Enhancing Prediction Accuracy in Gastric Cancer Using High-Confidence Machine Learning Models for Class Imbalance

Danish Jamil [1,2,*], Sellappan Palaniappan [1], Muhammad Naseem [2], and Asiah Lokman [1]

[1] Department of Information Technology, Malaysia University of Science and Technology, Kuala Lumpur, Malaysia; Email: sell@must.edu.my (S.P.), asiah@must.edu.my (A.L.)
[2] Department of Software Engineering, Sir Syed University of Engineering and Technology, Karachi, Pakistan
*Correspondence: danish.jamil@phd.must.edu.my, djamil@ssuet.edu.pk (D.J.)

***Abstract*—Gastric Cancer (GC) diagnosis and prognosis present significant challenges in the clinical industry. To address the issue of low prediction accuracy resulting from imbalanced positive and negative GC cases, this study proposes a medical Decision Support System (DSS) based on supervised Machine Learning (ML) methods. Four ML models, including Naïve Bayes (NB), Logistic Regression (LR), and Multilayer Perceptron (MLP), were employed in this study. The impact of data imbalance on GC prediction was assessed through two procedures. Among the ML models, the MLP model demonstrated the best performance in weighted GC prediction, achieving a sensitivity of 0.930 and a Positive Predictive Value (PPV) of 0.932 for balanced predictions, and a sensitivity of 0.918 and a PPV of 0.908 for unbalanced predictions. The NB model showed promise in handling the data imbalance issue, achieving a sensitivity of 0.722 and a PPV of 0.420 on the unbalanced dataset. Additionally, a DSS was developed specifically for the NB and LR models to improve prediction accuracy. The proposed method significantly improved the sensitivity of optimistic GC case prediction, with the Naïve Bayes model achieving a sensitivity of 0.936 and the Logistic Regression model achieving a sensitivity of 0.8306. These improvements enhance the reliability and efficiency of GC diagnostics, offering valuable decision support in healthcare. This research provides insights into addressing class imbalance in GC likelihood prediction and has potential implications for clinical practice.**

***Keywords*—class imbalance, gastric cancer, decision support system, machine learning, prediction accuracy, naive bayes, logistic regression, medical diagnostics, positive predictive value**

I. Introduction

Gastric cancer is a significant global health concern, requiring accurate diagnosis and prognosis for effective treatment [1]. According to the World Health Organization, there were an estimated 1.09 million new cases and 768,793 deaths worldwide in 2022 (World Health Organization, 2022) [2]. In 2023, the American Cancer Society projects approximately 26,500 new cases of gastric cancer and around 11,130 related deaths in the United States [3]. Despite advancements in treatment options, accurate prediction of gastric cancer occurrence remains challenging. Machine Learning (ML) and Artificial Intelligence (AI) techniques have emerged as valuable tools in healthcare research, enabling the analysis of large datasets and the development of predictive models to improve gastric cancer prediction accuracy [3]. However, the prediction of gastric cancer likelihood is hindered by the imbalanced distribution of positive and negative cases in the available data [3–5]. The purpose of this study is to address the gaps and challenges in gastric cancer prediction using ML techniques. Specifically, this study aims to tackle the issue of class imbalance in training datasets, where the number of positive gastric cancer cases is considerably lower than the number of negative cases. This class imbalance often leads to reduced prediction accuracy for the minority class.

Several studies have explored various approaches to address the class imbalance in gastric cancer prediction, including data sampling techniques [6]. However, there is a need to further investigate the reliability and confidence of ML-based predictions from a medical decision-making perspective [7, 8]. Incorporating high-confidence predictions into the decision-making process can provide valuable support to healthcare professionals and enhance the diagnostic process [9]. The significance of this research lies in its potential to improve gastric cancer prediction and subsequent decision-making [10]. By addressing the class imbalance issue and providing high-confidence ML-based predictions, this study can contribute to more accurate diagnostics, improved treatment planning, and better patient outcomes. Additionally, this research aims to bridge the gaps in the current literature by examining the medical decision-making perspective and providing insights into the reliability of ML predictions. The main developments in the research topic of gastric cancer prediction involve the application of ML and AI techniques to analyze large datasets and predict outcomes. Previous studies have explored data sampling techniques to address class imbalance, but there is a lack of research

Manuscript received May 23, 2023; revised June 25, 2023; accepted July 7, 2023; published December 19, 2023.

focusing on the medical decision-making perspective and the reliability of ML-based predictions [11]. By investigating and addressing these gaps, this study aims to contribute to the existing body of knowledge and provide valuable insights for healthcare professionals. The findings of this research can have implications for improving gastric cancer prediction accuracy, enhancing treatment strategies, and ultimately reducing the mortality rate associated with this disease.

The two studies, referred to as "Our Study 1" and "Study 2: Prediction Model for Gastric Cancer via Class Balancing Techniques," focus on the prediction of gastric cancer using machine learning techniques while addressing the challenge of class imbalance in training datasets. In Our Study 1, a medical decision support system is developed based on supervised machine learning models, including Naïve Bayes, Logistic Regression, and Multilayer Perceptron. The study evaluates the impact of data imbalance on gastric cancer prediction and highlights the effectiveness of the MLP model for weighted predictions. Additionally, the Naïve Bayes model shows promise in handling data imbalance. The proposed decision support system offers valuable insights for improving gastric cancer prediction accuracy.

In Study 2, "Prediction Model for Gastric Cancer via Class Balancing Techniques," the emphasis is on addressing class imbalance through class-balancing techniques before applying supervised learning strategies. Various classifiers, including Naive Bayes, Bayesian Network, Random Forest, and Decision Tree (C4.5), are utilized. The study evaluates the performance of these classifiers after employing oversampling, undersampling, and a hybrid approach. Notably, the classifiers created using the hybrid balancing method demonstrate the best performance, particularly the Bayesian Network model, which exhibits superior overall performance in terms of accuracy metrics such as the false positive rate and area under the ROC curve.

The findings of these studies have practical implications for gastric cancer prediction and decision-making in healthcare. The MLP and Naïve Bayes models proposed in Study 1 can be considered for developing decision support systems in clinical settings. These models have shown promising performance in accurately predicting gastric cancer cases and handling data imbalance. On the other hand, Study 2, "Prediction Model for Gastric Cancer via Class Balancing Techniques" emphasizes the effectiveness of the hybrid balancing method and the superiority of the Bayesian Network model in achieving accurate predictions. These findings highlight the importance of addressing class imbalance and utilizing appropriate machine-learning models and techniques to improve prediction accuracy. However, it is important to consider the limitations and specific context of each study, such as the characteristics of the datasets used, the sample size, and any assumptions made during the modeling process. Future research should further explore these aspects and conduct comparative studies to validate and generalize the findings. By doing so, the accuracy and reliability of machine learning-based gastric cancer prediction can be enhanced, ultimately leading to improved outcomes in clinical practice.

This paper makes significant contributions in addressing the problem of low prediction accuracy caused by an imbalanced distribution of positive and negative gastric cancer cases. The key contributions of this study are as follows:

- Evaluation of ML Model Performance: The study systematically evaluates the performance of various ML models in handling the class imbalance issue. Comparing and analyzing the results, it provides insights into the effectiveness of different ML approaches for gastric cancer likelihood prediction.
- Assessment of Data Imbalance Impact: The research investigates the impact of data imbalance on the accuracy of gastric cancer prediction. By examining the challenges and limitations posed by imbalanced datasets, the study sheds light on the importance of addressing this issue for reliable predictions.
- Development of an Effective Approach: This paper proposes an effective approach to improve the accuracy of ML-based decision-making processes. By implementing specific techniques or algorithms to tackle the data imbalance problem, the study enhances the prediction accuracy and offers a practical solution for improving gastric cancer diagnostics.

These contributions advance the field of gastric cancer prediction by providing insights into ML model performance, highlighting the impact of data imbalance, and offering an effective approach to enhance prediction accuracy. The findings of this study have implications for improving clinical decision support systems and ultimately contribute to the advancement of gastric cancer diagnostics and prognosis.

This paper consists of three main sections. Section II provides a comprehensive literature review, highlighting relevant studies and existing research in the field. In Section III, the proposed Medical Decision Support System (DSS) model is presented, detailing its components and functionality. Section IV showcases the experimental results, providing a detailed analysis and comparisons of the different ML models employed. Finally, in Section V, the research is concluded, summarizing the key findings and discussing the implications of the study.

II. Literature Review

Nowadays, artificial intelligence and machine learning are widely used and considered effective ways to analyze big data, which can play important roles in predicting the occurrence of gastric cancer [12–14]. Machine Learning (ML) algorithms have the ability to learn from large volumes of past data, predict outcomes for future data, and classify data into different categories, thus aiding decision-making in situations involving large-scale data analysis [15, 16]. This section provides a comprehensive overview of studies that support the background and hypothesis of this research, highlighting the current trends and progress in the field. The following studies have been reviewed as shown in Table I.

TABLE I. LITERATURE REVIEW AND RESEARCH GAP SUMMARY

Study	Objective	Methods	Finding	Research Gap
Ming *et al.* [17]	Compare breast cancer prediction performance	Compared ML models and BCRAT	ML models showed improved prediction accuracy compared to BCRAT	Not mentioned in the passage
Stark *et al.* [18]	Predict breast cancer risk	Used ML models on personal health data	ML models (neural network, logistic regression, linear discriminant analysis) outperformed BCRAT	Not mentioned in the passage
Rajendran *et al.* [19]	Address class imbalance issue	Used oversampling, undersampling, hybrid methods	Hybrid balancing method (SMOTE + Spread Subsample) facilitated fair predictive models	Previous studies did not examine approaches to address class imbalance issue from a medical decision-support
Yin *et al.* [20]	Address class imbalance issue	Compared ML classifiers on balanced and unbalanced datasets	Balanced dataset performed better for all ML models compared to unbalanced dataset	Previous studies did not examine approaches to address class imbalance issue from a medical decision-support perspective
Current Study	Address class imbalance and provide confidence	Used ensemble approach with Naive Bayes and Logistic Regression	Proposed approach aimed to provide high-confidence ML predictions for more reliable and efficient diagnostics	Research gap is addressed by proposing an ensemble approach that provides confidence measures for ML predictions and filters cases requiring further review by domain experts

Liu *et al.* [17] investigated the utility of ML algorithms, including logistic regression, random forest, and gradient boosting, in predicting lymph node metastasis in gastric cancer. The findings revealed that ML models could accurately predict lymph node involvement, aiding in treatment decision-making and patient management.

Zhou *et al*. [18] developed an ML-based prognostic model for gastric cancer patients using gene expression data. The model integrated multiple ML algorithms, such as random forest and gradient boosting, to predict patient survival outcomes. The results demonstrated the potential of ML in personalized prognostic prediction for gastric cancer patients.

Previous studies have examined the performance of various ML models in predicting the likelihood of gastric cancer while considering the issue of class imbalance in the data. Class imbalance refers to the considerably low percentage of positive gastric cancer cases in historical diagnosis data, which can lead to lower prediction accuracy for smaller categories with fewer cases compared to larger categories [6]. This study addressed this issue by utilizing various data sampling techniques to create more balanced data distributions. They applied ML methods, including Naïve Bayes, Bayesian Network, Random Forest, and Decision Tree, on the National Health Service (NHS) dataset after employing oversampling, under-sampling, and a hybrid balancing method. The hybrid method, which combined Synthetic Minority Over-sampling TEchnique (SMOTE) and Spread Subsample, resulted in a uniform distribution of cases across two classes, leading to the development of fair predictive models.

Most previous studies have shown better prediction performance of ML models on artificially created balanced datasets compared to naturally observed unbalanced datasets [18, 19]. However, applying these findings to real-world decision-making for predicting the likelihood of gastric cancer can be challenging due to the scarcity of positive gastric cancer cases in population health datasets. Additionally, from a medical decision-making perspective, the reliability and confidence of the ML predictions are crucial factors. Healthcare professionals typically incorporate ML predictions with manual analysis of diagnostic data to ensure accurate decisions [18, 19]. Therefore, a machine learning-based decision support system that provides confidence or reliability measures for predictions can effectively support healthcare professionals in incorporating ML predictions into the decision-making process. For example, a stronger prediction (with a high confidence measure) by the ML model may indicate that the diagnostic data strongly supports the predicted outcome, while a weaker prediction (with a low confidence measure) may necessitate more detailed manual analysis.

The present methodology in this study has limitations that highlight areas for improvement and alternative approaches that could be considered.

One limitation of the study is the utilization of a limited set of ML models in the methodology. Exploring alternative ML models, such as Support Vector Machines, Random Forests, or Deep Learning models, could provide different perspectives and potentially enhance the performance of gastric cancer prediction in this study. Another limitation is the approach used to address class imbalance in the dataset. Exploring advanced sampling techniques, such as SMOTER (SMOTE with Edited Nearest Neighbor) or adaptive sampling methods, could be beneficial in this study [20]. These techniques can be explored to generate balanced datasets and mitigate the impact of class imbalance on the predictive models. The absence of external validation is a significant limitation in this study. Incorporating validation on independent datasets from diverse populations would strengthen the methodology employed here. External validation provides an assessment of the generalizability and robustness of the proposed approach, enhancing its credibility and applicability in this study.

The methodology could benefit from the integration of ensemble methods in this study. Leveraging the strengths of multiple ML models through ensemble approaches,

such as Random Forests, Gradient Boosting, or Stacking, could improve prediction accuracy in this study. Ensemble methods combine the predictions of multiple models to generate more reliable gastric cancer predictions. Considering the advancements in deep learning, the integration of deep learning techniques should be discussed in this study. Models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown promising results in medical prediction tasks. Exploring the benefits of these models, specifically designed for image or sequential data analysis, could contribute to the advancement of gastric cancer prediction in this study [21]. The interpretability of predictive models in healthcare is essential, and it can be further addressed in this study. Incorporating explainable AI techniques, such as feature importance analysis, SHapley Additive exPlanations (SHAP) values, or attention mechanisms, would enhance the interpretability of predictions in this study. This would improve the trustworthiness and acceptance of the methodology by healthcare professionals involved in this study. By addressing these limitations and considering alternative approaches, the proposed methodology in this study can be strengthened and contribute to the field of gastric cancer prediction [22].

In this ensemble approach, the stacking technique is employed to combine the predictions of two ML models, Naïve Bayes and Logistic Regression, to provide high-confidence and low-confidence predictions. During the training phase of this proposed method, the available data is divided into a training set and a validation set [23, 24]. The Naïve Bayes model and the Logistic Regression model are trained using the training set independently. Predictions are then made on the validation set using both models. In the meta-model training phase, the predictions from the base models (Naïve Bayes and Logistic Regression) are combined with the original features as input. A meta-model, such as another Logistic Regression model or a suitable alternative, is trained using the combined predictions as input and the target variable labels as the output. The meta-model is trained using the validation set. In the prediction phase, the predictions from the Naïve Bayes and Logistic Regression models are combined. These combined predictions are used as input for the trained meta-model to generate the final prediction. To achieve high-confidence and low-confidence predictions, a threshold is set on the predicted probabilities. Predictions with probabilities above the threshold are considered high-confidence predictions, while those below the threshold are categorized as low-confidence predictions. The threshold can be adjusted based on the desired level of confidence. By utilizing this stacking ensemble technique in the proposed method, the ensemble model can leverage the strengths of both the Naïve Bayes and Logistic Regression models and potentially enhance the overall performance by combining their predictions. The inclusion of a meta-model further improves the prediction accuracy and enables the generation of high-confidence and low-confidence predictions based on the chosen threshold.

However, previous studies have not extensively explored approaches to address class imbalance issues in gastric cancer likelihood prediction from a medical decision-support perspective [6]. This study aims to bridge the research gap in gastric cancer likelihood prediction. The approach taken in this study distinguishes itself by addressing the broader issue of gastric cancer likelihood prediction, compared to previous studies that focused on specific aspects such as prognostic prediction or lymph node metastasis. An ensemble approach is proposed, using two ML models, Naïve Bayes and Logistic Regression, to provide high-confidence and low-confidence predictions. The approach considers the class imbalance issue, incorporates confidence measures for ML predictions, and emphasizes the collaboration between ML models and domain experts. By doing so, the study aims to develop a more comprehensive and reliable decision support system for gastric cancer prediction, with the objective of enhancing patient management and treatment decision-making.

III. Materials and Methods

This study aimed to predict the likelihood of gastric cancer using machine-learning models. The research objective was to develop an approach that addresses the challenge of imbalanced data and provides confidence measures for predictions, allowing for more reliable and efficient diagnostics. The study utilized the National Health Service (NHS) hospital dataset as the primary data source.

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. 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 II 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.

A. 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 Eqs. (1) and (2).

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

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 Eq. (2).

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

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.

B. 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. The algorithm explains the relationship between a dependent variable with two categories and one or more other independent variables [26]. Logistic regression is fundamentally represented by the logistic function and the conditional probability distribution, as seen as in Eqs. (3)–(5).

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

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

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

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.

C. 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 [27]. 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).

D. 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 [28]. 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).

E. 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 [29, 30]. 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. The equations for calculating Positive Predictive Value (PPV) and sensitivity are, as seen as in Eqs. (7) and (8).

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

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

The variable "gastric_cancer_history" refers to the previous gastric cancer diagnosis history and was considered as the main variable to be predicted using various ML models. It served as the dependent or class variable. From the original dataset, as shown in Table II, "gastric_cancer_history" had three possible recorded values: 0, 1, and 6. A value of 0 indicated that the patient had no history of gastric cancer (referred to as "Class 1"), a value of 1 indicated that the patient had previously been diagnosed with gastric cancer (referred to as "Class 0"), and a value of 6 meant that the status was unknown. Since the study aimed to predict whether a patient had a positive or negative gastric cancer diagnosis, the cases with "gastric_cancer_history" = 6 were removed before further processing. Please see the attached supplementary file.

TABLE II. NATIONAL HEALTH SERVICE (NHS) SYNTHETIC GASTRIC CANCER DATASET

Feature	Description	Measurement Years	Values code Numerial 2009−2021
Acetylsalicylic Acid (ASA)	Decrease the risk of gastric caner	Boolean	
High_blood_ pressure	If a patient is hypotensive	Boolean	0 = No 1 = Yes 6 = Not known
Body Mass Index (BMI)	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 6 = Not known
Diabetes	If a patient is diabetic	Boolean	0 = No 1 = Yes 6 = Not known
Diarrheoa<6 months	Inadequate sanitation and insufficient hygiene	Boolean	0 = No 1 = Yes 6 = Not known
Medical_ history_IBD	Medical history IBD	Boolean	0 = No 1 = Yes 6 = Not known
Serum sodium	Level of sodium in blood	mEq/L	0 = No 1 = Yes 6 = Not known
Smoking	If the patient smokes	Boolean	0 = No 1 = Yes 6 = Not known
gastric_cancer_history	If the patient diagnosis with gastric cancer	Boolean	0 = No 1 = Yes 6 = Not known
Medical_history_IBD	Medical history IBD	Boolean	0 = No 1 = Yes 6 = Not known

The final dataset contained 145,789 cases after removing the cases with "gastric_cancer_history" = 6 (unknown). Out of these, there were 131,210 cases for Class 0 and 14,579 cases for Class 1. This indicates that approximately 91.8% of the records were non-cancer cases, and 9.8% were positive cancer cases. While this distribution reflects real-world population data, from a machine learning perspective, it highlights the severely skewed class distribution in the NHS dataset. This class imbalance situation makes it difficult for machine learning models to make accurate predictions for the smaller category, positive gastric cancer (Class 1) cases. In this context, obtaining good prediction performance for positive gastric cancer cases using machine-learning models was expected to be extremely challenging.

To address the challenges of overfitting and ensure that the model generalizes well to unseen data, it is crucial to split the dataset into a training set and a testing set. The model is trained on the training set, and its performance is evaluated on the testing set. This approach helps prevent the model from simply memorizing the training data and allows it to learn patterns that can be applied to new, unseen data. When working with synthetic datasets, it is important to ensure that the generated data accurately represents the real-world data to which the model will be applied [31]. To assess this, Pearson correlation analysis can be employed. This analysis helps identify which features have the strongest correlation with the target variable in the synthetic gastric dataset and whether these correlations align with those observed in real-world data. By comparing the correlations, we can validate if the synthetic dataset captures the essential relationships between features and the target variable. Furthermore, class imbalance can still pose a challenge in synthetic datasets. In this case, it is necessary to identify which features are most informative for predicting the minority class, which is gastric cancer in this scenario. Using Pearson correlation analysis makes it easier to identify features that are highly correlated with the target variable for the minority class, even if the majority class is more prevalent in the dataset [32]. This process ensures that the resulting model is optimized for predicting the minority class and is less influenced by the class imbalance issue, as shown in Figs. 1 and 2. In this study, we utilized two supervised machine learning models: Naïve Bayes classifier and Logistic Regression, based on previous literature research. For this, the learning library WEKA was used to train the ML models and analyze the prediction performances. We used Sensitivity (also known as Recall) and Positive Predictive Value (PPV, also known as Precision) [33, 34].

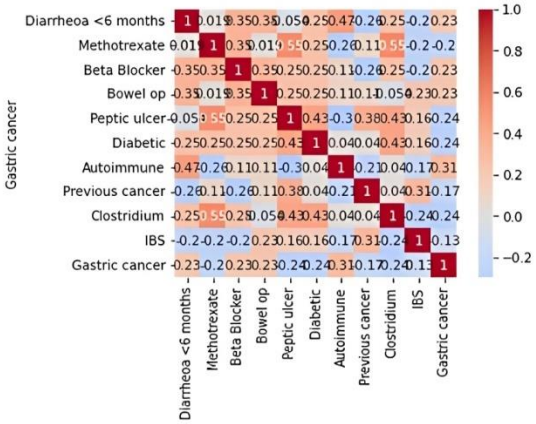

Figure 1. Pearson correlation coefficient of balanced dataset heatmap.

To train and test the above two machine learning models, we partitioned the NHS dataset into distinct training and testing sets. To study the impact of class imbalance, we built two sets of training datasets: "balanced" and

"unbalanced" using a stratified sampling method. The "unbalanced" training dataset consisted of 98,000 cases, with 88,200 cases of Class 0 and 9,800 cases of Class 1, reflecting the 90–10 ratio of the two categories in the original NHS dataset. The "balanced" training dataset consisted of 25,658 cases, with an equal distribution of 12,829 cases for both Class 0 and Class 1. The testing dataset contained 12,000 cases, which did not overlap with either the balanced or unbalanced training sets. It consisted of 11,000 cases of Class 0 and 1,000 cases of Class 1, reflecting the original distribution of data among the categories.

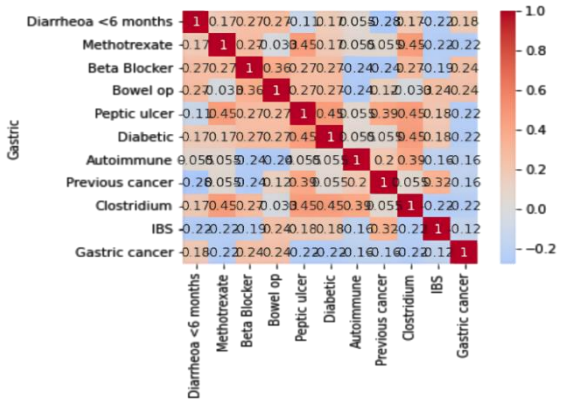

Figure 2. Pearson correlation coefficient of unbalanced dataset heatmap.

As stated earlier, our objective was to use the two machine learning models and sampled training sets (balanced and unbalanced) to improve prediction performance and decision support for the testing set. The testing set represented the real-life distribution of negative and positive gastric cancer cases that medical professionals are likely to encounter in practice. We performed a case-by-case analysis, comparing the predictions made by different ML models trained on distinct model-and-training set combinations (NB-Balanced, NB-Unbalanced, LR-Balanced, and LR-Unbalanced) on the same testing dataset of 11,000 cases. We examined the disagreements in predictions of different ML models and the prediction probabilities to explore approaches to improve the accuracy of Class 1 predictions. Fig. 3 shows the architecture design of the study, including the machine learning models, sampled training sets, and case-by-case analyses we conducted to address the class imbalance issue in the data. Since the smaller class, Class 1, had a very small percentage of cases in the unbalanced training dataset but a sizeable percentage of cases in the balanced training dataset, we hypothesized that the prediction performance of Class 1 would be considerably better with the balanced training dataset compared to the unbalanced training dataset for both the machine learning models NB and LR.

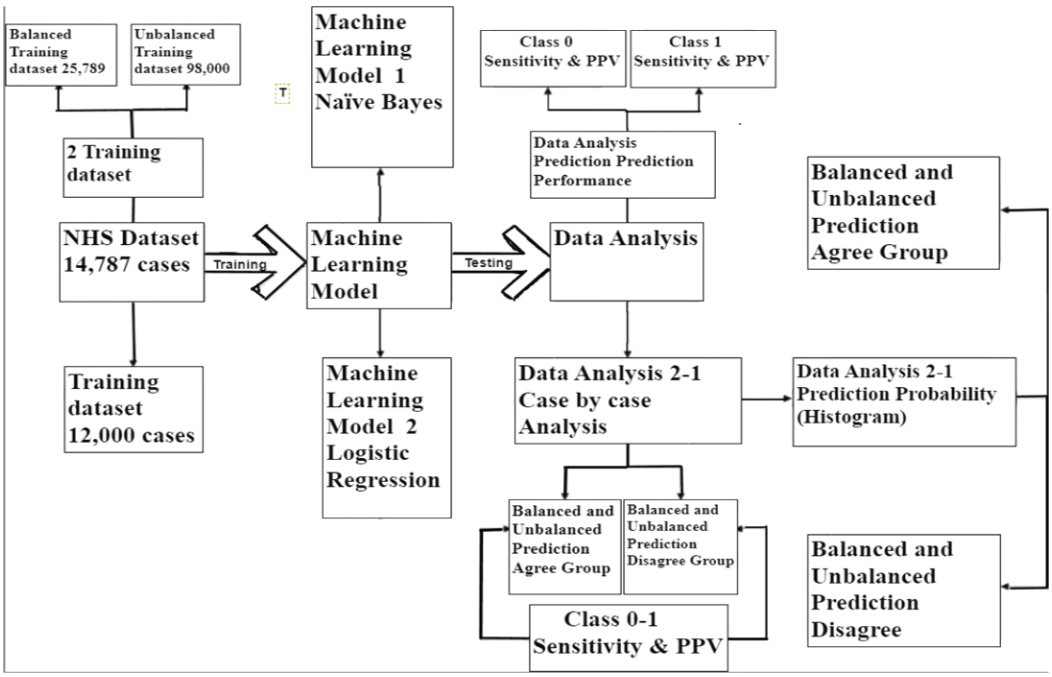

Figure 3. The architecture design of the study.

As shown in Fig. 3, the architecture design of the study addresses class imbalance in predicting gastric cancer using machine-learning methods. The process begins with the preparation of balanced and unbalanced training datasets based on the original NHS dataset, along with a separate testing dataset. Naïve Bayes and Logistic Regression algorithms are then applied to the training datasets. The testing results are divided into two groups: "Agree" and "Disagree" based on whether the predictions made by the balanced and unbalanced models agree or disagree. Sensitivity and Positive Predictive Value (PPV) are calculated separately for each group to evaluate the models' performance. Fig. 4 provides a visual representation of the steps involved in training, testing, and evaluating the models in the context of class imbalance in gastric cancer prediction.

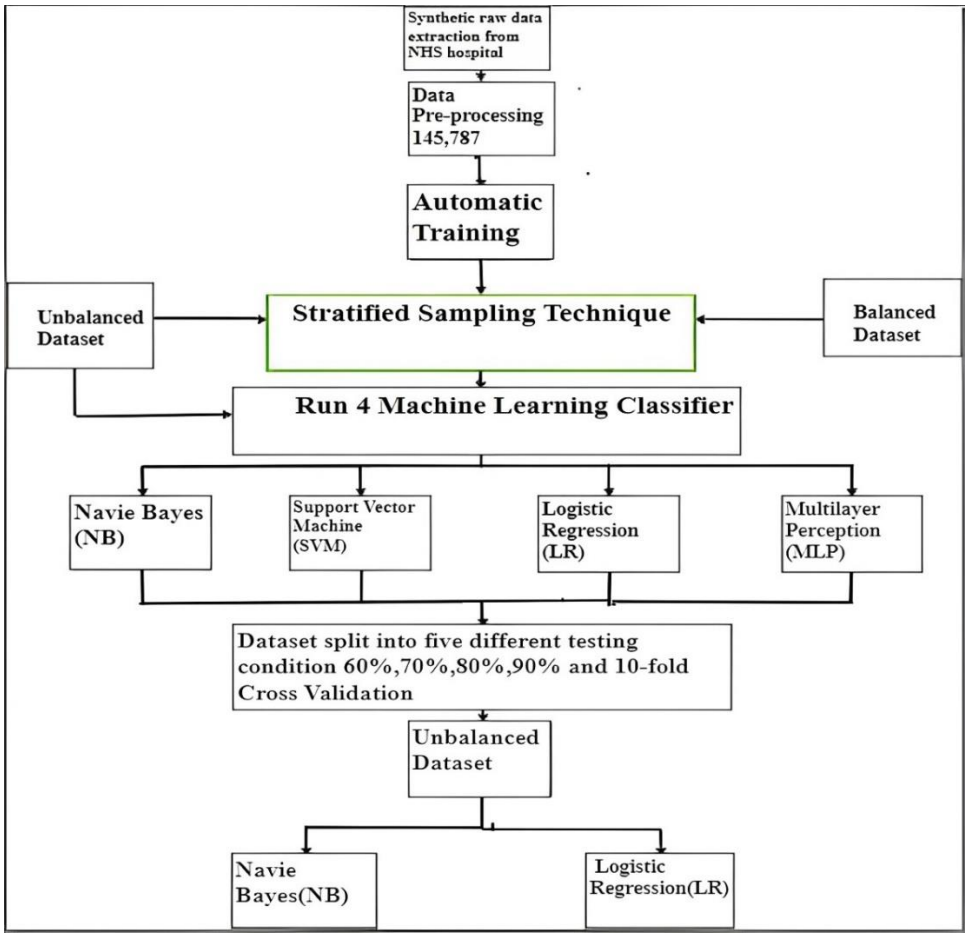

Figure 4. Flowchart of various steps performed in the study.

The gastric cancer dataset used in this study is not publicly available and was obtained from the NHS Liverpool University Hospital with approval from the responsible surgeon (coauthor). The data are anonymous and consist of records observed from 2009 to the 2021 calendar year. The dataset includes personal features, systemic conditions, gastric conditions, and separate fields for "diet" and "food information" about the individuals, as listed in Table II, along with their possible values. The features are categorized into four groups: personal characteristics, behavior, systemic features, and the gastric condition.

Table II illustrates the 21 different types of fields present in the dataset. The field "gastric cancer history" (previous gastric cancer diagnosis) was chosen as the dependent variable to be predicted based on the values of the other fields. The "year" variable was excluded from the study as it only indicated the year the record was created and did not provide relevant health or history context. The following 11 fields were considered as independent variables: "gastric_cancer_history", "previous cancer history", "Clostridium", "methotrexate", "diabetic", "autoimmune", "peptic_ulcer", "bowl op", "betablocker", "current diarrhea" (frequency count of this combination of covariates), and "cholecystectomy" (chole) (frequency count of this combination of "methotrexate"). The "count" column was the only numeric variable, while the remaining fields and variables were nominal. This study hypothesis is that training both models on a balanced dataset can improve their performance by mitigating bias towards the majority class and allowing equal representation of both classes in the training data. This can result in more accurate and unbiased predictions for both classes, including the minority class (Class 1). To test this hypothesis, we developed a Decision Support System (DSS) using two supervised machine learning models: Naive Bayes and Logistic Regression.

Overall, the results demonstrate that using a balanced training dataset can enhance the performance of the Naive Bayes and Logistic Regression models in predicting the likelihood of having gastric cancer. The original NHS dataset was divided into distinct training and testing sets to train and test the two selected ML models. Two sets of training datasets—"balanced" and "unbalanced"—were created using a stratified sampling method to examine the impact of class imbalance. The "unbalanced" training dataset consisted of 98,000 cases, with 88,200 cases of Class 0 and 9,800 cases of Class 1, reflecting the 90–10 ratio of the two categories in the original NHS dataset. In contrast, the "balanced" training set comprised 12,829 cases for both Classes 0 and 1. The testing dataset, separate from the balanced and unbalanced training sets, contained 12,000 cases with 11,000 cases of Class 0 and 1,000 cases of Class 1, reflecting the original distribution of data among the categories. The study aimed to enhance the prediction performance and decision support for the testing

set by utilizing the two machine learning models and sampled training sets (balanced and unbalanced). The testing set was designed to simulate the distribution of negative and positive gastric cancer cases that medical professionals are likely to encounter in practice.

In this research, WEKA 3.8 and Python machine learning software were used to evaluate four ML models (NB, LR, SVM, and MLP) for prediction purposes. The WEKA machine learning libraries offer various classification, clustering, and preprocessing classifiers to analyze different datasets [34]. Moreover, both Python and WEKA provide well-defined frameworks for building and testing models. The selected ML models were trained and tested under five different conditions, including 10-fold cross-validation and percentage splits ranging from 60% to 90% in 10% increments [34]. This study employed the WEKA library to evaluate the performance of the constructed machine-learning models and assess their accuracy, sensitivity, and specificity. 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 [35].

The primary objective of this research is to address the challenge of low prediction accuracy in gastric cancer cases caused by an imbalanced distribution of positive and negative instances. The aim is to develop an effective approach that enhances the accuracy of the machine learning-based decision-making process in predicting the likelihood of gastric cancer.

To achieve this objective, the proposed approach incorporates several key components. Firstly, a comprehensive data preprocessing stage is conducted, including data cleaning, feature extraction, and transformation. This stage ensures the quality and suitability of the data for subsequent analysis. Next, feature selection methods play a crucial role in identifying the most relevant and informative features for gastric cancer prediction [36]. These methods eliminate redundant or irrelevant features, reducing the dimensionality of the dataset and improving the efficiency and accuracy of the predictive models [37]. In this study, the researchers apply the Pearson Correlation Balanced method as a feature selection technique to identify the most significant features for gastric cancer prediction. This method calculates the correlation between each independent variable and the target variable while considering the class imbalance in the dataset. By considering the imbalanced distribution, the Pearson Correlation Balanced method measures the linear association between variables and helps identify the strength and direction of their relationship. By applying the Pearson Correlation Balanced method, the researchers can determine the features that exhibit a strong correlation with gastric cancer likelihood, taking into account the class imbalance issue [38]. Features with higher correlation coefficients are considered more relevant and informative for predicting the occurrence of gastric cancer. This feature selection process prioritizes the most influential features, allowing for more focused and accurate predictions. Following feature selection, multiple machine learning algorithms are employed to train predictive models, including Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and Multilayer Perceptron (MLP). Each algorithm possesses unique strengths and characteristics, enabling the model to capture different aspects of the data and make accurate predictions. In terms of complexity, the time and space complexities of the chosen algorithms vary. Naive Bayes has a time complexity of O (N * d) for training, where N represents the number of training instances and d denotes the number of features. Logistic Regression has a time complexity of O (k * N * d) for training, with k as the number of iterations. The time complexity of SVM depends on the chosen kernel function, ranging from O (N * d) for linear SVM to higher complexities for non-linear SVM with kernel functions like the Gaussian RBF. MLP's time complexity depends on the number of layers, the number of neurons, and the optimization algorithm used, ranging from O (N * e * L) to O (N * e * L^2). The space complexities vary as well, ranging from O (N * d) for Naive Bayes and SVM to O(d) for Logistic Regression and O (L * M) for MLP.

By implementing this proposed approach, it is expected that the prediction accuracy of gastric cancer likelihood can be significantly improved compared to traditional methods. The model considers the imbalanced distribution of positive and negative cases, utilizes effective data preprocessing techniques, employs feature selection to focus on relevant features, and utilizes multiple machine learning algorithms to capture different aspects of the data. These improvements enhance the decision-making process and provide more accurate predictions for gastric cancer likelihood. Regarding the complexity of model evaluation, based on the provided information:

Group: "Agree"

Number of cases: 9784

Calculation of sensitivity and PPV: Since we compare the model predictions with the originally assigned class labels in the testing set, the complexity of calculating sensitivity and PPV for this group is linear with respect to the number of cases. Therefore, the complexity is O(n), where n is the number of cases in the "Agree" group (9784).

Group: "Disagree"

Number of cases: 1218

Calculation of sensitivity and PPV: Similar to the "Agree" group, calculating sensitivity and PPV for the "Disagree" group is also linear with respect to the number of cases. Therefore, the complexity is O(m), where m is the number of cases in the "Disagree" group (1218).

To summarize, the complexity of the model evaluation process can be approximated as follows:

For the "Agree" group: O(n), where n is the number of cases (9784).

For the "Disagree" group: O(m), where m is the number of cases (1218).

Since the calculations for each group are performed separately, the overall complexity would be the sum of the complexities for each group:

Total complexity = O(n) + O(m)

IV. Result and Discussion

A. Supervised Machine Learning Performance

The predictive performance of Naïve Bayes and Logistic Regression models was evaluated on both balanced and unbalanced training datasets, as shown in Table III. Overall, the Naïve Bayes model outperformed the Logistic Regression model in both datasets. These results suggest that the proposed ensemble approach using Naive Bayes and Logistic Regression models can effectively address the issue of class imbalance and provide accurate predictions for gastric cancer likelihood.

For both ML models, the balanced training set resulted in better sensitivity and PPV values for Class 0 and better sensitivity values for Class 1 compared to the unbalanced training set. However, for Class 1, the PPV values for the unbalanced training set were higher for both Naïve Bayes (0.659) and Logistic Regression (0.755) models than the corresponding values for the balanced training set (0.422-NB and 0.553-LR). Overall, the balanced training dataset predicted Class 1 more accurately than the unbalanced training dataset, while also performing better for Class 0 and Class 1 in general.

B. Case-by-Case Scanning Followed by Prediction Probability Analysis

After examining the prediction performances of different ML models and training set combinations, a case-by-case analysis was performed on both Naïve Bayes and Logistic Regression models. For each ML model, the testing results were partitioned into two groups based on the predictions made for each case by the two models trained on balanced and unbalanced training sets. The first group, "Agree" included cases where predictions made by the balanced and unbalanced models (for both NB and LR) agreed with each other, and the second group, "Disagree" included cases where the predictions made by the balanced and unbalanced models (for both LR and NB) disagreed with each other. In the NB model, the "Agree" group consisted of 9,784 cases (89%), and the "Disagree" group consisted of 1,218 cases (11%) out of the total 11,000 cases in the testing set. For the LR model, the "Agree" group consisted of 9,800 cases (89%) and the "Disagree" group consisted of 1,200 cases (11%). Finally, sensitivity and PPV values were calculated separately by comparing the model predictions with the originally assigned class labels in the testing set for each group. For both ML models, the sensitivity and PPV of Class 1 for unbalanced predictions were 0, indicating that none of the predictions were correct. Similarly, the sensitivity and PPV for Class 0 were 0 for the balanced predictions for both ML models. Class 0 PPV also decreased significantly for the 88 Disagree-Unbalanced group (0.78-NB and 0.53-LR) for both models compared to their balanced (0.995-NB and 0.992-LR) and unbalanced (0.969-NB and 0.939-LR) training sets as shown in Tables III and IV.

As shown in Table V, when the ML models trained on balanced and unbalanced training datasets predicted differently, the Class 1 PPV dropped for both ML models. For Naïve Bayes, it decreased to 0.2186 (Disagree-Balanced) from 0.422 (Balanced), and for Logistic Regression, it decreased to 0.4738 (Disagree-Balanced) from 0.553 (Balanced).

TABLE III. Naïve Bayes and Logistic Regression Prediction Performances

		Naive Bayes		Logistic Regression	
		Balanced	Unbalanced	Balanced	Unbalanced
Class 0	Sensitivity	0.869	0.964	0.925	0.988
	PPV	0.995	0.969	0.992	0.939
Class 1	Sensitivity	0.953	0.687	0.927	0.358
	PPV	0.422	0.659	0.553	0.755

TABLE IV. Prediction Performances When Balanced and Unbalanced Predictions Agree

		Naive Bayes	Logistic Regression
Class 0	Sensitivity	0.9607	0.9876
	PPV	0.9946	0.9922
Class 1	Sensitivity	0.9360	0.8306
	PPV	0.6587	0.7553

TABLE V. Prediction Performances When Balanced and Unbalanced Predictions Disagree

		Naive Bayes		Logistic Regression	
		Balanced	**Unbalanced**	**Balanced**	**Unbalanced**
Class 0	Sensitivity	0.00	1.00	0.00	1.00
	PPV	0.00	0.7814	0.00	0.5262
Class 1	Sensitivity	1.00	0.00	1.00	0.00
	PPV	0.2186	0.00	0.4738	0.00

In this study, the distribution of prediction probabilities of the two ML models for the Agree and Disagree groups was analyzed to examine their correlation with prediction accuracy. Figs. 5–8 illustrates the prediction probability distributions when the balanced and unbalanced prediction performances agreed and disagreed for both the Naïve Bayes and Logistic Regression algorithms. Fig. 5(a) presents the distribution of prediction probabilities for the NB-Balanced model, while Fig. 5(b) displays the distribution for the NB-Unbalanced model in the Agree group. Similarly, Fig. 6(a) shows the distribution of prediction probabilities for the LR-Balanced model, and Fig. 8 presents the distribution for the LR-Unbalanced model in the Agree group. As shown in Figs. 5 and 6 demonstrates that when both the balanced and unbalanced datasets predict the same outcome for both Naïve Bayes and Logistic Regression algorithms, a similar pattern emerges. More than 90% of the cases exhibit prediction probabilities over 0.97, while only a few cases have prediction probabilities ranging from 0.5 to 0.9. Table IV indicates that the Sensitivity and PPV values for both Class 0 and Class 1 were significantly high in the Agree group. Thus, there appears to be a correlation between high prediction probability values and better prediction performance for both the LR and NB models.

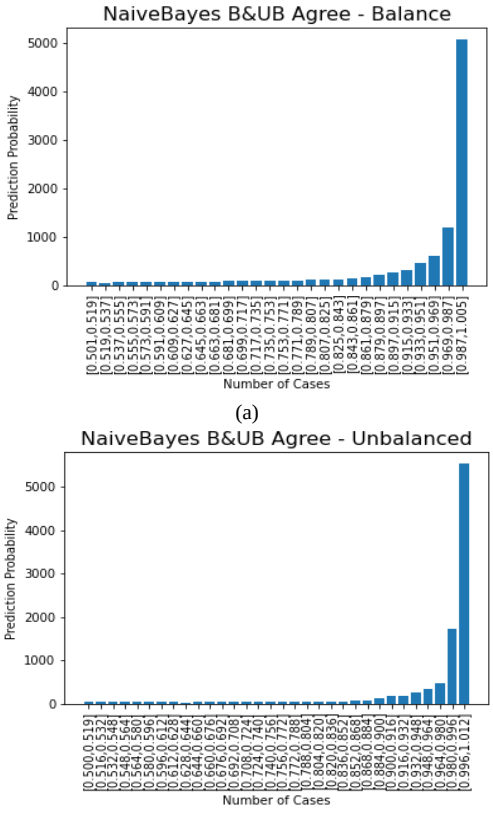

Figure 5. Prediction probability distribution of models for agree group: (a) NB-Balanced model; (b) NB-unbalanced model.

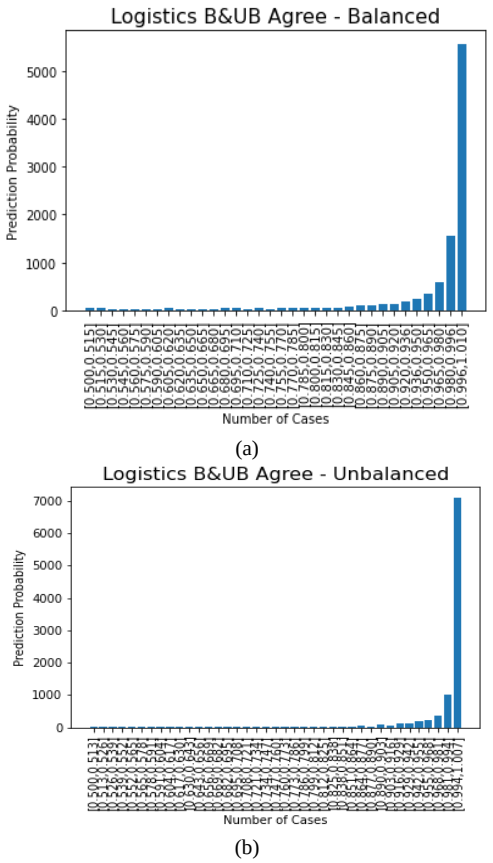

Figure 6. Prediction probability distribution of models for agree group: (a) LR-balanced model; (b) LR-unbalanced model.

As shown in Fig. 5(a), the distribution of prediction probabilities for the NB-Balanced model, and Fig. 5(b), for the NB-Unbalanced model, are presented for the Disagree group, where the predictions of NB-Balanced and NB-Unbalanced disagree. Similarly, Fig. 6(a) displays the distribution of prediction probabilities for the LR-Balanced model, and Fig. 8(b) illustrates the distribution for the LR-Unbalanced model in the disagree group.

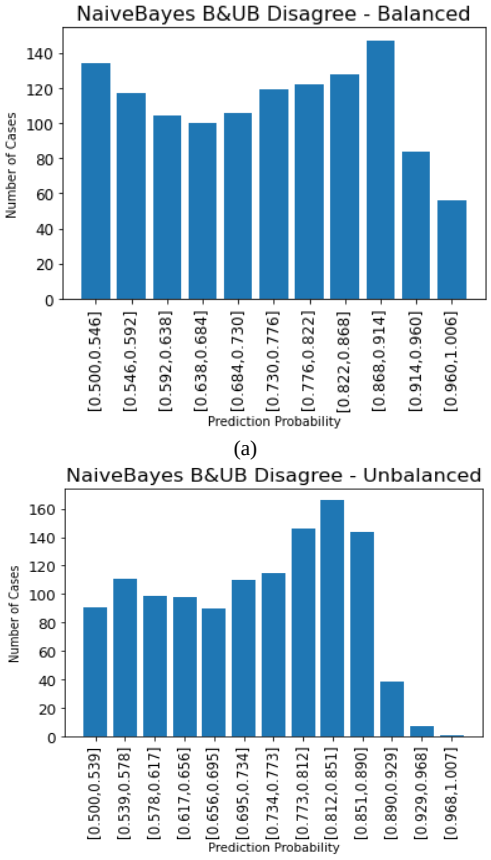

Figure 7. Probability distribution of models for disagree group: (a) NB-balanced model; (b) NB-unbalanced model.

From the four plots in as shown in Figs. 7 and 8, it can be observed that the prediction probabilities were uniformly distributed between 0.5 and 0.95 when the predictions of the balanced and unbalanced models disagreed. This is in stark contrast to the Agree set, where more than 90% of the cases exhibited extremely high prediction probabilities (as shown in Figs. 5 and 6. As shown in Table V, the prediction accuracy (Sensitivity and PPV) for the Disagree group was significantly lower compared to the balanced and unbalanced NB and LR models. This indicates that the prediction probabilities for both models were more evenly distributed when the model predictions were inaccurate. From a medical decision support perspective, the above findings suggest that the predictions made by the ML model for the Agree group can be considered reliably accurate and labeled as high-confidence ML predictions. On the other hand, cases in the Disagree group can be labeled as low-confidence ML predictions. Furthermore, providing the prediction probability values would be beneficial from a decision

support perspective, as low prediction probability values can indicate missing values or cases where ML models are unable to make clear predictions due to uncommon combinations of variable values, as shown in Fig. 8(b).

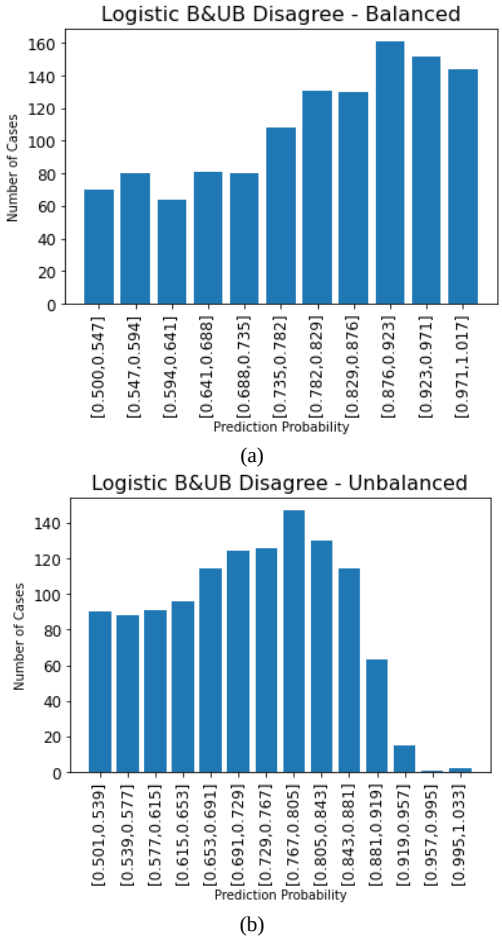

Figure 8. Prediction probability distribution of models for disagree group: (a) LR-balanced model; (b) LR-unbalanced model.

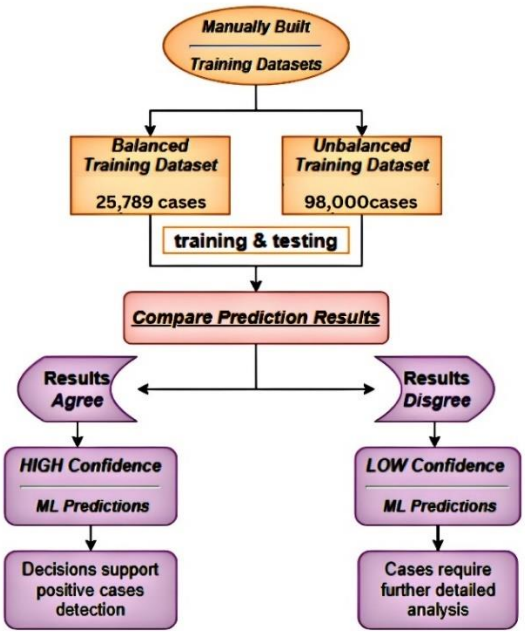

Figure 9. Decision-making logic derived from the study.

The results also indicate that solely providing prediction probabilities along with the prediction may not be entirely reliable, as there were cases in the Disagree group with low Sensitivity and PPV values but high prediction accuracy, particularly depicted in Figs. 6(a), 8, and 9. Overall, the findings of this study suggest that balancing the training dataset can enhance the performance of prediction models for gastric cancer. However, it is crucial to carefully consider the trade-off between sensitivity and PPV. Additionally, case-by-case analysis and prediction probability distribution analysis can offer valuable insights into model performance and help identify areas for improvement.

C. Results

The study aimed to evaluate the performance of Naïve Bayes and Logistic Regression models in predicting gastric cancer likelihood using supervised machine learning. Two datasets, one balanced and one unbalanced, were used for training the models. The results demonstrated that the Naïve Bayes model outperformed the Logistic Regression model in both datasets, indicating its effectiveness in addressing class imbalance and achieving accurate predictions for gastric cancer likelihood. When comparing the balanced and unbalanced datasets, it was observed that the balanced training set yielded better sensitivity and Positive Predictive Value (PPV) values for Class 0, whereas the unbalanced training set performed better in terms of sensitivity and PPV values for Class 1. This highlights the need to carefully consider the trade-off between sensitivity and PPV when selecting the training dataset. A case-by-case analysis was conducted to examine the predictions made by the models trained on the balanced and unbalanced datasets. Cases were categorized as "Agree" or "Disagree" based on the alignment or discrepancy between the predictions of the two models. The "Agree" group exhibited higher sensitivity and PPV values, indicating more accurate predictions. Conversely, the "Disagree" group had lower PPV values and included cases where there were no correct predictions for positive instances. The analysis of prediction probabilities revealed that cases with consistent predictions between the balanced and unbalanced datasets had higher prediction probabilities, suggesting more accurate predictions. On the other hand, when the predictions differed, the prediction probabilities were uniformly distributed between 0.5 and 0.95, indicating less reliable predictions. The proposed method of relying on high-confidence ML-based predictions and filtering out weaker predictions proved to be effective in addressing the class imbalance issue and improving the prediction accuracy of ML-based gastric cancer likelihood decision-making processes. This approach can provide more reliable diagnoses and be valuable for decision support to healthcare professionals and researchers.

D. Interpretation and Discussion

In the context of previous studies, the findings align with research indicating the successful application of Naïve Bayes models in medical classification tasks, including cancer prediction. Logistic Regression, although

commonly used, may struggle with class imbalance, leading to lower performance. The superior performance of the Naïve Bayes model can be attributed to its assumption of feature independence, which may align well with the dataset's characteristics and the patterns of gastric cancer likelihood. On the other hand, Logistic Regression's performance may have been hindered by its inability to effectively handle class imbalance. The results emphasize the importance of consistency in model predictions for improved accuracy. Cases with consistent predictions showed higher sensitivity and PPV values, indicating their reliability in predicting gastric cancer likelihood. In contrast, cases with inconsistent predictions demonstrated lower PPV values and less confidence in their predictions.

The choice of Logistic Regression (LR) and Naive Bayes (NB) models in this study is justified based on several reasons. First, LR and NB were selected due to their effectiveness in handling class imbalance issues in datasets. The gastric cancer dataset used in this research exhibits significant class imbalance, with a 9:1 ratio of non-cancer to gastric cancer-diagnosed cases. It is crucial to employ models that can adapt to such imbalanced distributions. LR and NB have built-in mechanisms to adjust their predictions and account for the class imbalance, making them suitable choices for this specific problem. Additionally, LR and NB models are known for their interpretability, providing insights into the factors influencing the predictions. This characteristic allows for a better understanding of the underlying relationships between the features and the target variable. In medical applications, interpretability is of great importance as it enables healthcare professionals to trust and make informed decisions based on the predictions. Deep learning techniques, on the other hand, often lack interpretability and operate as black boxes, which can be a limitation in medical contexts. LR and NB models have been extensively studied and have shown good performance in various domains, including medical applications. They have a solid track record of delivering reliable results in a wide range of scenarios. While there are many ML and deep learning techniques available, LR and NB have proven to be effective and efficient in different classification tasks. Their simplicity and effectiveness make them practical choices, especially when working with limited data or computational resources. Furthermore, LR and NB models offer computational efficiency compared to more complex deep learning architectures. Deep learning models often require large amounts of data and computational resources to train and optimize. In contrast, LR and NB models can deliver accurate predictions with relatively lower training and inference times. In medical settings where real-time predictions and quick response times are crucial, the computational efficiency of LR and NB models can be advantageous.

Considering these factors, the use of Logistic Regression (LR) and Naive Bayes (NB) models in this study is justified to address the class imbalance issue in gastric cancer likelihood prediction. These models provide a balance between interpretability, performance, and computational efficiency, making them well-suited for the research objective and the specific characteristics of the dataset.

E. Implications and Limitations

The study highlights the significance of addressing class imbalance in gastric cancer prediction using machine learning [9]. The superiority of the Naïve Bayes model suggests its value as a tool in this domain. The findings have implications for researchers and practitioners working on predictive models for cancer diagnosis and risk assessment [39]. However, it is crucial to acknowledge the limitations of the study. The evaluation was restricted to Naïve Bayes and Logistic Regression models, and other algorithms may yield different results. The use of manually created balanced and unbalanced datasets may not fully capture the complexity of real-world data. Therefore, further research with larger and more diverse datasets is necessary to validate the findings and enhance the generalizability of the results. Additionally, the study did not explore other factors that could impact prediction accuracy, such as additional features or preprocessing techniques, which should be considered in future research.

V. Conclusion

The study highlights the importance of addressing class imbalance in predictive models for gastric cancer likelihood using machine learning. The findings demonstrate the advantages of using a balanced training dataset in terms of sensitivity and Positive Predictive Value (PPV) for Class 0, while the unbalanced training dataset performs better in sensitivity and PPV for Class 1. However, the trade-off between ruling out negative cases and accurately identifying positive cases should be carefully considered. The case-by-case analysis reveals that consistent predictions between the balanced and unbalanced models result in higher sensitivity and PPV values, indicating more accurate predictions. On the other hand, cases with inconsistent predictions show lower PPV values and even reach 0 sensitivity, indicating incorrect predictions for positive instances. Furthermore, the prediction probability analysis demonstrates that cases with consistent predictions have significantly higher prediction probabilities, suggesting more accurate predictions. Conversely, when the predictions differ, the prediction probabilities are uniformly distributed, indicating less reliable predictions. Based on these findings, a decision logic is proposed to classify predictions into high-confidence and low-confidence categories. High-confidence predictions show improved accuracy in predicting gastric cancer likelihood, while low-confidence predictions require further testing by professionals for accurate diagnosis.

In light of these results, future research could explore alternative machine learning algorithms beyond Naïve Bayes and Logistic Regression models. Additionally, the validation of findings and improving generalizability can be achieved through larger and more diverse datasets. Further investigation into additional features and

preprocessing techniques should also be considered to enhance prediction accuracy. In summary, the proposed method offers a valuable approach to address the challenges posed by class imbalance in gastric cancer prediction. Future research should focus on exploring alternative algorithms, incorporating larger datasets, and investigating additional features and preprocessing techniques to further improve the accuracy and reliability of machine learning-based decision-making processes in gastric cancer likelihood assessment.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Danish Jamil played a major role in the conceptualization of the study, design of the decision support system, and methodology development. Sellappan Palaniappan contributed to the methodology development, particularly in refining the ML models (Naïve Bayes and Logistic Regression) and their implementation. Asiah Lokman and Danish Jamil were responsible for data collection and preparation. Sellappan Palaniappa provided input and expertise in the methodology development, particularly in the evaluation of prediction performance measures and statistical analysis. All authors had approved the final version.

REFERENCES

[1] R. Cao *et al*., "Artificial intelligence in gastric cancer: Applications and challenges," *Gastroenterol. Rep.*, vol. 10, 2022.

[2] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics," *CA. Cancer J. Clin.*, vol. 73, no. 1, pp. 17–48, 2023.

[3] B. S. Chhikara and K. Parang, "Global cancer statistics 2022: The trends projection analysis," *Chem. Biol. Lett.*, vol. 10, no. 1, 451, 2023.

[4] B. Park, S. Yang, J. Lee, I. J. Choi, Y.-I. Kim, and J. Kim, "Gastric cancer risk prediction using an epidemiological risk assessment model and polygenic risk Score," *Cancers (Basel).*, vol. 13, no. 4, 876, 2021.

[5] D. Banik and D. Bhattacharjee, "Mitigating data imbalance issues in medical image analysis," *Data Preprocessing, Active Learning, and Cost Perceptive Approaches for Resolving Data Imbalance*, IGI Global, pp. 66–89, 2021.

[6] A. Lokman *et al*., "Prediction model for gastric cancer via class balancing techniques," *Int. J. Comput. Sci. Netw. Secur.*, vol. 23, no. 1, pp. 53–63, 2023.

[7] Y. Lyu, Q. Xu, Z. Yang, and J. Liu, "Prediction of patient choice tendency in medical decision-making based on machine learning algorithm," *Front. Public Heal.*, vol. 11, 2023.

[8] L. E. Cipriano, "Evaluating the impact and potential impact of machine learning on medical decision making," *Medical Decision Making*, 2023.

[9] S. S. Martinez *et al.*, "Machine learning for clinical decision-making: Challenges and opportunities in cardiovascular imaging," *Front. Cardiovasc. Med.*, vol. 8, 2022.

[10] P. H. Niu, L. L. Zhao, H. L. Wu, D. B. Zhao, and Y. T. Chen, "Artificial intelligence in gastric cancer: Application and future perspectives," *World J. Gastroenterol.*, vol. 26, no. 36, 5408, 2020.

[11] P. Y. Zhou and A. K. C. Wong, "Explanation and prediction of clinical data with imbalanced class distribution based on pattern discovery and disentanglement," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–15, 2021.

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

[13] D. B. Demirhan, H. Yilmaz, H. Erol, H. M. Kayili, and B. Salih, "Prediction of gastric cancer by machine learning integrated with mass spectrometry-based N-glycomics," *Analyst*, no. 9, 2023.

[14] D. Jamil, S. Palaniappan, S. S. Zia, A. Lokman, and M. Naseem, "Reducing the risk of gastric cancer through proper nutrition-a meta-analysis.," *Int. J. Online & Biomed. Eng.*, vol. 18, no. 7, 2022.

[15] D. Jamil, "Diagnosis of gastric cancer using machine learning techniques in healthcare sector: A survey," *Informatica*, vol. 45, 2022.

[16] D. Jamil *et al.*, "Diagnosis of gastric cancer using machine learning techniques in healthcare sector: A survey," *Informatica*, vol. 45, no. 7, 2022.

[17] D. Liu *et al.*, "Machine learning-based model for the prognosis of postoperative gastric cancer," *Cancer Manag. Res.*, pp. 135–155, 2022.

[18] C. Zhou *et al.*, "A machine learning-based predictor for the identification of the recurrence of patients with gastric cancer after operation," *Sci. Rep.*, vol. 11, no. 1, pp. 1–7, 2021.

[19] X. Yuan, S. Chen, C. Sun, and L. Yuwen, "A novel early diagnostic framework for chronic diseases with class imbalance," *Sci. Rep.*, vol. 12, no. 1, pp. 1–16, 2022.

[20] L. M. Terracciano *et al.*, "Opportunities and challenges for machine learning in rare diseases," *Frontiers in Medicine*, vol. 8, 747612, 2021.

[21] S. Shilaskar, A. Ghatol, and P. Chatur, "Medical decision support system for extremely imbalanced datasets," *Inf. Sci. (Ny).*, vol. 384, pp. 205–219, 2017.

[22] P. Ö. Kavas *et al*.,"Machine learning-based medical decision support system for diagnosing HFpEF and HFrEF using PPG," *Biomed. Signal Process. Control*, vol. 79, 104164, 2023.

[23] M. Rashid, J. Kamruzzaman, T. Imam, S. Wibowo, and S. Gordon, "A tree-based stacking ensemble technique with feature selection for network intrusion detection," *Appl. Intell.*, vol. 52, no. 9, pp. 9768–9781, 2022.

[24] B. A. Tama, S. Im, and S. Lee, "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble," *Biomed Res. Int.*, vol. 20, 2020.

[25] A. Mortezagholi, O. Khosravizadehorcid, M. B. Menhaj, Y. Shafigh, and R. Kalhor, "Make intelligent of gastric cancer diagnosis error in Qazvin's medical centers: Using data mining method," *Asian Pacific J. Cancer Prev.*, vol. 20, no. 9, pp. 2607–2610, 2019.

[26] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, "Supervised data mining approach for predicting student performance," *Indones. J. Electr. Eng. Comput. Sci.*, pp. 1584–1592, 2019.

[27] M. Zikeba, J. M. Tomczak, M. Lubicz, and J. Świkatek, "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients," *Appl. Soft Comput.*, vol. 14, pp. 99–108, 2014.

[28] M. N. Hasnine, G. Akcapinar, B. Flanagan, R. Majumdar, K. Mouri, and H. Ogata, "Towards final scores prediction over clickstream using machine learning methods," in *Proc. 6th International Conference on Computers in Education*, 2018, pp. 399–404.

[29] J. Brownlee, "How to use classification machine learning algorithms in weka," *Mach. Learn. Mastery*, vol. 2, 2016.

[30] R. Hasan, S. Palaniappan, S. Mahmood, K. U. Sarker, and A. Abbas, "Modelling and predicting student's academic performance using classification data mining techniques," *Int. J. Bus. Inf. Syst.*, vol. 34, no. 3, pp. 403–422, 2020.

[31] D. Chicco and G. Jurman, "The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, pp. 1–13, 2020.

[32] H. Felippe *et al.*, "Threshold-free estimation of entropy from a Pearson matrix," *Europhys. Lett.*, vol. 141, no. 3, 31003, 2023.

[33] M. Das and R. Dash, "A comparative study on performance of classification algorithms for breast cancer data set using WEKA tool," *Intelligent Systems*, Springer, pp. 289–297, 2022.

[34] P. Fergus and C. Chalmers, "Performance evaluation metrics," *Applied Deep Learning: Tools, Techniques, and Implementation*, Springer, pp. 115–138, 2022.

[35] Y. Li and Z. Chen, "Performance evaluation of machine learning methods for breast cancer prediction," *Appl Comput Math*, vol. 7, no. 4, pp. 212–216, 2018.
[36] M. J. Iqbal *et al.*, "Clinical applications of artificial intelligence and machine learning in cancer diagnosis: Looking into the future," *Cancer Cell Int.*, vol. 21, no. 1, pp. 1–11, 2021.
[37] Z. Xiao, D. Ji, F. Li, Z. Li, and Z. Bao, "Application of artificial intelligence in early gastric cancer diagnosis," *Digestion*, vol. 103, no. 1, pp. 69–75, 2022.
[38] S. A. Mahmoodi, K. Mirzaie, M. S. Mahmoodi, and S. M. Mahmoudi, "A medical decision support system to assess risk factors for gastric cancer based on fuzzy cognitive map," *Comput. Math. Methods Med.*, vol. 2020, 2020.
[39] U. Shaham *et al.*, "A deep learning approach to unsupervised ensemble learning," in *Proc. International Conference on Machine Learning*, 2016, pp. 30–39.