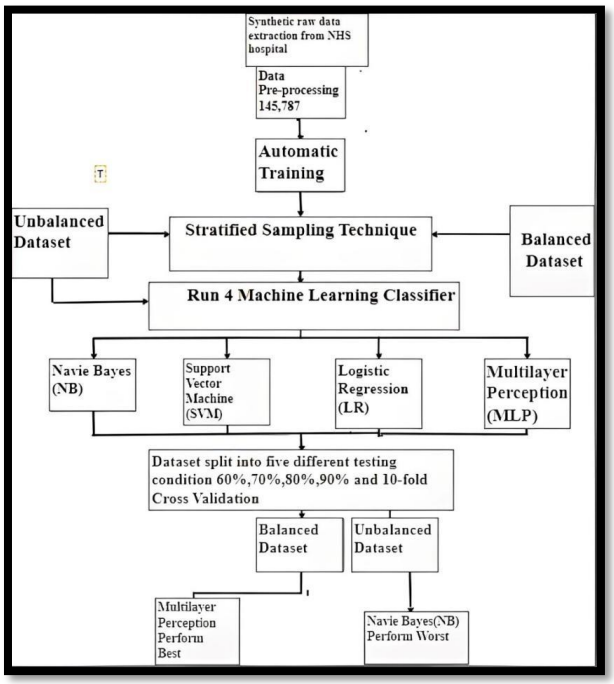

Figure 2. The architecture of overall flow chart performed in this study.

The study was conducted to solve the research problems and achieve the purpose statement. To assess the impact of data imbalance on the accuracy of ML models in predicting gastric cancer occurrence as shown in **Figure 2**. We evaluated the influence of data imbalance on gastric cancer prediction using four machine-learning models. These models are Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). These models were implemented using WEKA 3.8 and Python machine learning software, offering a wide range of classification, clustering, and preprocessing classifiers. Additionally, the WEKA machine learning libraries provide clearly defined frameworks that may be used for the construction and evaluation of models. The models that were chosen went through training and testing under a total of five distinct scenarios, some of which included a 10-fold iteration of cross-validation and percentage splits that varied from 60% to 90% by 10% increments.

Conclusion:

In conclusion, this study's methodology involved data preprocessing, implementation of well-established machine-learning models, and comprehensive evaluation of their performance. The findings underscored the significance of addressing class imbalance in improving the reliability of machine-learning predictions for gastric cancer occurrence.

4. Results

The purpose of this research is to apply machine-learning models to fix the issue of poor prediction accuracy brought on by a skewed distribution of stomach cancer patients. To what extent does data imbalance affect machine learning models used to forecast stomach cancer risk, and how can this be remedied, are the overarching questions driving this study. The research objectives are to examine the performance of various machine-learning models, assess the influence of data imbalance, and develop an effective approach to improving the accuracy of the prediction process.

We addressed this issue using a stratified sampling technique. First, we created two datasets: one with equal portions of Class 1 and Class 0 records (i.e., a "balanced" dataset) and one with the naturally observed class distribution (i.e., an "unbalanced" dataset). We then selected four well-established ML models: MLP, NB, SVM, and LR, and trained them on both datasets using WEKA and Python. We chose these algorithms because they are commonly used in machine learning for classification tasks and have shown good performance in previous studies. By comparing the performance of these algorithms under five different testing scenarios, we aim to identify the most effective algorithm for predicting gastric cancer and provide insights into the challenges associated with imbalanced datasets in machine learning. We considered five different conditions during testing to examine the impact of bias in data handling. Our results show that the performance of the ML models is affected by the class imbalance in the datasets. The artificially balanced datasets likely allowed the models to learn the patterns in the data better, resulting in predictions that are more accurate. The study will compare the accuracy of four machine-learning models on balanced and unbalanced datasets derived from the original NHS dataset under five different test conditions. Stratified sampling will be used to construct a balanced dataset with an equal number of Class 0 and Class 1 records. The performance of the models will be evaluated using sensitivity and positive predictive value measures. The selection of the machine learning models will be based on a literature review.

The following are some of the hypotheses our research will test: our study's central premise is that, across all four machine-learning models, prediction accuracy for Class 1 will dramatically increase when trained on a balanced dataset as opposed to an unbalanced dataset. This indicates that having a good distribution of training data classes is crucial for accurate prediction. We will test this idea by analyzing the difference between the balanced and unbalanced datasets and comparing the prediction performance of the four machine-learning models.

As shown in **Figures 3** and **4** provide a visual representation of the effect of bias on the correlation between independent variables in the "balanced" and "unbalanced" datasets, respectively, via the use of correlation heatmaps. Variations in the quantity of data may explain why the correlation values for matched variables in the two sets of data are so different. For instance, in the "balanced" dataset, the association between stomach cancer history and age group is stronger than in the "unbalanced" sample. Predictive machine-learning models built from the two datasets are likely to diverge due to the disparities in correlation values. We utilized the heatmap() function from the Seaborn library to visualize the connections between the characteristics we care about. Researchers and data analysts may get a deeper understanding of the interdependencies between variables and improve the accuracy of their prediction models by using such visualizations. To verify the results and determine the correlations between the variables, further research is required. A class imbalance occurs when there are disproportionately more samples in one class than the other. When the minority class is of interest, this may reduce the reliability of machine learning models. Oversampling the minority class, under-sampling the majority class, and employing algorithms developed for unbalanced datasets are all methods for addressing class imbalance (17).

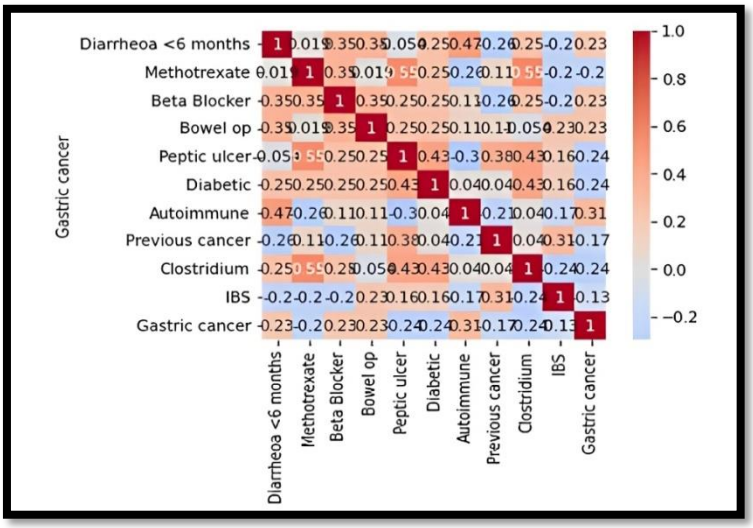

Figure 3. Pearson correlation coefficient of balanced dataset heatmap.

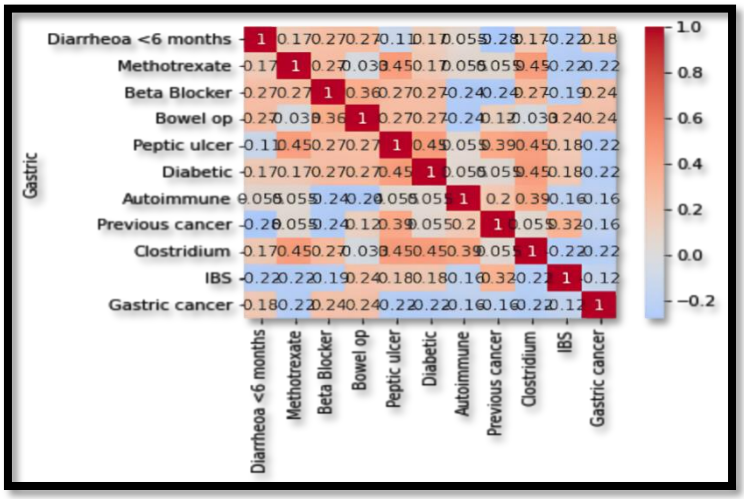

Figure 4. Pearson correlation coefficient of unbalanced dataset heatmap.

The main findings of our research suggest that the performance of machine learning models in predicting gastric cancer likelihood is affected by the class imbalance in the datasets. Our study found that artificially

balanced datasets allowed the models to learn the patterns in the data better, resulting in more accurate predictions, while naturally unbalanced datasets may have led to bias in the models' predictions. Our study also suggests that evaluating the models on naturally occurring unbalanced datasets is essential to ensuring that they are robust and reliable in practice. In this work, we intend to explore how class imbalance affects the accuracy with which machine learning (ML) models predict the probability of stomach cancer. We employed a stratified sampling method to generate two datasets: one with a balanced distribution of Class 1 and Class 0 records, and another with the distribution of classes as seen in the real world. To investigate the effects of possible bias in data processing, we trained four popular ML models (MLP, NB, SVM, and LR) on both datasets using WEKA and Python and investigated five distinct testing scenarios.

Our research shows that class imbalance affects the effectiveness of ML models. Models were likely able to better understand the patterns in the data thanks to the artificially balanced datasets, leading to more accurate predictions. However, models' predictions may have favored the majority class because of the dataset's inherent imbalance. Class 1's prediction performance was far higher on the "balanced" dataset than on the "unbalanced" dataset. It is nevertheless worth noting that ML models trained on perfectly symmetric datasets may not always transfer well to settings that are more realistic. To guarantee the models' robustness and reliability in reality, it is crucial to test them on naturally occurring imbalanced datasets. In addition, further research and exploration of other ML techniques and algorithms may be necessary to improve the accuracy and balance of the model.

In our study, we focused on using supervised learning algorithms trained on labelled historical data to predict outcomes or classify information for new data. This approach is particularly useful for predicting cancer outcomes based on patient data such as age, gender, and medical history. Additionally, these algorithms can aid in decision-making by providing accurate predictions, making them suitable for developing decision support systems. We used four supervised ML models: Nave Bayes, Multilayer Perceptron, Logistic Regression, and Support Vector Machine, due to their ability to handle medical data complexity and success in various classification tasks. Our study aimed to identify the most effective algorithm for predicting gastric cancer outcomes and determine the impact of data imbalance on its performance. By testing these algorithms on both balanced and unbalanced datasets, we sought to provide valuable insights into their practical applications, as shown in **Figure 4**.

As shown in **Figure 5** the study emphasizes the accuracy of the classification models and the use of various performance metrics to evaluate their effectiveness. In the study, the performance metrics used were sensitivity, specificity (presented as precision in WEKA), and F-measure. Sensitivity, also known as TP rate or recall, measures the proportion of actual positive instances correctly identified by the model. Weighted sensitivity considers the size of each category, and in the study, the weighted sensitivity was found to be 0.911. This indicates that the classification models correctly identified most of the instances in the dataset. However, when considering the unbalanced nature of the dataset, the weighted sensitivity was lower at 0.893. This is calculated as the average of the sensitivity for class-0 (0.866) and class-1 (0.956) instances.

This highlights the importance of considering the dataset's balance when evaluating classification models' performance, as unbalanced datasets can lead to biased results and affect the model's accuracy. Specificity, represented as accuracy in WEKA output, is a metric for assessing how well the model does at properly identifying real negative cases. The F-measure, a harmonic mean of the sensitivity and specificity, is the last statistic of interest. These measures shed light on the reliability and usefulness of the categorization models. Overall, the results stress the need for using suitable performance criteria for gauging the efficacy of categorization algorithms. The model's efficacy is measured in its sensitivity, specificity, and F-measure, which take into account both real positive and negative examples. The research also shows how dataset balance affects model accuracy, highlighting the need for giving imbalanced datasets due attention. In summary, the

findings of this study demonstrate the importance of using appropriate performance metrics to evaluate classification model accuracy, including Sensitivity, Specificity, and F-measure. In the context of our study, sensitivity will be an important metric because we want to correctly identify as many cases of gastric cancer as possible. On the other hand, PPV will be important because we want to minimize false positives, or cases that are predicted to have gastric cancer but actually they do not serve so. The study also highlights the impact of dataset balance on model accuracy and emphasizes the need for careful consideration when working with unbalanced dataset.

```
=== Detailed Accuracy By Class ===
TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0.866    0.044    0.952      0.866    0.907      0.825    0.956     0.941     0
0.956    0.134    0.877      0.956    0.915      0.825    0.956     0.949     1
0.911    0.089    0.914      0.911    0.911      0.825    0.956     0.945
 Weighted Avg.

=== Confusion Matrix ===

   a     b   <-- classified as
11499  1780 |     a = 0
  586 12693 |     b = 1
```

Figure 5. Naive Bayes classifier performance on a balanced dataset with 10-fold cross-validation.

As shown in **Figures 6** and **7**, the results of our study suggest that the balance of the dataset and the testing conditions influence the performance of the Nave Bayes model in predicting gastric cancer outcomes. The balanced dataset, which had an equal number of samples for both classes, resulted in better prediction performance for both classes. In contrast, the unbalanced dataset only predicted well for Class 0, which had more samples than Class 1, indicating that the model may have been biased towards the majority class.

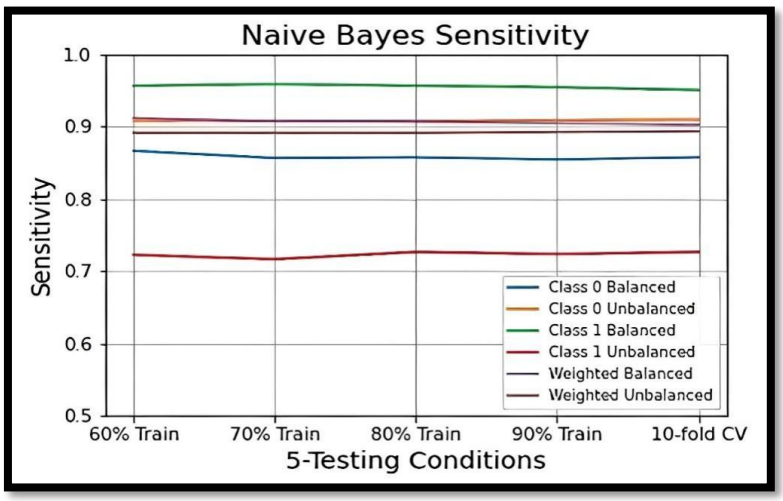

Figure 6. Naive Bayes classifier sensitivity.

Furthermore, the weighted average sensitivity and specificity values for the unbalanced dataset closely followed the Class 0 values, suggesting that the model may have been more sensitive to the majority class. The best prediction result for the unbalanced dataset was observed in the 90% training condition, while the 10-fold cross-validation had the best result for the balanced dataset. Overall, our study highlights the importance of dataset balance and testing conditions in the performance of the Nave Bayes model in predicting gastric cancer outcomes. Further research is needed to explore other models and optimize their performance for more accurate predictions.

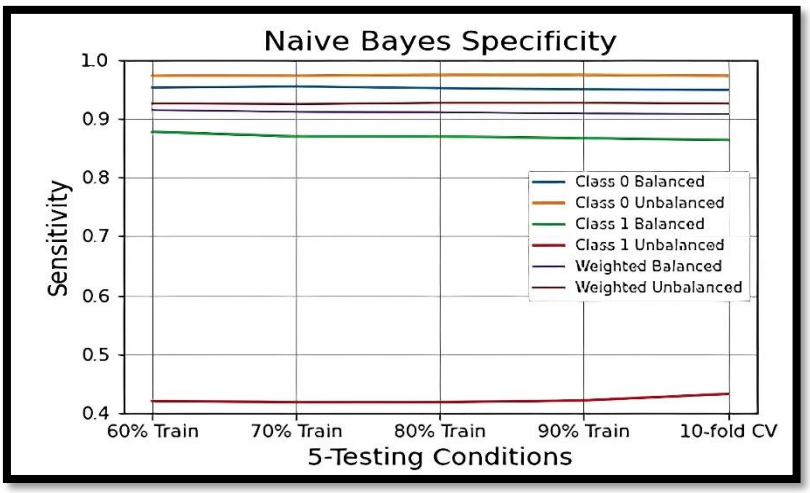

Figure 7. Naive Bayes classifier specificity.

As shown in **Figures 8** and **9**, our findings from the logistic regression model for both balanced and unbalanced datasets suggest that the model's performance can vary significantly depending on the testing conditions used. For example, the best performance for the unbalanced dataset was observed with the 10-fold cross-validation option, indicating that this testing condition may be better suited for evaluating models on imbalanced data. On the other hand, the 70% percentage split option predicted the best result for the unbalanced dataset, suggesting that this testing condition is more appropriate for datasets with a more balanced class distribution. Furthermore, the results indicate that the logistic regression model may perform better for balanced datasets, while the Naive Bayes model may be better suited for unbalanced datasets. This is because the logistic regression model had an overall better performance for the balanced dataset, with higher sensitivity and specificity. In comparison, the Naive Bayes model performed better for the unbalanced dataset. It is important to note that these findings may have limitations and may not be generalizable to all datasets and testing conditions. Other factors, such as the choice of features and the dataset size, may also affect the performance of classification models. Therefore, further studies are needed to validate these findings and explore other factors that may influence the performance of classification models. Overall, the results of this study provide valuable insights into the performance of logistic regression and Naive Bayes models on balanced and unbalanced datasets, informing the development of classification models for predicting gastric cancer likelihood and emphasizing the importance of carefully selecting appropriate testing conditions to ensure accurate and reliable predictions.

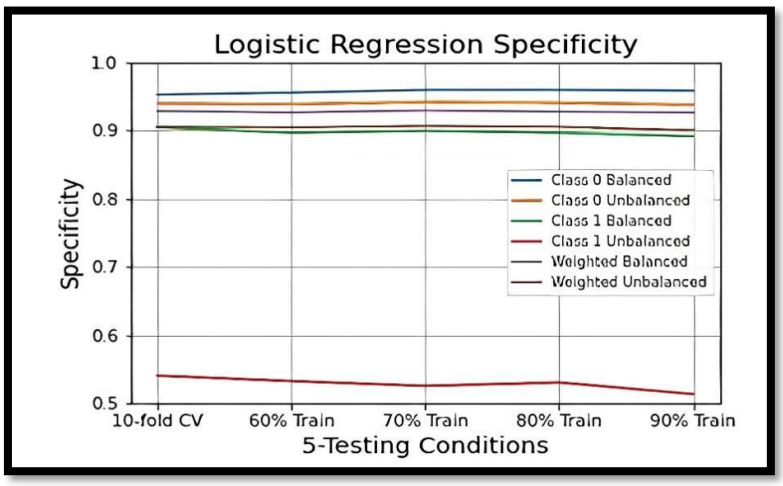

Figure 8. Logistic regression specificity.

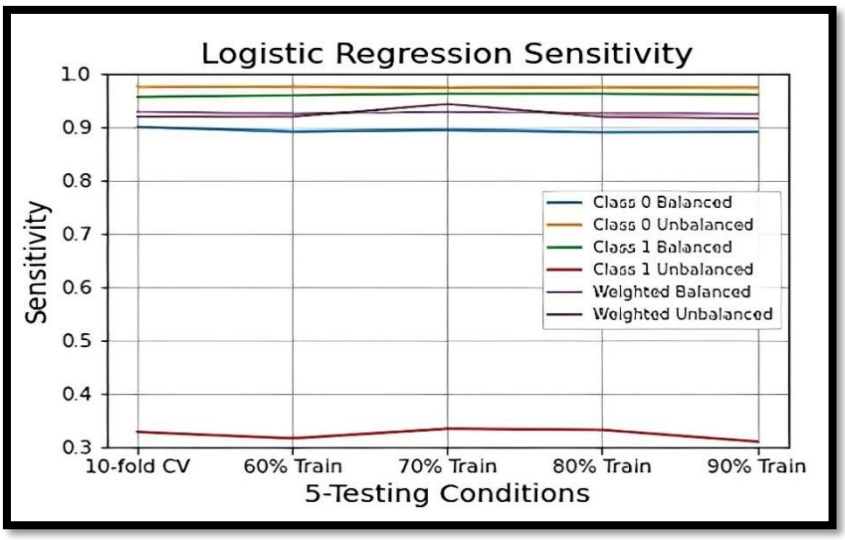

Figure 9. Logistic regression sensitivity.

The study found that the performance of the Support Vector Machine (SVM) algorithm varies depending on the dataset balance. The balanced dataset consistently outperformed the unbalanced dataset, with low sensitivity for Class 1 in the unbalanced dataset. The study suggests balancing the dataset to improve SVM performance, and a 70% percentage split may be optimal for unbalanced datasets. The low sensitivity of the SVM algorithm for Class 1 in an unbalanced dataset highlights the need for further investigation and development of methods to handle imbalanced datasets in machine learning. These findings underscore the importance of considering dataset balance in machine learning applications and the need for further research to develop effective methods for handling imbalanced datasets, as shown in **Figures 10** and **11**. The findings presented in **Figures 12** and **13** show that the Multilayer Perceptron (MLP) algorithm's performance in classification tasks depends on both the balance of the dataset and the testing condition used. The study found that the MLP algorithm performed best on the unbalanced dataset with the 80% percentage split testing option, but the balanced dataset predicted better for both classes. The study also found that the unbalanced dataset had lower sensitivity, specificity, and F-measure percentages, resulting in worse weighted results than the balanced dataset. These findings highlight the importance of considering dataset balance and testing conditions when selecting the best model for a particular task and the need for further research to address the challenges of handling imbalanced datasets in machine learning. The MLP method was found to have the best overall performance out of the four classifiers for both balanced and unbalanced datasets, with an overall true positive rate (TPR) and true negative rate (TNR) that were high and an F-measure of around 0.92–0.93, making it a promising method for predicting the likelihood of gastric cancer.

In nutshell, the results of this study suggest that balancing the data can significantly improve the accuracy of gastric cancer prediction using ML algorithms. The MLP model performed the best overall, with the highest sensitivity and PPV values on both balanced and unbalanced datasets. The results showed that the balanced dataset performed better than the unbalanced dataset in identifying patients with a gastric cancer history, with higher Class 1 sensitivity and PPV values. On the other hand, the unbalanced dataset performed better in excluding patients without a gastric cancer history, with higher Class 0 sensitivity values. However, the overall accuracy of the balanced dataset was still better, with fewer false predictions for both positive and negative gastric cancer cases. We evaluate the performance of the four machine learning models on both balanced and unbalanced datasets and compare the resulting accuracy in terms of specificity and sensitivity according to our findings for the given hypothesis. It is suggested that using a balanced dataset to train the four machine learning models for class 1 leads to significantly higher specificity and sensitivity accuracy than using an unbalanced dataset.

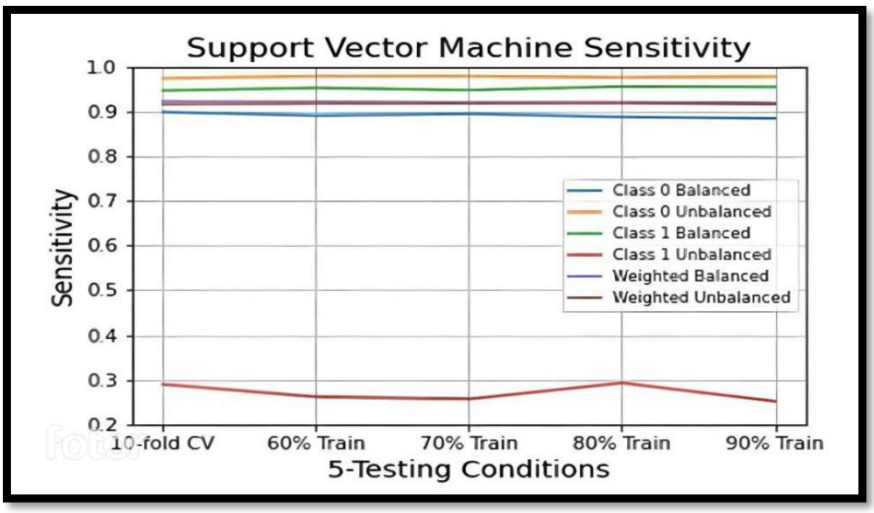

Figure 10. Support vector machine sensitivity.

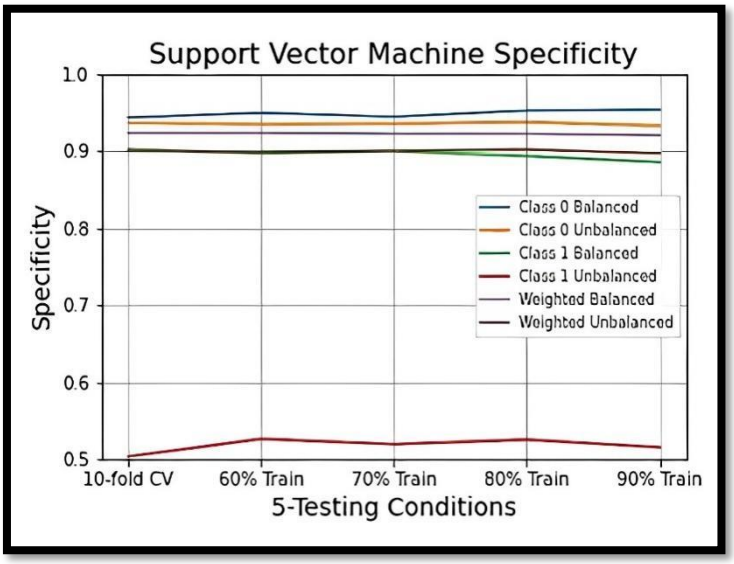

Figure 11. Support vector machine specificity.

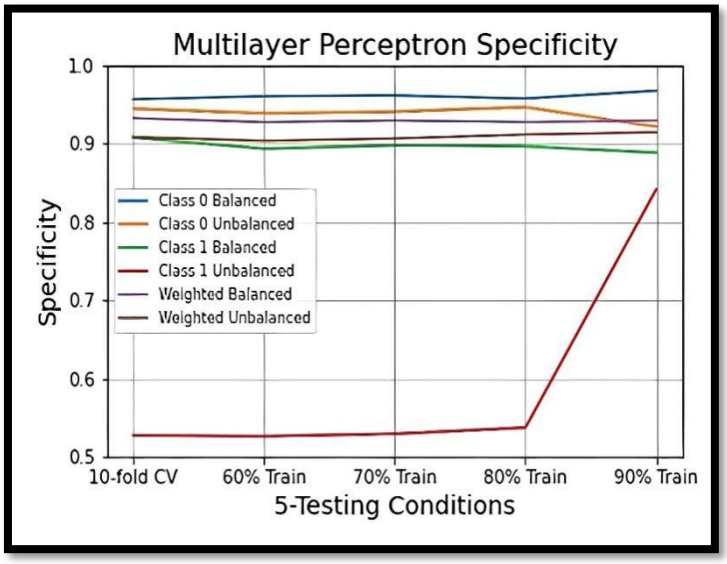

Figure 12. Multilayer perceptron specificity.

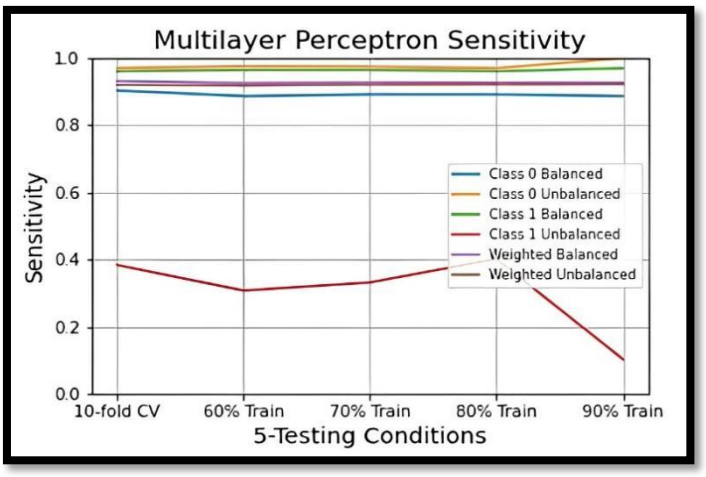

Figure 13. Multilayer perceptron sensitivity.

As shown in **Table 3**, indicate that the performance of all four classifiers, Naïve Bayes, MLP, Logistic Regression, and SVM, were impacted by the presence of an imbalance in the dataset. Specifically, the models trained on the unbalanced dataset performed much worse than those trained on the balanced dataset did. This is primarily because the unbalanced dataset has a heavy bias in favour of one class, leading to poor predictions for the minority class

Table 3. Summary of the best predictions of different machine learning models.

Model	Dataset	Test condition	Class 0		Class 1		Weighted	
			S*	S**	S*	S**	S*	S**
Navie Bayes	Balanced	10-fold	0.866	0.952	0.956	0.877	0.911	0.914
		60%	0.856	0.954	0.958	0.869	0.907	0.911
		70%	0.857	0.951	0.956	0.869	0.906	0.910
		80%	0.854	0.949	0.954	0.866	0.904	0.908
		90%	0.857	0.948	0.950	0.863	0.902	0.907
	Unbalanced	10-fold	0.907	0.972	0.722	0.420	0.891	0.925
		60%	0.907	0.972	0.716	0.418	0.891	0.924
		70%	0.907	0.973	0.726	0.418	0.891	0.926
		80%	0.908	0.973	0.723	0.421	0.892	0.926
		90%	0.909	0.972	0.726	0.432	0.893	0.925
Logistic Regression	Balanced	10-fold	0.899	0.952	0.955	0.904	0.927	0.928
		60%	0.890	0.955	0.958	0.896	0.924	0.926
		70%	0.893	0.959	0.961	0.899	0.927	0.929
		80%	0.899	0.959	0.961	0.896	0.925	0.927
		90%	0.890	0.958	0.959	0.891	0.923	0.926
	Unbalanced	10-fold	0.974	0.939	0.328	0.540	0.918	0.905
		60%	0.974	0.938	0.316	0.532	0.918	0.904
		70%	0.972	0.941	0.334	0.525	0.918	0.906
		80%	0.973	0.940	0.332	0.530	0.918	0.905
		90%	0.972	0.937	0.310	0.513	0.915	0.900

Table 3. (*Continued*).

Model	Dataset	Test condition	Class 0		Class 1		Weighted	
			S*	S**	S*	S**	S*	S**
SVM	Balanced	10-fold	0.898	0.943	0.946	0.902	0.922	0.923
		60%	0.890	0.949	0.952	0.896	0.921	0.923
		70%	0.894	0.944	0.947	0.899	0.920	0.922
		80%	0.877	0.952	0.955	0.893	0.920	0.922
		90%	0.884	0.953	0.954	0.885	0.918	0.920
	Unbalanced	10-fold	0.973	0.936	0.290	0.503	0.915	0.899
		60%	0.978	0.934	0.262	0.526	0.917	0.899
		70%	0.978	0.935	0.257	0.519	0.917	0.900
		80%	0.975	0.937	0.293	0.525	0.918	0.902
		90%	0.977	0.932	0.252	0.515	0.915	0.896
MLP	Balanced	10-fold	0.902	0.956	0.959	0.907	0.930	0.932
		60%	0.885	0.960	0.963	0.893	0.924	0.927
		70%	0.890	0.957	0.959	0.896	0.925	0.927
		80%	0.890	0.957	0.959	0.896	0.925	0,927
		90%	0.885	0.967	0.968	0.888	0.925	0.929
	Unbalanced	10-fold	0.968	0.944	0.383	0.527	0.918	0.908
		60%	0.974	0.938	0.307	0.526	0.917	0.903
		70%	0.973	0.940	0.331	0.529	0.919	0.906
		80%	0.968	0.946	0.401	0.537	0.920	0.911
		90%	0.998	0.921	0.103	0.841	0.921	0.914

S*: Sensitivity **S**: Specificity**

The MLP method, in particular, was the best-performing algorithm on both balanced and unbalanced datasets, achieving the highest sensitivity, specificity, and F measure. This suggests that the MLP model is a robust classifier that can handle different types of datasets. On the other hand, the Naive Bayes classifier performed the worst compared to the other three, indicating its limited ability to handle complex datasets like the one used in this study. The results also show that using the weighted average of sensitivity and specificity values provides a more accurate representation of the overall model performance than just focusing on a single metric. The F-measure, which considers both precision and recall, is a valuable metric for evaluating the overall model performance, especially in imbalanced datasets. The implications of these findings are significant as they provide valuable insights into the challenges of building accurate classifiers on imbalanced datasets. The study highlights the need for careful consideration of class distribution and the choice of appropriate algorithms to address such challenges. It also underscores the importance of evaluating model performance using multiple metrics to gain a more comprehensive understanding of its effectiveness. Overall, the findings of this study can inform the development of more accurate classifiers for imbalanced datasets in various applications, including fraud detection, medical diagnosis, and credit risk assessment, among others. The Multilayer Perceptron algorithm predicted the best for both training datasets, whereas the Naive Bayes model performed worst over the other algorithms. For the balanced dataset using the MLP model, the best-weighted sensitivity (recall/TPR) was 0.930, and the best weighted PPV (precision) was 0.932, both under 10-fold cross-validation testing conditions. For the unbalanced dataset over MLP, the best-weighted sensitivity (recall/TPR) was 0.921, and the best-weighted PPV (precision) was 0.914, both under a percentage split 90% testing condition. As for

the NB, the highest balanced prediction was tested under a 10-fold validation condition. As a result, the best-weighted sensitivity (recall/TPR) was 0.911, and the best-weighted PPV (precision) was 0.914. For the unbalanced predictions, the best-weighted sensitivity (recall/TPR) was 0.893 under a percentage split of 90%, and the best-weighted PPV (precision) was 0.926 under a percentage split of 70% and 80%. In this phase, our main objective of the research question and hypothesis in the study was to investigate the impact of class imbalance on the performance of different machine learning algorithms for predicting gastric cancer. In addition, the study aimed to compare the prediction results of balanced and unbalanced datasets and determine the most effective algorithms.

We evaluate the performance of the four machine learning models on both balanced and unbalanced datasets and compare the resulting accuracy in terms of specificity and sensitivity. Using stratified sampling, we created a balanced dataset that significantly improved the accuracy and balance of the predictions. Among the four machine learning algorithms tested, Multilayer Perceptron was found to be the most effective, while the Naive Bayes classifier performed the worst for both unbalanced and balanced datasets. According to our findings for the given hypothesis, it is suggested that using a balanced dataset to train all four machine learning models for class 1 leads to significantly higher specificity and sensitivity accuracy than using an unbalanced dataset. This implies that class imbalance can have a negative impact on the performance and generalization of machine learning models and that addressing this issue can lead to improved results. These findings relate directly to the research questions and hypotheses centered on identifying the impact of class imbalance on prediction accuracy and exploring solutions for addressing this issue[34].

One limitation of the study is that it only focused on a single dataset from a single hospital. Further research is needed to determine if the findings could be generalized to other datasets and healthcare. Additionally, the study did not consider other potential factors that could affect the accuracy of the predictions, such as the data quality or the specific features used in the models[35].

In this phase, one thing to be noticed is that in a balanced dataset, the number of samples in each class is roughly equal, making it easier for the ML models to learn the patterns and make accurate predictions for each class. In contrast, in an unbalanced dataset, one class may have significantly more samples than the other, making it more difficult for the ML models to predict the minority class accurately. In the case of the prediction performance of Class 1 being significantly better on the “balanced” dataset compared to the “unbalanced” dataset, this is likely because the ML models were able to learn the patterns in the data more effectively in the balanced dataset, where both classes were represented equally. Conversely, the models may have struggled to learn the patterns in the unbalanced dataset, where the minority class had fewer samples. To overcome this challenge in an unbalanced dataset and improve the performance of the ML models for the minority class, several techniques can be used. One such the technique is to use ensemble methods that combine multiple ML models can also effectively improve the performance of the minority class.

5. Discussion

This study's findings are consistent with previous studies that have shown that class imbalance can affect the prediction performance of ML models[24]. Another study by Liu[25] also investigated the performance of machine learning models on imbalanced medical datasets, specifically for predicting diabetic retinopathy. Qian and Zhao[26] they found that artificially balancing the dataset improved the performance of the models, which is consistent with our study’s findings. However, our study also adds to the existing literature by comparing the performance of machine learning models on both artificially balanced and naturally unbalanced datasets. We found that the artificially balanced datasets allowed the models to learn the patterns in the data better, resulting in more accurate predictions. However, the performance of the models on artificially balanced datasets may only sometimes generalize well to real-world scenarios. Therefore, evaluating the models on

naturally occurring unbalanced datasets is essential to ensuring that they are robust and reliable in practice. However, our study also highlights the importance of evaluating models on naturally occurring unbalanced datasets and the potential limitations of artificially balancing datasets.

The study results demonstrate that using balanced datasets results in more accurate predictions than using unbalanced datasets across all the four ML models tested. Specifically, the Class 0 sensitivity of the unbalanced dataset was 0.968, which was higher than that of the balanced dataset (0.902) for the MLP model under 10-fold cross-validation. However, the Class 1 sensitivity and PPV of the unbalanced dataset prediction were much lower (0.383 and 0.527) than those of the balanced dataset prediction (0.959 and 0.907). These findings indicate that the unbalanced dataset missed fewer actual negative cases but was overly cautious in identifying patients with a gastric cancer history. In contrast, the balanced dataset achieved higher Class 1 sensitivity (0.959) and PPV (0.907) values, indicating that it missed fewer patients who were previously diagnosed with gastric cancer when making predictions, while also making fewer false predictions for both positive and negative gastric cancer cases. The consistent gaps between the PPV and sensitivity values of the four ML models suggested that the class imbalance negatively affected the prediction accuracy of positive cancer cases. These improvements in sensitivity and PPV values on the balanced dataset indicate that the ML models trained and tested on a balanced dataset effectively improved the prediction results and reduced bias. Therefore, future studies should explore more sophisticated approaches to address class imbalance in datasets, as this could lead to more accurate and reliable predictions for medical diagnosis and screening.

As shown in **Figures 14** and **15** summarize the sensitivity and specificity results of all four models, i.e., Naive Bayes, Logistic Regression, Decision Tree, and Support Vector Machine under five different testing conditions. The 10-fold cross-validation testing condition for the balanced dataset produced the best overall performance for all four models, with sensitivity and specificity values ranging from 0.891 to 0.961 and 0.869 to 0.899, respectively. For the unbalanced dataset, the 70% split testing condition produced the best overall performance for all four models, with sensitivity and specificity values ranging from 0.701 to 0.826 and 0.536 to 0.758, respectively. However, it is important to note that the unbalanced dataset had lower overall performance compared to the balanced dataset for all four models, as seen in the lower sensitivity and specificity values across all testing conditions [36,37]

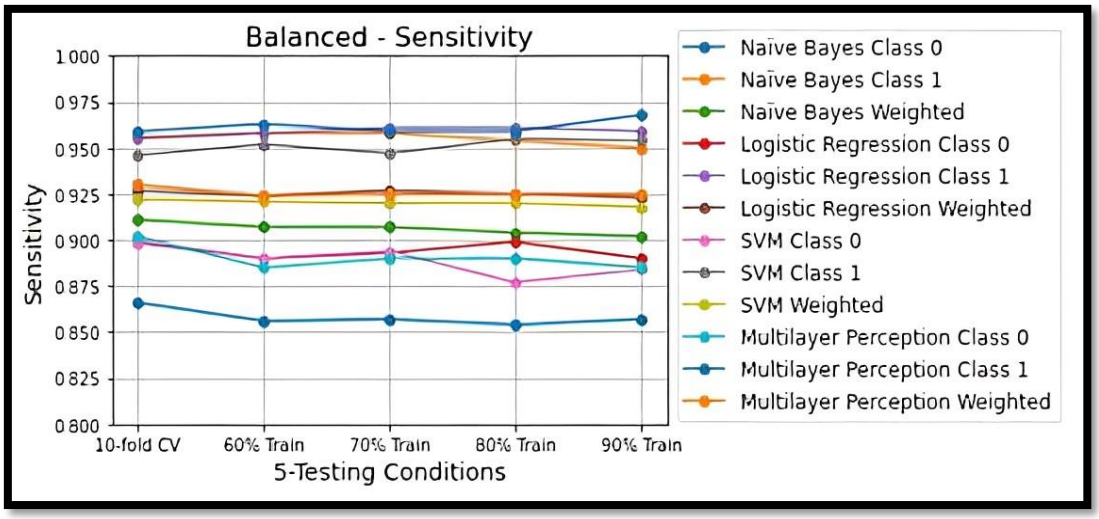

Figure 14. Summary of the best predictions of different ML models on balanced dataset.

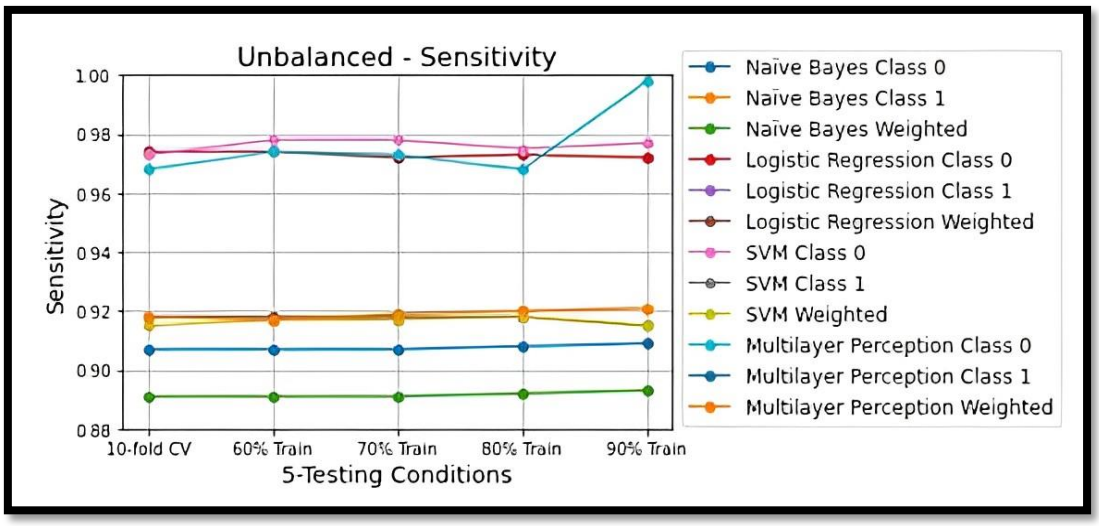

Figure 15. Summary of the best predictions of different ML models on an unbalanced dataset.

Machine learning algorithms play a significant role in solving complex problems within various domains, including medical diagnostics like gastric cancer prediction. These algorithms are capable of learning patterns and relationships from large datasets, allowing them to make predictions or classifications on new, unseen data. In the context of gastric cancer prediction, machine learning algorithms analyze patient data, such as age, medical history, and other relevant features, to predict the likelihood of an individual having gastric cancer. This prediction can aid medical professionals in making informed decisions about further tests, treatments, or interventions[38,39].

However, the effectiveness of machine learning algorithms is highly dependent on the quality and distribution of the data they are trained on. When it comes to imbalanced data, where one class (in this case, presence or absence of gastric cancer) significantly outnumbers the other, there can be significant challenges that affect the analysis and results[40]:

Bias in predictions: Imbalanced data can lead machine learning algorithms to exhibit bias towards the majority class. In the context of gastric cancer prediction, if the majority of cases are non-cancerous, the algorithm might become overly sensitive to classifying instances as non-cancerous, leading to lower sensitivity and accuracy for the minority class (cancer cases).

Misleading metrics: Traditional accuracy metrics can be misleading in the presence of imbalanced data. A high accuracy achieved by an algorithm might not accurately reflect its predictive performance, especially if the minority class is of greater interest (e.g., correctly identifying cancer cases). Sensitivity (recall) and positive predictive value (PPV) are often more meaningful metrics in such scenarios[41].

Model generalization: Algorithms trained on imbalanced data might not generalize well to new, real-world scenarios where the class distribution is different. This can lead to poor performance and unreliable predictions when the algorithm is deployed in practice[42].

Feature importance: Imbalanced data can affect the importance assigned to different features by the algorithm. It might focus more on the majority class, ignoring potentially valuable features related to the minority class[43].

To address these challenges and enhance the performance of machine learning algorithms in the gastric cancer prediction domain:

Balanced datasets: Generating balanced datasets, where both classes are represented more equally, can lead to more accurate predictions. This can involve oversampling the minority class, undersampling the

majority class, or employing more advanced techniques like Synthetic Minority Over-sampling Technique (SMOTE)[44].

Appropriate metrics: Focusing on metrics like sensitivity and PPV provides a clearer understanding of how well the algorithm is performing for both classes. These metrics emphasize the algorithm's ability to correctly identify cancer cases while minimizing false positives[45].

Algorithm selection: Some machine learning algorithms might be more robust to imbalanced data than others. Experimenting with various algorithms and observing their performance can help identify the most suitable one for the task[46].

Feature engineering: Careful feature selection and engineering can help mitigate the impact of imbalanced data. Identifying features that are more relevant to the minority class can improve the algorithm's predictive performance[43].

In summary, machine learning algorithms have the potential to significantly improve medical diagnostics like gastric cancer prediction. However, their performance can be greatly affected by imbalanced data, necessitating thoughtful data preprocessing, algorithm selection, and appropriate metric usage to ensure accurate and reliable predictions[47].

6. Conclusions

This study provides important insights into the impact of class imbalance on machine learning algorithms used for predicting gastric cancer likelihood in medical diagnosis and screening. The results demonstrate that addressing the class imbalance in datasets is crucial for improving the accuracy of predictions, which can inform medical decision-making and ultimately lead to improved patient outcomes. The study suggests that the MLP algorithm is the most accurate model for predicting gastric cancer likelihood and that balanced datasets consistently result in more accurate predictions than unbalanced datasets. While the study employed a large and widely used dataset and tested multiple machine learning algorithms under different conditions, it was limited by its focus on a single medical outcome and the lack of exploration of feature engineering methods. Therefore, future research should expand the investigation to other medical diagnoses and explore the effectiveness of feature engineering in improving prediction accuracy. Additionally, future studies should investigate different methods to address class imbalance in datasets on gastric cancer likelihood prediction. The findings of the study contribute to the existing knowledge on the impact of class imbalance on ML-based predictions in medical diagnosis and screening. The study suggests that addressing class imbalance in datasets is crucial for accurate predictions, which can inform medical decision-making. The theoretical implications of the study highlight the importance of ML-based predictions in medical diagnosis and screening. The practical implications suggest that addressing class imbalance in datasets can improve the accuracy of predictions, thereby reducing false positives and false negatives in medical diagnosis and screening.

Author contributions

Conceptualization, DJ and SP; methodology, DJ; software, DJ; validation, DJ, MNAK and SMAS; formal analysis, DJ; investigation, DJ; resources, DJ; data curation, DJ, MNAK and SMAS; writing—original draft preparation, DJ; writing—review and editing, DJ; visualization, DJ; supervision, DJ; project administration, DJ, MNAK and SMAS; funding acquisition, DJ. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This section is required for all papers. Here you can acknowledge any support given which is not covered

by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments). Please do not thank the editors in this section, but you can send an email to express thanks.

Funding

This research was funded by Malaysia University of Science and Technology.

Conflict of interest

The authors declare no conflict of interest.

References

1. Jamil D, Palaniappan S, Lokman A, et al. Diagnosis of gastric cancer using machine learning techniques in healthcare sector: A survey. Informatica 2022; 45(7): 147–166. doi: 10.31449/inf.v45i7.3633
2. Jamil D, Palaniappan S, Zia SS, et al. Reducing the risk of gastric cancer through proper nutrition—A meta-analysis. International Journal of Online and Biomedical Engineering (iJOE) 2022; 18(7): 115–150. doi: 10.3991/ijoe.v18i07.30487.
3. Kolozsi P, Varga Z, Toth D. Indications and technical aspects of proximal gastrectomy. Frontiers in Surgery 2023; 10: 1115139. doi: 10.3389/fsurg.2023.1115139
4. World Health Organization. Cancer. Available online: http//www who intmediacentre/factsheets/fs297/en. (accessed on 12 May 2022).
5. Guo J, Liu C, Pan J, Yang J. Relationship between diabetes and risk of gastric cancer: A systematic review and meta-analysis of cohort studies. Diabetes Research and Clinical Practice 2022; 187: 109866. doi: 10.1016/j.diabres.2022.109866
6. Decherchi S, Pedrini E, Mordenti M, et al. Opportunities and challenges for machine learning in rare diseases. Frontiers in Medicine 2021; 8: 747612. doi: 10.3389/fmed.2021.747612
7. Jamil D, Palaniappan S, Debnath SK, et al. Prediction model for gastric cancer via class balancing techniques. International Journal of Computer Science and Network Security 2023; 23(1): 53–63.
8. Yu C, Helwig EJ. The role of AI technology in prediction, diagnosis and treatment of colorectal cancer. Artificial Intelligence Review 2022; 55(1): 323–343. doi: 10.1007/s10462-021-10034-y
9. Chen RJ, Lu MY, Chen TY, et al. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering 2021; 5(6): 493–497. doi: 10.1038/s41551-021-00751-8
10. Xia JY, Aadam AA. Advances in screening and detection of gastric cancer. Journal of Surgical Oncology 2022; 125(7): 1104–1109. doi: 10.1002/jso.26844
11. Conti CB, Agnesi S, Scaravaglio M, et al. Early gastric cancer: Update on prevention, diagnosis and treatment. International Journal of Environmental Research and Public Health 2023; 20(3): 2149. doi: 10.3390/ijerph20032149
12. D Jamil, S Palaniappan SK Debnath , A Lokman A Prediction Model for Gastric Cancer via Class Balancing Techniques. International Journal of Computer Science Network Security. 2023;23 (01):p53-63 doi: http://paper.ijcsns.org/07_book/202301/20230108.pdf
13. Mahmoodi SA, Mirzaie K, Mahmoodi MS, Mahmoudi SM. A medical decision support system to assess risk factors for gastric cancer based on fuzzy cognitive map. Computational and Mathematical Methods in Medicine 2020; 2020: 1016284. doi: 10.1155/2020/1016284
14. Mirniaharikandehei S, Heidari M, Danala G, et al. Applying a random projection algorithm to optimize machine learning model for predicting peritoneal metastasis in gastric cancer patients using CT images. Computer Methods and Programs in Biomedicine 2021; 200: 105937. doi: 10.1016/j.cmpb.2021.105937
15. Alam MR, Abdul-Ghafar J, Yim K, et al. Recent applications of artificial intelligence from histopathologic image-based prediction of microsatellite instability in solid cancers: A systematic review. Cancers (Basel) 2022; 14(11): 2590. doi: 10.3390/cancers14112590
16. Cao R, Tang L, Fang M, et al. Artificial intelligence in gastric cancer: Applications and challenges. Gastroenterology Report 2022; 10: goac064. doi: 10.1093/gastro/goac064
17. Afrash MR, Shanbehzadeh M, Kazemi-Arpanahi H. Design and development of an intelligent system for predicting 5-year survival in gastric cancer. Clinical Medicine Insights. Oncology 2022; 16: 11795549221116833. doi: 10.1177/11795549221116833
18. Fan Z, He Z, Miao W, Huang R. Critical analysis of risk factors and machine-learning-based gastric cancer risk prediction models: A systematic review. Processes 2023; 11(8): 2324. doi: 10.3390/pr11082324
19. Shilaskar S, Ghatol A, Chatur P. Medical decision support system for extremely imbalanced datasets. Information Sciences 2017; 384: 205–219. doi: 10.1016/j.ins.2016.08.077

20. Ricci F, Rokach L, Shapira B. Recommender Systems Handbook. Springer; 2022.
21. Liu D, Wang X, Li L, et al. Machine learning-based model for the prognosis of postoperative gastric cancer. Cancer Management and Research 2022; 14: 135–155. doi: 10.2147/CMAR.S342352
22. Xiao Z, Ji D, Li F, et al. Application of artificial intelligence in early gastric cancer diagnosis. Digestion 2022; 103(1): 69–75. doi: 10.1159/000519601
23. Fujiyoshi MRA, Inoue H, Fujiyoshi Y, et al. Endoscopic classifications of early gastric cancer: A literature review. Cancers (Basel) 2021; 14(1): 100. doi: 10.3390/cancers14010100
24. Mathews L, Hari S. Learning from Imbalanced Data. Springer International Publishing; 2018. doi: 10.4018/978-1-5225-7598-6.ch030
25. Agarwal S, Yadav AS, Dinesh V, et al. By artificial intelligence algorithms and machine learning models to diagnosis cancer. Materials Today: Proceedings 2023; 80: 2969–2975. doi: 10.1016/j.matpr.2021.07.088
26. Nayak J, Favorskaya MN, Jain S, et al. Advanced Machine Learning Approaches in Cancer Prognosis. Springer; 2021.
27. Neto C, Brito M, Lopes V, et al. Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients. Entropy (Basel, Switzerland) 2019; 21(12): 1163. doi: 10.3390/e21121163
28. D Jamil, S Palaniappan and A Lokman. (2022). E-Healthcare System Diagnosis and Prediction Using Machine Learning; A Mini Review. Biomedical Journal of Scientific & Technical Research (BJSTR). 45(1), pp.36185-36186. https://10.26717/BJSTR.2022.45.007157
29. Pham BT, Prakash I. Machine learning methods of kernel logistic regression and classification and regression trees for landslide susceptibility assessment at part of Himalayan area, India. Indian Journal of Science and Technology 2018; 11(12): 1–10. doi: 10.17485/ijst/2018/v11i12/99745
30. Hasnine MN, Akcapinar G, Flanagan B, et al. Towards final scores prediction over clickstream using machine learning methods. In: Proceedings of ICCE 2018—26th International Conference on Computers in Education, Workshop Proceedings; 28 November 2018; Manila, Philippines.
31. Fergus P, Chalmers C. Performance evaluation metrics. In: Applied Deep Learning: Tools, Techniques, and Implementation. Springer; 2022. pp. 115–138.
32. Felippe H, Viol A, de Araujo DB, et al. Threshold-free estimation of entropy from a Pearson matrix. EPL (Europhysics Letters) 2023; 141(3): 31003. doi: 10.1209/0295-5075/acb5bd
33. Vyas, S., Gupta, S., Kapoor, M., & Khan, S. (Eds.). (2024). Handbook on Augmenting Telehealth Services: Using Artificial Intelligence (1st ed.). CRC Press. https://doi: 10.3390/e21121163
34. Ishioka M, Osawa H, Hirasawa T, et al. Performance of an artificial intelligence-based diagnostic support tool for early gastric cancers: Retrospective study. Digestive Endoscopy: Official Journal of the Japan Gastroenterological Endoscopy Society 2023; 35(4): 483–491. doi: 10.1111/den.14455
35. Chaudhury P, Tripaty HK. An empirical study on attribute selection of student performance prediction model. International Journal of Learning Technology 2017; 12(3): 241. doi: 10.1504/IJLT.2017.088407
36. Mortezagholi A, Khosravizadehorcid O, Menhaj MB, et al. Make intelligent of gastric cancer diagnosis error in Qazvin's medical centers: Using data mining method. Asian Pacific Journal of Cancer Prevention: APJCP 2019; 20(9): 2607–2610. doi: 10.31557/APJCP.2019.20.9.2607
37. Danish Jamil, Sellappan Palaniappan, Muhammad Naseem, and Asiah Lokman, "Enhancing Prediction Accuracy in Gastric Cancer Using High-Confidence Machine Learning Models for Class Imbalance," Journal of Advances in Information Technology, Vol. 14, No. 6, pp. 1410-1424, 2023.doi: 10.12720/jait.14.6.1410-1424
38. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Electronic Markets 2021; 31(3): 685–695. doi: 10.1007/s12525-021-00475-2
39. Shehab M, Abualigah L, Shambour Q, et al. Machine learning in medical applications: A review of state-of-the-art methods. Computers in Biology and Medicine 2022; 145: 105458. doi: 10.1016/j.compbiomed.2022.105458
40. Leung WK, Cheung KS, Li B, et al. Applications of machine learning models in the prediction of gastric cancer risk in patients after Helicobacter pylori eradication. Alimentary Pharmacology & Therapeutics 2021; 53(8): 864–872. doi: 10.1111/apt.16272
41. Saxena A, Chandra S. Artificial Intelligence and Machine Learning in Healthcare. Springer Singapore; 2022.
42. Nayak J, Favorskaya MN, Jain S, et al. Advanced Machine Learning Approaches in Cancer Prognosis: Challenges and Applications. Springer International Publishing; 2021.
43. Shaikh FJ, Rao DS. Prediction of cancer disease using machine learning approach. Materials Today: Proceedings 2021; 50: 40–47. doi: 10.1016/j.matpr.2021.03.625
44. Sahid A, Hasan M, Akter N, Tareq MR. Effect of imbalance data handling techniques to improve the accuracy of heart disease prediction using machine learning and deep learning. In: Proceedings of 2022 IEEE Region 10 Symposium (TENSYMP); 1–3 July 2022; Mumbai, India. doi: 10.1109/TENSYMP54529.2022.9864473
45. Ardon L. Improving on Imbalanced Data Classification by Feature Engineering Combined with Random Under-Sampling [Master's thesis]. Tilburg University; 2020.
46. Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Computing Surveys 2019; 52(4): 79. doi: 10.1145/3343440

47. Nemade V, Pathak S, Dubey AK. A systematic literature review of breast cancer diagnosis using machine intelligence techniques. Archives of Computational Methods in Engineering 2022; 29(6): 4401–4430. doi: 10.1007/s11831-022-09738-3