Prediction Model for Gastric Cancer via Class Balancing Techniques

Danish Jamil [1††], Sellappan Palaniappan[2††], Sanjoy Kumar Debnath [3††], Muhammad Naseem[4††]
Susama Bagchi [5††] and Asiah Lokman [6††]

[1††,2††],[6††]Department of Information Technology, Malaysia University of Science and Technology, Petaling Jaya, Malaysia
[3††,5††] Chitkara University Institute of Engineering and Technology, Chitkara University Punjab, India

[1††,4††] Department of Software Engineering, Syed University of Engineering and Technology, Karachi, Pakistan
*Correspondence: danish.jamil@phd.must.edu.my

Abstract
Many researchers are trying hard to minimize the incidence of cancers, mainly Gastric Cancer (GC). For GC, the five-year survival rate is generally 5–25%, but for Early Gastric Cancer (EGC), it is almost 90%. Predicting the onset of stomach cancer based on risk factors will allow for an early diagnosis and more effective treatment. Although there are several models for predicting stomach cancer, most of these models are based on unbalanced datasets, which favours the majority class. However, it is imperative to correctly identify cancer patients who are in the minority class. This research aims to apply three class-balancing approaches to the NHS dataset before developing supervised learning strategies: Oversampling (Synthetic Minority Oversampling Technique or SMOTE), Undersampling (SpreadSubsample), and Hybrid System (SMOTE + SpreadSubsample). This study uses Naive Bayes, Bayesian Network, Random Forest, and Decision Tree (C4.5) methods. We measured these classifiers' efficacy using their Receiver Operating Characteristics (ROC) curves, sensitivity, and specificity. The validation data was used to test several ways of balancing the classifiers. The final prediction model was built on the one that did the best overall.

Keywords
Gastric Cancer, Imbalance dataset, Synthetic Minority Oversampling Technique, Receiver Operating Characteristics (ROC)

I. INTRODUCTION

In 2019, nearly 28,000 cases of GC were reported (17,750 males and 10,250 females). About 10,960 people are known to have died of cancer (6,720 men and 4,240 women). Almost 70% to 90% of all GCs start with h.pylori infection. It circulates in the human body through uncooked or unwashed food.

Salty foods are more likely to cause an increase in GC, which can develop into a tumour. In Japan, approximately 30 out of every 100,000 people have been diagnosed with GC at some point. There is no way to avoid GC; if the doctor finds the patient has severe symptoms, GC becomes a tumour. Operations, chemotherapy, therapy, and radiation therapy are the best treatments for patients[1]. This period of delay may cause cancer to deteriorate to the point that it is too late[2] for patients to get comprehensive treatment. Researchers have proposed using an intelligent decision-support system that recognizes the different cancer forms. It will be helpful for both patients and doctors in terms of the treatments they can choose from and how much they will cost [6]. So, if gastric cancer is found and treated earlier, the patient may have a much better chance of beating the disease. Several risk factors must be examined to predict the chance of getting gastric cancer. Identifying a person's risk of developing stomach cancer necessitates familiarity with gastric cancer risk factors. These include gender, age, BMI, previous history of gastric cancer, and others[3].

H. pylori, ASA, ethnicity, diarrhea lasting less than six months, and other risk factors have also been observed. All of these possible risks are taken into account in our analysis. This analysis does not look at other things that can cause gastric cancer, like genetics and lifestyle choices like smoking and drinking. Different stages of stomach cancer have very different survival rates, so it's important to get a diagnosis as soon as possible. Those who have the condition at an earlier, less invasive stage are more likely to survive it than those who have it at a later, more invasive stage. When a patient has stomach cancer, doctors must make a correct diagnosis and avoid false positive findings[4][5].

This research aims to help doctors make fair and accurate decisions about how likely a patient is to have stomach cancer. examine the performance of various ML models with regards to class imbalance issues for gastric cancer prediction by using various data sampling techniques to create more balanced data distribution, including, oversampling of smaller category cases, under-sampling of larger category cases, and others[6][7]. As previously stated, the fundamental challenge that this article seeks to address is the class imbalance problem that occurs in the dataset provided by NHS Liverpool hospital. As it is a problem that reaches across domains, our work focuses on developing and applying methods for several practical problems heavily encumbered by class imbalance. The aim of this paper is as follows:

Manuscript received January 5, 2023
Manuscript revised January 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.1.8

- To make a fair decision support system for doctors that will help them find stomach cancer early and increase the number of people who survive it.
- Development of a prediction model for the occurrence of gastric cancer by applying class-balancing methods to data on gastric cancer risk factors.

Paper Outline. The remaining portion of the paper is structured as follows: Section 2 describes the literature review and major research gap in previous studies. Section 3 discusses the background of the study. It focuses on the GC risk factors associated with synthetic data and highlights critical risk factors with GC treatment. Then, it describes the material and methods used to explain the methodological procedure. Section 4 discusses the findings. Section 5 concludes and provides direction for future work.

II. Literature Review

This paper's objective was to perform research in diagnosing pathological samples for imbalanced classes. The study may be broken down into three primary research lines: the resampling approach, the optimization of multi-class classifiers, and the selection of performance measurements. The suggested resampling method is straightforward. Oversampling and under-sampling are used to control the size of the dataset. The two steps of SVMFS are the grid search and the vector search, which are part of the hybrid filter wrapper technique[8]. The M-PSO algorithm utilizes a swarm-based optimization method and a randomly generated feature vector. Both methods enhance the accuracy of categorization. When comparing M-PSO to SVMFS, M-PSO has a lower classifier set training time. The proposed classification methods, M-PSO, SVMFS, and synthetic sampling, perform very well on multi-class classification tasks. The empirical findings demonstrate the efficacy of the suggested categorization algorithms. We use a set of nine metrics to evaluate the efficacy of classifier algorithms. This analysis requires testing holdouts or previously unknown data. The suggested algorithms will eventually be implemented in actual clinical diagnostic systems[9].

This article looks at how well different classifiers can predict the type of breast cancer that will come back and finds that neural networks do the best. [10] In this work provides a method for boosting the efficiency and precision of three popular classifiers: the Decision Tree (J48), the Naive Bayes (NB), and the Sequential Minimal Optimization (SMO) (SMO). To verify and compare the classifiers, use two standard sets, Wisconsin Breast Cancer (WBC) and Breast Cancer dataset. K-Nearest Neighbors, Support Vector Machine (SVM), Logistic Regression, Decision Tree (C4.5), and SVM were the five machine learning techniques used (KNN). The study [11] gives an overview of AI methods, feature predictors, typical training and testing methodologies, assessment metrics, and systems application in clinical practice for estimating recurrence risk in breast cancer, as provided in a recent article. Despite many publications over the previous decade, this issue is still not fully resolved. It is difficult to train surgical instrument detectors because of the class imbalance between different types of surgical tools. This work addresses this by proposing a semi-supervised learning-based training strategy[12]. To begin, we annotated recordings of 24 instances of robotic gastrectomy for stomach cancer to identify the initial bounding box of the surgical tools. Next, unlabeled movies were separated using a trained instrument detector, and new labels were added to the tools, leading to class imbalance based on the statistics of the labeled videos. In this study[13] found that the GNB, XGBoost, and random forest algorithms were the most effective in predicting overall survival (OS), distant metastases (DM), and peritoneal metastases (PM), respectively. More precise machine learning research, in many instances, is required to find the most accurate algorithm and make tailored therapies available in the next few years. This study's [14] goal is to provide a soft computing-based medical decision support system that makes use of fuzzy cognitive mapping (FCMs) to aid doctors in selecting the most effective course of treatment for each individual patient, taking into account their unique illness risk profile. FCMs are widely regarded as one of the most powerful AI methods for modeling complicated systems. The purpose [15] of this research is to examine if data mining methods and the features of diseases associated with increased risk may be used to make accurate predictions and diagnoses of stomach cancer. The SVM algorithm produced the highest quality classifications when applied to test samples. Therefore, this smart technology may be employed as a physician assistant in facilities that teach future doctors to diagnose patients. In this study[16] focuses on addressing the issue of imbalanced data in predictive models for breast cancer. The researchers apply three class balancing techniques (SMOTE, SpreadSubsample, and a hybrid method) on the BCSC dataset to create more balanced datasets. Four classifiers (Naïve Bayes, Bayesian Network, Random Forest, and Decision Tree) were then used on the balanced datasets to create predictive models. The best-performing model was determined by evaluating the classifiers' performance using ROC curve, sensitivity, and specificity. The aim is to have a more accurate and efficient way of diagnosing and treating breast cancer, which is a leading cause of fatality among women

From the above discussion, it is evident from the previous research that most of the ML algorithms used in gastric cancer prediction have provided erroneous and unbalanced prediction outcomes. As a result, patients are forced to depend on doctors for cancer diagnosis, even though doing so might take weeks or months, contributing to a higher risk of the disease progressing and, ultimately, a worse prognosis. Better prediction models may analyze enhanced medical data to provide cutting-edge health informatics, allowing for more expedited and effective medical treatment. This research fills this gap in the literature. The goals of our research are to identify the most risk factor related that are involved in the development of GC and also effective machine learning (ML) method for predicting gastric cancer, to assess how class-imbalanced data affects ML-based gastric cancer prediction, and to propose a strategy for addressing this problem so that more accurate predictions may be made.

A. The Constraints of Existing Models

Machine learning models' diagnostic efficacy and accuracy are profoundly affected by the attributes or features taken from datasets[17]. Even though several investigations have been conducted on the issue of an attribute or feature selection and the extraction of data from several well-organized datasets, it is still crucial to choose the best attributes without modifying them, as this dramatically decreases the computational complexity and training time of the model and increases its accuracy[7][18]. Previous studies have not considered common but significant issues such as outliers, noise, unnormalized data, and high computing costs. This is the case even if these issues have been identified. Also, it is important to keep the level of computer complexity as low as possible[3].The computational complexity is proportional to the number of trained attributes. For this reason, it is crucial to determine the bare minimum of features necessary for reliable tumour classification. In addition to selecting features, we urgently need new or custom-tailored model structures to improve diagnosis even more.

III. Materials and Methods

The method involves several important steps, such as choosing the target data, preprocessing the chosen data, putting the data into a structured and easy-to-understand format, balancing the dataset, using supervised learning techniques, and evaluating the performance of machine learning using evaluation measures. These steps ultimately lead to the extraction of knowledge from the target dataset, where new insights and ideas can be developed to enhance business operations or, in this case, to assist in the early diagnosis and prediction of diseases such as gastric cancer[19].

A. Dataset Selection

The Gastric Cancer Dataset is not publicly available; it is from the NHS Liverpool University Hospital and has been used with approval from the responsible surgeon (coauthor), and the data are all anonymous. The study used the NHS Liverpool hospital dataset, with records observed from 2009 to the 2021 calendar year. The dataset contained unique features, systemic conditions, stomach conditions, and diet food about the individuals in separate fields, as listed in Table I, with possible values for each field. The features are arranged into four groups: personal characteristics, behavior, systemic features, and the stomach condition. The original NHS Liverpool hospital dataset includes 1255,789 records observed over 12 years from 2009 to 2021. The original dataset contains 40 variables representing patients' physiological and biographical information. Table I below shows the names and descriptions of the variables and their measurement values after pre-processing the dataset. Many risk factors were deleted because they did not contain any relevant information regarding the patient's health. After pre-processing the original GC dataset, this study had 145,789 records observed over; a period 12-year from 2009 to 2021 was provided by an NHS Liverpool hospital and is constituted of data from GC patients. The dataset contains 18 variables representing various clinical information.

TABLE I. NHS HOSPITAL DATASET DESCRIPTION

S.No	Variables Names	Coded Values Indicate the stage of GC
1	Years	Numerical 2009-2021
2	Age	1= Age 18-29 2= Age 30-34 3= Age 35-39 4= Age 40-44 5= Age 45-49 6= Age 50-54 7= Age 55-59 8= Age 60-64 9= Age 65-69 10= Age 70-74 11= Age 75-79 12= Age 80-84 13= Age >85
3	High_blood_pressure	0= No 1= Yes 9 = Not known
4	Diarrheoa <6 months	0= No 1= Yes 9 = Not known
5	Medical_history_IBD	0= No 1= Yes 9 = Not known
6	Serum sodium	0= No 1= Yes 9 = Not known
7	gastric_cancer_history	0= No 1= Yes 9 = Not known
8	Associated_factor Methotrexate	1= Almost entirely 2= Scattered fibroglandular 3= Heterogeneously 4= Extremely 9 = Not known or different measurement system
9	BMI group	1= 10-24.99 2= 25-29.99 3= 30-34.99 4= 35 or more 9 = Not known
10	Smoking	0= No 1= Yes 9 = Not known

B. Data Pre-Processing and Transformation

The pre-processing phase involves cleaning the chosen data of outliers, missing numbers, and other irregularities. A model that produces inaccurate findings or incorrect diagnoses of test data due to inconsistencies in the selected data might have catastrophic consequences[20]. The deletion of irrelevant variables is one of the procedures that take place during the pre-processing stage. This is because the purpose of the research may be accomplished without using unrelated variables. In addition, missing values or anomalies might arise because of a lack of information and approximate measurement values, resulting in insufficient precision and a higher percentage of error throughout the data assessment process. Before applying the model, imputation has to be done since cancer datasets often include missing values[21]. This makes it challenging to manage missing data. The missing values for the nominal and numerical characteristics in the dataset were filled in using the modes and means taken from the training data. Since all of the variables were determined to be of the nominal (categorical) type, modes, which are values that occur most often, were taken from the training data and used to impute missing values.

In order to continue processing the data, it needs to be converted into an acceptable format that can be read and is compatible with the data mining methods that have been applied to the dataset[22]. Transformation is required to meet the requirements of various types of data mining methods. One example of this transformation is converting numerical values into nominal ones.

C. Feature Selection

Selecting the features is the second phase, which utilizes a few different filter-based approaches. The recursive feature elimination approach comes after the study that uses correlation analysis to determine which predictors need to be addressed as the most important. This method is superior to other nonparametric methods, such as the K-Nearest Neighbors, which cannot rank predictors according to the value they play in the overall prediction when selecting the most accurate predictors[23].Because this will result in a better selection of optimum features, we suggest combining a correlation-based elimination approach and recursive feature elimination. This will bring about the desired effect. Even after features are removed using correlation, there is still a possibility that there are characteristics that are not particularly valuable; as a result, a second step employing recursive feature elimination will guarantee that the appropriate features are selected[24].

D. Classification

The classification stage is the third phase. The features chosen in the previous phase are used as input for the classification model at this stage. A fivefold cross-validation is performed, which means that 70% per cent of the total data is used during the training phase, but only 30 % is utilised during the testing phase. This dataset is then classified by the machine learning model used to diagnose gastric cancer. Within this part, the specifics of the machine learning model and the classification results are broken down in great depth[25].

E. Class Balancing

This imbalance is typical of disease-related datasets like the one used for this investigation, which focused on stomach cancer cases. In such cases, the larger class is referred to as the "majority class." In comparison, the smaller class is referred to as the "minority class." If the unbalanced dataset is employed, classifiers will lean toward the majority class, resulting in poor minority class classification performance. It is also possible that the classifiers will incorrectly assume that everyone belongs to the majority group and ignore the minority[26].The patient with the condition is often a demographic minority in medical databases. As a result, medical databases need a reliable sampling strategy. To address this problem, researchers have developed sample procedures that either remove data from the dominant group (undersampling) or add data to the underrepresented group (oversampling) using artificial means. Different sampling strategies, such as under-sampling, oversampling, and a hybrid of the two, have been developed to address the issue of class imbalance[27].A well-balanced data set is essential for developing a reliable prediction model from the training set. However, in the NHS gastric cancer dataset utilized for this analysis, the class labels of the target variable are not even. Especially when the data is very unbalanced (90.2% No and 9.8% Yes), this might lead to average performance from the classifiers on the Yes label, the minority class. This is because classifiers often optimize for overall accuracy rather than considering how each class is distributed individually[27].

An oversampling technique increases the proportion of individuals from the underrepresented group within the sample used for training. All the observations from the majority and minority classes are kept. Therefore, no information is lost from the original training dataset during oversampling. Since the training set size is dramatically increased, there is a risk of over-fitting and longer training times when using this method. To oversample the minority class, a popular oversampling method called Synthetic Minority Oversampling Technique (SMOTE) is implemented. SMOTE involves making fake instances of the minority classes that are the same as the real ones. This is to add more instances of the minority classes to the training set. The number of instances (n) and closest neighbours are used to generate these synthetic instances (k). Overfitting is avoided because new minority examples are generated by interpolating between nearby existing minority examples. Under-sampling, in which the number of samples from the majority class is reduced to equalize the class distribution between the minority and majority classes, is another method for dealing with the class imbalance issue[28]. The advantages of this method include less time spent training and better efficiency due to the drastically decreased size of the training dataset. The strategy has the potential drawback of losing valuable context in the training data. To achieve a more equitable class distribution, SpreadSubsample may reduce the number of samples from the majority class in the original dataset. Class values may be distributed from 0 to 9, depending on the spread value. A uniform distribution is achieved when the class distribution parameter is set to 1, with all class labels evenly spaced. Sometimes, a mix of oversampling and under-sampling is better because it clarifies the data space and stops people from making too many assumptions. The cancer target variable is underrepresented in the training dataset used for subsequent research. The values in the skewed data set are more heavily weighted toward the negative than the positive. In this data collection, there is a significant imbalance between the number of "no" and "yes" responses. There are "no" values (90.2%) and "yes" values (9.8%) in the 145,789 observations dataset. The results may lean toward the more common answer (No) because of the large discrepancy between the values in the class variable. This reduces the effectiveness of the outcomes and reflects doubt about whether machine learning algorithms produce the best model for making predictions. The training dataset is used without class balancing when doing classification tasks on the data. By applying SMOTE to the training dataset, we can oversample the minority class label and generate a new dataset with more evenly distributed classes. The data was resampled such that the minority class value (yes) had more occurrences. All instances of SMOTE were generated using WEKA's default settings, and its closest neighbor[29].The use of the SMOTE filter has increased the minority class value (yes) while leaving the majority unchanged. This strategy for racial equality has resulted in doubling the minority's share of wealth. It was found that there were 145,789 total cases in the dataset.

Under-sampling is the second way of class balancing. It was applied to the training dataset to construct a new dataset using this approach. This new training dataset was then created. This was accomplished via the SpreadSubsample function, in which the class distribution spread was set to 1.0 to facilitate a uniform distribution between the two class values (yes and no). As a direct consequence, the value of the majority class was reduced to equal the worth of the minority class. After making use of the SpreadSubsample filter, the number of occurrences in the class with the majority value of "no" (which is the same as the value held by the minority class) was found to be lower than before (which is yes). Within the training dataset, the class values "Yes" and "No" each have an equal proportion of distribution for the target variable "cancer," which is represented by the value "50%" accordingly[30].

Following this, the oversampling and under-sampling methods were merged to resample the unbalanced dataset. A training dataset using this approach was then constructed. To resample the distribution of the class values in the target variable, cancer, first, the oversampling approach known as SMOTE was used, and then the under-sampling method known as SpreadSubsample was employed. Because of this, the class that represented the minority, denoted by the yes value, was oversampled first, followed by the class that represented the majority, shown by the no value, which was under-sampled. In order to produce a result that is consistent when comparing these three ways of class balancing, the parameters used in the SMOTE and SpreadSubsample methods of class balancing, which came before, were also used in this approach[31].

F. Data Mining Techniques

To make timely decisions and uncover previously unknown facts, data miners use various techniques. The use of data mining methods allows for the discovery of previously unseen patterns within the data, which in turn aids data professionals in elucidating the interconnections between the data and allowing for more evidence-based and well-informed decision-making. Data mining techniques are becoming increasingly important in medical diagnosis, especially for predicting the likelihood of a patient surviving a cancer diagnosis. This is because they enable clinicians to make prompt decisions about the most effective treatment methods, early detection, and prediction of cancer and other diseases, increasing patient survival rates and decreasing treatment costs. In the healthcare industry, data mining methods such as classification, clustering, association, and regression are frequently used for diagnosis and illness prediction [32][33].

Classifying novel objects requires two steps: a training phase and a validation phase. In the first phase, a model is built using the training dataset to characterize a collection of data classes or concepts. This is an example of supervised learning since the categories into which the training sample fits have already been specified. Next, the model is implemented to make forecasts about the types of incoming data or objects. This method is commonly used in research aimed at the early detection and prognosis of cancer and is gaining popularity. Naive Bayes, Bayesian networks, decision trees, and association-based classification are only a few classification algorithms used for stomach cancer prediction research. When doing classification, the data is often split into two sets: the training and testing sets. During the training phase, the classifier is used to build a model, and the model's accuracy in making predictions or assigning labels to test data is verified in the testing phase[34].

Bayesian network A Bayesian network is a graphical model used to depict the probability correlations between the study's variables. The Bayesian model provides insight into the probability distribution represented by the network by assuming conditional independence over the numerous random variables. From a Bayesian perspective, the classification issue is the difficulty in determining the class with the highest probability given a collection of observed variable values. Using the available data, this probability is estimated using the Bayesian theorem. It is referred to as the class's posterior probability [35]. This classifier needs an extensive training set to properly explore all possible permutations and accurately estimate the training set's probability distribution. This is potentially time-consuming, which is a drawback of this data mining method. The general use of conditional independence is one of the Bayesian Network's main strengths, allowing for a compact and cost-effective representation of the joint probability distribution. Aside from that, the classifier is resistant to noise and other non-classification-related confounding effects.

Previous literature research has shown that Bayesian networks are widely used in numerous medical diagnostics, notably for cancer prediction. The usage of Bayesian network classifiers in stomach cancer prediction is rising [36]. Whenever the class variable and the attribute set do not have a deterministic connection, this classifier has proven helpful in medical diagnostics. Since K2 is a learning method for Bayesian networks, it has been used to categorize stomach cancer. Bayesian networks learn their structure from the data using search methods. K2 is a widely used heuristic algorithm in cancer classification that uses a greedy search strategy. It is one of several kinds of learning algorithms, including AD (All Dimensions) Trees and TAN (Tree Augmented Naive Bayes). The K2 method heuristically generates numerous different acyclic digraphs, and their data-interpreting prowess is evaluated based on these. Iterative permutations of the ordering are performed throughout the model-building process, with the network with the most significant probability being chosen. The Bayesian Network's conditional probabilities are calculated directly from the data using a Simple Estimator after the structure has been understood[37].

Random Forest is a tree-based approach that uses ensemble learning to produce predictions by aggregating the results of several classifiers that it develops and then using those aggregated results. These ensembles of classifiers employ a random tree generator to produce their tree-based components. The training data is randomly sampled. The resulting classification and regression trees (CART) might number hundreds or thousands. Random Forest is a machine learning method with a common ancestor with the CART approach. However, it distinguishes itself by its non-deterministic development through a two-level randomization mechanism. In order to identify the split at the node level, each tree is generated using a bootstrap sample of the training data, which is then explored through a randomly selected collection of features (input variables). The random feature selection increases the prediction power and efficiency since it decreases the correlation between the trees. Because of the bagging phenomenon, the forest ensemble has a minimal standard deviation[38].The Gini impurity measure is used as the splitting criteria in the Random Forest method, with the most negligible impurity value being calculated at each node for a given set of variables [39]. Random Forest's ability to offer a measure of variable importance—the extent to which a given feature is associated with the classification outcome—is a powerful tool for classifiers. The out-of-bag samples give an unbiased test-set error estimate and a variable importance measure, which may be used to evaluate the bootstrap-derived trees. Random Forest has been the go-to approach for classification problems like those used in stomach cancer prediction research because of its numerous advantages, list the following benefits of this method:

- High-dimensional data with missing values and continuous, binary, and categorical variables are no problem.
- It is resilient enough to prevent data over-fitting; hence, it does not call for tree-shaking pre-processing.
- A straightforward non-parametric technique works well. It is easy to understand and can be used to analyze a wide variety of datasets.
- It is more generalizable and produces predictions that are more accurate.

Decision tree A decision tree is a supervised strategy that employs the reasoning approach to discover answers for a given issue. This approach helps ensure that the best possible decisions are made. This data mining technique is very flexible and straightforward, making it an appealing option for applications in various domains. This is especially true given that it uses advice-oriented visualization to allow users to make prediction decisions based on the outcomes that have been observed. It is standard practice in medicine to use a decision tree as part of the decision-making process when diagnosing diseases or generating predictions about cancer. In a decision tree, the tree-shaped structures represent decision sets that are straightforward to interpret and comprehend, allowing decision-makers to evaluate and select the most appropriate course of action based on the risks and benefits associated with each possible outcome for various options. Following is a list of the components that make up the fundamental framework of a decision tree[3].

- Internal nodes, in which each node contains one incoming branch and two or more outgoing branches.
- A root node does not have any incoming branches but does include zero or more outward branches as part of its structure.
- Leaf or terminal nodes, in which each node comprises one incoming branch but no outgoing branches.

Each node in the decision tree represents an attribute in the input attribute space. Each branch indicates a condition value for the node it corresponds to. The non-terminal nodes are equipped with attribute test criteria, which are used to categorize the records by their differentiating features, shown by the branches[22].

The C4.5 algorithms are an extension of the ID3 algorithms. They are one of the prominent classification types used in the decision tree. These algorithms are used to predict and detect stomach cancer. C4.5 uses the notion of information entropy to create decision trees, beginning with a set of predefined training data and moving forward from there. This strategy uses the fact that every variable in the data is relevant to the decision-making process by dividing the records into a more significant number of smaller groups. C4.5 determines the process of selecting an attribute to divide the data using the normalized information gain, also known as the difference in entropy. Suppose no information is gained from any of the characteristics. In that case, the C4.5 will build a decision node based on the class expected value from nodes higher up the tree[4]. The decision node is determined by selecting the property that significantly increases normalized information gain. In the decision tree, the node that corresponds to the branch with an entropy value of 0 is the leaf node. This method is executed recursively on subsets that are not leaf nodes and have an entropy value that is not zero. When all of the samples in a particular subset or node belong to the same category, the process of splitting will end. After that, a lead node is created to facilitate the class selection. The C4.5 algorithm has a few benefits, including the fact that it is straightforward to construct in a comprehensible format, that it can be applied to data that has discrete and continuous attributes, that it can handle attributes that have missing values and differing costs in the training data, and that it has greater precision because of the pruning procedure. The large amount of processing time and the exorbitant costs incurred are two drawbacks of the C4.5 classifier.

IV. Result and Discussion

The WEKA software, part of the Waikato Environment for Knowledge Analysis, was used to experiment[40]. The model was validated on the class-imbalanced and class-balanced training sets using k-fold (10 folds) cross-validation. In this study, the recommended 10-fold cross-validation method was used to check the accuracy of the classifier model made from the training dataset. This method was used to diagnose and predict stomach cancer. The 145,789 instances that comprise the training dataset were subjected to class-balancing techniques. Using these techniques, we addressed the problem of class imbalance in the cancer-focused dependent variable. Multiple balancing strategies were used, including SMOTE (for oversampling), SpreadSubsample (for under-sampling), a combination of SMOTE and SpreadSubsample, and a spread of 1.0 for the distribution. The best prediction model for the stomach cancer dataset was determined after comparing the performance of several classifiers based on their respective sampling strategies. Accuracy, ROC, PRC Area, FP Rate, Specificity, Precision, Recall, and F-measure were only a few assessment metrics used to evaluate the classifiers' results. Due to the nature of the medical data included in this study's breast cancer dataset, specific assessment metrics are crucial for gauging the efficacy of the algorithm-based prediction model. These metrics include accuracy, TP rate (or sensitivity or recall), FP rate, precision, ROC area, and PRC area.

A prediction model must accurately detect the existence of a disease without room for error. If a model can successfully predict which patients will get stomach cancer, the TP rate will be high. Although a higher TN rate is desired, it is not given the same weight as the TP rate. In an illness like cancer, where an early diagnosis significantly improves the prognosis, the false negative and false favourable rates may be deadly. The classifier's efficiency improves as the false positive and false negative rates decrease and the true favourable and accurate negative rates increase. The dataset and classifiers may alter these general requirements for a disease prediction model. Two sample approaches were shown to provide superior measurements for evaluating the performance of the classifiers, and this was true across all four classifiers. The original training set is used in these techniques, and SMOTE and SpreadSubsample are applied to the training set in conjunction with one another. When used together, it was shown that SMOTE and SpreadSubsample outperform balance within each classifier on several metrics. When comparing the accuracy, sensitivity, and precision of the Bayesian Network, Random Forest, and Decision Tree C4.5 models, the latter's technique of combination yields better results. When the FP rates of the four classifiers we have discussed so far are compared, the hybrid approach has a lower FP rate, making it better for diagnosing illnesses.

The overall conclusion is that even though all four of these classifiers perform effectively without a class balancing mechanism, the results are likely to be skewed due to the uneven distribution of class value. There will be a slant toward the No Cancer Class value regarding accuracy, sensitivity, and specificity. A validation (test) dataset may be employed to verify these four classifiers, and a better model can be obtained by combining SMOTE with the SpreadSubsample approach. This hybrid balancing approach evenly distributes the Yes and No cancer class values, which will assist in the creation of a balanced prediction model.

With the NHS dataset in hand, we investigated additional performance parameters used to assess the classifiers to find the best classification model for the prediction of gastric cancer. Classifiers built on the training set were validated using the validation dataset. The results showed that the two sets of classifiers were quite close in assessment metrics. This demonstrates that all classifiers perform well when evaluated on the test set. However, to choose the best or most robust classifier among the four suggested classifiers, we evaluated them using several standard assessment measures used in medical diagnosis.

The Bayesian Network classifiers in Table II have a 99.31% accuracy rate. The lowest accuracy was achieved by Random Forest, at 93.8%. The Bayesian network has been proven to have the lowest FP rate for the Yes class label, with 0.0011% of FP being predicted. In order to prevent the unnecessary suffering of patients whom a correct diagnosis of stomach cancer may have saved, it is crucial to achieve the lowest possible FP rate. The Bayesian network was the most accurate classifier when comparing the class labels and the weighted average. The Bayesian Network has the most significant reported sensitivity (99.1% on average) and ROC (93.6% on average) of all classifiers. In conclusion, our research employing NHS data on stomach cancer demonstrates that a Bayesian network may be used as the prediction model. The Bayesian Network was chosen as the optimal classification model for this research due to its proven track record of success as a prediction model for cancer studies and its widespread usage in cancer diagnostics. Because it can be represented graphically, the Bayesian network model is also more accessible for the human mind to grasp. Table III compares the assessment metrics used to assess the performance of the models used in this research with those used in earlier publications on the same NHS dataset, providing additional evidence that the Bayesian Network model used here provides a superior prediction.

The Bayesian network model offers the best ROC and accuracy compared to other models. Therefore, the Bayesian Network is a more accurate prediction model for categorising breast cancer incidence based on associated risk variables. Aside from that, class balancing methods were used, which was different from the case in earlier research. None of these other studies used the variety of strategies used in this one to address the class imbalance problem in the NHS. This research has shown that a superior prediction model could be obtained by combining the Bayesian network with a hybrid balancing strategy.

Table II. MEASURES OF PERFORMANCE VALUATION PRESENTED

Classifier	**Class label**	**Performance Evaluation Metrics**				
		Accuracy	**FP Rate**	**Precision**	**Sensitivity or Recall**	**ROC**
Decision Tree C4.5 (DT)	Yes	0.973	0.205	0.891	0.921	0.906
	No		0.006	0.833	0.686	0.814
	Weighted average		0.204	0.975	0.976	0.912
Bayesian Network	Yes	0.993	0.217	0.991	1.011	0.935
	No		0.001	1.010	0.792	0.936
	Weighted average		0.211	0.990	0.990	0.936
Random Forest (RF)	Yes	0.938	0.196	0.971	0.944	0.911
	No		0.044	0.432	0.793	0.914
	Weighted average		0.189	0.959	0.953	0.914
Naïve Bayes (NB)	Yes	0.948	0.208	0.992	1.010	0.936
	No		0.001	1.000	0.791	0.935
	Weighted average		0.212	0.992	0.992	0.936

Table III. CONTRAST TO PREVIOUS STUDIES

Previous literature	Predictive model	Evaluation measure	Scope of study
[40]	Fuzzy cognitive maps (FCM)	95.7	This work proposes a synthetic sampling technique to balance dataset along with modified particle swarm optimization (m-PSO technique
[14]	Fuzzy cognitive maps (FCM)	95.83%	The FCM-based model is comprehensive, transparent, and more effective than previous models for assessing the risk of GC
[14]	Association rule mining with SVM	Accuracy = 98%	An association rule model with feature selection on the dataset
Our System	Bayesian network	Accuracy 0.993 ROC= 0.936	Determine GC on the basics of critical risk factor associated with

V. CONCLUSION

This research was carried out using a dataset from the NHS that included 145,789 different stomach cancer patients. SMOTE, SpreadSubsample, and a mixture of SMOTE and SpreadSubsample were the three strategies used to achieve class balance in the training dataset, which allowed for the problem of class imbalance to be resolved. The Bayesian Network, Random Forest, and Decision Tree C4.5 classification models were all constructed using these approaches. When the different sampling strategies were compared across each classifier using the performance assessment metrics, the findings showed that the classifiers created by employing the hybrid balancing method had the best performance regarding the false positive rate and the area under the ROC. Because of this, it was concluded that the method of class balancing most appropriate for the BCSC dataset was a hybrid strategy, which was statistically demonstrated to perform well in comparison to other sampling strategies. Based on the findings, the Bayesian network that was produced from the class-balanced NHS data by applying the hybrid technique had superior overall performance in terms of ROC (0.937), sensitivity (78.1%), and false-positive rate (0%). It also had 100% specificity. By forecasting the incidence of stomach cancer based on the risk variables, this research demonstrates that the Bayesian Network model may serve as a better decision support system for doctors and a method for early detection and treatment for patients. In conclusion, the findings of this research showed that the hybrid balancing approach combined with the Bayesian Network algorithm was the one that produced the highest level of accuracy in predicting the likelihood of Developing GC, given a set of risk variables. Patients with gastric cancer can learn more about the disease and what puts them at risk because of this method. It also helps doctors make decisions about diagnosis and treatment that are objective and backed up by statistics. Future work might entail selecting features from the NHS data set and then segmenting the variables into groups based on their similarities. A predictive model thatwas developed using feature selection and variables that are similar to one another might produce a generalised model.This would reduce the number of risk factors that needed to be diagnosed. It is not guaranteed that the findings of this study could be generalised to other GC datasets with different properties; therefore, it would also be interesting to apply this methodology to other data with features such as shape, location, tumour size, or radiation intensity. This is because it is not guaranteed that the findings of this study could be generalised.

ACKNOWLEDGMENTS

We would like to express our sincere appreciation to the NHS Liverpool hospital in England for making the dataset available to us to enable us to conduct this study. We would like to extend our sincere gratitude to Dr. Shakil Ahmed, a consultant surgeon at the Royal Liverpool hospital.

REFERENCES

[1] D. Jamil, “Diagnosis of Gastric Cancer Using Machine Learning Techniques in Healthcare Sector: A Survey,” *Informatica*, vol. 45, 2022, doi: 10.31449/inf.v45i7.3633.

[2] L. Goshayeshi *et al.*, “Predictive model for survival in patients with gastric cancer,” *Electron. physician*, vol. 9, no. 12, p. 6035, 2017.

[3] 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, doi: 10.31557/APJCP.2019.20.9.2607.

[4] M. S. Mohammad Reza Afrash and and H. Kazemi-Arpanahi, "Design and Development of an Intelligent System for Predicting 5-Year Survival in Gastric Cancer," *Clin. Med. Insights Oncol.*, vol. 16, no. 1, pp. 1–13, 2022, doi: DOI: 10.1177/11795549221116833.

[5] S. Bagchi, K. G. Tay, A. Huong, and S. K. Debnath, "Image processing and machine learning techniques used in computer-aided detection system for mammogram screening-A review," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 3, p. 2336, 2020.

[6] S. S. Z. Danish Jamil, Sellappan Palaniappan, Asiah Lokman, Muhammad Naseem, "Diagnosis of Gastric Cancer Using Machine Learning Techniques in Healthcare Sector: A Survey," *Informatica*, 2022.

[7] 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.

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

[9] A. Mahani and A. R. B. Ali, "Classification problem in imbalanced datasets," *Recent Trends Comput. Intell.*, pp. 1–23, 2019.

[10] S. S. ZIA, P. AKHTAR, and T. J. A. MUGHAL, "Case Retrieval Process of CBR Technique Implements on Knowledge-Based Clinical Decision Support Systems (KBCDSS) for Diagnosis of Breast Cancer Disease," *Sindh Univ. Res. Journal-SURJ (Science Ser.*, vol. 47, no. 2, 2015.

[11] C. Mazo, C. Aura, A. Rahman, W. M. Gallagher, and C. Mooney, "Application of Artificial Intelligence Techniques to Predict Risk of Recurrence of Breast Cancer: A Systematic Review," *J. Pers. Med.*, vol. 12, no. 9, p. 1496, 2022.

[12] J. Yoon, J. Lee, S. Park, W. J. Hyung, and M.-K. Choi, "Semi-supervised learning for instrument detection with a class imbalanced dataset," in *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, Springer, 2020, pp. 266–276.

[13] P. Melek Akcay, MD, Durmus Etiz, MD, and Ozer Celik, "Prediction of Survival and Recurrence Patterns by Machine Learning in Gastric Cancer Cases Undergoing Radiation Therapy and Chemotherapy," *Adv. Radiat. Oncol.*

[14] 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.

[15] M. A. Mohammed *et al.*, "Retraction Note: Decision support system for nasopharyngeal carcinoma discrimination from endoscopic images using artificial neural network," *J. Supercomput.*, pp. 1–2, 2022.

[16] K. Rajendran, M. Jayabalan, and V. Thiruchelvam, "Predicting breast cancer via supervised machine learning methods on class imbalanced data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, 2020.

[17] S. A. Mahmoodi, K. Mirzaie, and S. M. Mahmoudi, "A new algorithm to extract hidden rules of gastric cancer data based on ontology," *Springerplus*, vol. 5, no. 1, p. 312, 2016.

[18] P. Sahu, P. K. Sarangi, S. K. Mohapatra, and B. K. Sahoo, "Detection and Classification of Encephalon Tumor Using Extreme Learning Machine Learning Algorithm Based on Deep Learning Method," in *Biologically Inspired Techniques in Many Criteria Decision Making*, Springer, 2022, pp. 285–295.

[19] J. Yuan, Q. Wang, Z. Li, C. Dong, P. Zhang, and X. Ding, "Domain-knowledge-oriented data pre-processing and machine learning of corrosion-resistant $γ$-U alloys with a small database," *Comput. Mater. Sci.*, vol. 194, p. 110472, 2021.

[20] V. et al Lysaght, T., Lim, H.Y., Xafis, "AI-Assisted Decision-making in Healthcare," *Asian Bioeth. Rev.*, no. 11, pp. 299–314, 2019, doi: https://doi.org/10.1007/s41649-019-00096-0.

[21] K. J. Cios, B. Krawczyk, J. Cios, and K. J. Staley, "Uniqueness of Medical Data Mining: How the new technologies and data they generate are transforming medicine," *arXiv Prepr. arXiv1905.09203*, 2019.

[22] L. M. Terracciano *et al.*, "Opportunities and Challenges for Machine Learning in Rare Diseases," *Front. Med. | www.frontiersin.org*, vol. 8, p. 747612, 2021, doi: 10.3389/fmed.2021.747612.

[23] W. Albattah, R. U. Khan, M. F. Alsharekh, and S. F. Khasawneh, "Feature Selection Techniques for Big Data Analytics," *Electronics*, vol. 11, no. 19, p. 3177, 2022.

[24] J. Yang, J. Zhou, Z. Zhu, X. Ma, and Z. Ji, "Iterative ensemble feature selection for multiclass classification of imbalanced microarray data," *J. Biol. Res.*, vol. 23, no. 1, pp. 1–9, 2016.

[25] R. Chauhan, R. Jangade, and R. Rekapally, "Classification model for prediction of heart disease," in *Soft Computing: Theories and Applications*, Springer, 2018, pp. 707–714.

[26] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *J. Big Data*, vol. 5, no. 1, pp. 1–30, 2018.

[27] H. et al. Iqbal, M.J., Javed, Z., Sadia, "Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future," *Cancer Cell Int 21*, vol. 270, 2021.

[28] S. Sharma, A. Gosain, and S. Jain, "A Review of the Oversampling Techniques in Class Imbalance Problem," in *International Conference on Innovative Computing and Communications*, 2022, pp. 459–472.

[29] N. V Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *European conference on principles of data mining and knowledge discovery*, 2003, pp. 107–119.

[30] J. Zhang, L. Chen, and F. Abid, "Prediction of breast cancer from imbalance respect using cluster-based undersampling method," *J. Healthc. Eng.*, vol. 2019, 2019.

[31] Z. Z. R. Al-Shamaa, S. Kurnaz, A. D. Duru, N. Peppa, A. H. Mirnezami, and Z. Z. R. Hamady, "The use of hellinger distance undersampling model to improve the classification of disease class in imbalanced medical datasets," *Appl. Bionics Biomech.*, vol. 2020, 2020.

[32] R. Raja, K. K. Nagwanshi, S. Kumar, and K. R. Laxmi, *Data Mining and Machine Learning Applications*. John Wiley \& Sons, 2022.

[33] S. K. Mohapatra, R. K. Kanna, G. Arora, P. K. Sarangi, J. Mohanty, and P. Sahu, "Systematic Stress Detection in CNN Application," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, 2022, pp. 1–4.

[34] D.-C. Li, C.-W. Liu, and S. C. Hu, "A learning method for the class imbalance problem with medical data sets," *Comput. Biol. Med.*, vol. 40, no. 5, pp. 509–518, 2010.

[35] C.-C. Kuo, H.-H. Wang, and L.-P. Tseng, "Using data mining technology to predict medication-taking behaviour in women with breast cancer: A retrospective study," *Nurs. Open*, vol. 9,

no. 6, pp. 2646–2656, 2022.

[36] A. Sheidaei, A. R. Foroushani, K. Gohari, and H. Zeraati, "A novel dynamic Bayesian network approach for data mining and survival data analysis," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 1–15, 2022.

[37] 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, p. 5408, 2020.

[38] R. Hasan, S. Palaniappan, A. R. A. Raziff, S. Mahmood, and K. U. Sarker, "Student Academic Performance Prediction by using Decision Tree Algorithm," in *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, Aug. 2018, pp. 1–5, doi: 10.1109/ICCOINS.2018.8510600.

[39] Q. Gu, J. Tian, X. Li, and S. Jiang, "A novel Random Forest integrated model for imbalanced data classification problem," *Knowledge-Based Syst.*, p. 109050, 2022.

[40] M. Das and R. Dash, "A Comparative Study on Performance of Classification Algorithms for Breast Cancer Data Set Using WEKA Tool," in *Intelligent Systems*, Springer, 2022, pp. 289–297.

Danish Jamil is currently a PhD scholar in the Department of Information Technology, at Malaysia University of Science and Technology, Petaling Jaya, Malaysia. He is also working as a Senior Lecturer in the Department of Software Engineering at Sir Syed University of Engineering and Technology, Karachi, Pakistan. His area of expertise is Data mining, Machine learning, Health Informatics, Clinical Informatics, and Deep Learning. He is the author of various articles and publications pertaining to his field of study.