Comparative Evaluation of SMOTE Algorithms in Predictive Modelling of Cirrhosis

MSc Research Project
MSc. Data Analytics

Anjana Sasanka Wedage
Student ID: X23131934

School of Computing
National College of Ireland

Supervisor: Dr. Bharat Agarwal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Anjana Sasanka Wedage
Student ID:	X23131934
Programme:	MSc. Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr. Bharat Agarwal
Submission Due Date:	12/08/2024
Project Title:	Comparative Evaluation of SMOTE Algorithms in Predictive Modelling of Cirrhosis
Word Count:	2855
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	14th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparative Evaluation of SMOTE Algorithms in Predictive Modelling of Cirrhosis

Anjana Sasanka Wedage
X23131934

Abstract

Cirrhosis has become a major medical issue in the world and people suffer from it in every corner of the world. The medical world is working on many treatments and procedures to save people from loosing lives. In the modern world, machine learning has given hope by developing models to detect the disease early. This paper focuses on using Synthetic Minority Over-sampling Technique (SMOTE) to improve predictions from the machine learning models. Mostly medical data has imbalanced data in the datasets, and it affects the prediction ability and SMOTE helps to get an even balanced dataset for the analysis. The study focuses on 9 SMOTE techniques and evaluated using 5 machine learning models. The results showed that Adaptive SMOTE technique showed better performance than other SMOTE techniques and specifically Adaptive SMOTE with K-Nearest Neighbors showed the better metrics accuracy and the recall of the stages combined.

1 Introduction

In recent years, the integration of machine learning in medical research has become more advanced and delivered some promising results. Using the large amount of data available, the advanced algorithms and the computational power, the complex problems in the medical domain are being tackled everyday. These solutions and answers helps to diagnose diseases and enables quality life and healthcare. Diagnosis is important in medical field as it can be life-threatening for patients. Research by Graber et al. (2005) shows that the diagnostic errors leads to patient harm and that can be a leading cause to malpractice claims. This can happen by the errors of classification of data, which can be fatal. This explains why the errors needs to be reduced and the accuracy of the results are important for the lives of the humans and financially for the institutes to avoid lawsuits.

Liver disease and cirrhosis represent significant health challenges globally, affecting millions of people and posing substantial burdens on healthcare systems.Schuppan and Afdhal (2008)When the injured tissue in your liver is replaced by a scar tissue, it is called fibrosis. This healing journey in the liver can be abnormal and can lead to excessive fibrogenesis and can be in various stages. Cirrhosis is the advanced level of liver fibrosis and studies shows that even though earlier that was shown that Cirrhosis cannot be reversed now it is possible to be reversed or regressed. This paper is focused on Cirrhosis identification, and it is important since early detection can save lives and reverse the effect. Figure 1 shows the overview of how the medical research and the data analysis moves together to improve the medical data analysis and the idea behind this study that is presented on this paper.

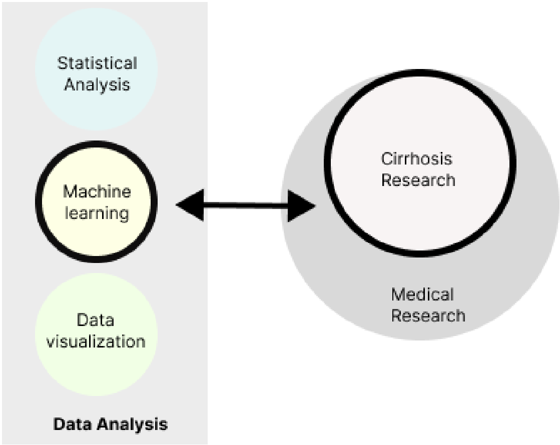

Figure 1: Overview of Data Analysis Techniques and Their Application in Medical Research

As mentioned above, classification of the patients are really important. There are various models that are developed to predict different classifications such as stages of the diseases, whether the patients going to survive or not, what type of medicine that needs to be prescribed etc. But one of the major problems that analysts face is the imbalanced datasets. Mostly medical data has imbalanced data which leads to inaccurate classifications when models been developed and run.

There are a good bit of research done to tackle the issues caused by the imbalanced data. Araf et al. (2024a) study compares 173 papers online in five libraries to analyse and compare the techniques used to handle imbalanced data. This further shows the techniques such as resampling, ensemble learning and cost-effective learning are used to handle the data and the metrics that are used to analyse the effect of them. Finally, it discusses the importance of comprehensive studies to understand the applications and the models that are used to handle the imbalanced data.

Study done by Roy et al. (2024) shows that SMOTE (Synthetic Minority Oversampling Technique) as one of the main techniques that can be used to handle the imbalanced data as a data level approach. It further explains that using data level approach such as SMOTE can be a better option because its independent of the classifiers used, and it doesn't modify the underlying algorithm and shows the applications of SMOTE in medical research.

Fernández et al. (2018)Last 20 plus years there were studies done to extend the SMOTE technologies to different levels where some of the techniques focuses on oversampling, under sampling, different insertions done to generate the synthetic samples, changing the dimensionality, relabelling and filtering the noisy instances are some theories behind the extensions oof SMOTE. Below are the other extensions of SMOTE techniques that were considered for the study, out of 85 more SMOTE extensions that are available online.

Even though there are studies done on individual SMOTE techniques, there are not many studies done to show the comparison of the SMOTE techniques and which techniques perform better than the others. This study is to focus on the SMOTE techniques available and pick a few that are relevant to handling medical data and see the effectiveness of those techniques. The question that is addressed in this paper is **What is the**

comparative impact of various Synthetic Minority Oversampling Technique (SMOTE) algorithms on the predictive performance of cirrhosis prediction models using machine learning techniques? The dataset that is used in the study is the Cirrhosis dataset that is available on Kaggle, and the output of the study is to see how the performances of Cirrhosis prediction models changes with different SMOTE techniques.

2 Related Work

Bowyer et al. (2011)In 2000 Chawla faced a classification problem where the non-majority classes of the problem were cancerous pixels and the basic decision tree he analysed got 97% accuracy, which led to him exploring the SMOTE technique. The method was inspired by handwritten character recognition, and SMOTE created synthetic examples of data rather than making samples of the data that is already in the dataset.

SMOTE does not consider the specific distribution and there is a risk of overlapping classes which can cause poor classification performance, particularly near the decision boundary. To avoid this issue, Borderline SMOTE was developed.

Han et al. (2005) Borderline SMOTE focuses on the instances of the minority class that are near the borderline of the classes. These borderline instances are considered more critical for defining the decision boundary and improving classification performance. Borderline SMOTE generates synthetic samples for the minority class, but only near the borderline instances, thereby creating a more effective separation between classes. Furthermore, it shows that the true positive rates and F values are better compared to basic SMOTE.

Developing the basic SMOTE there are different other techniques that have been introduced. Safe-Level-SMOTE is one of the techniques that is discussed in this paper. According to the study by Bunkhumpornpat et al. (2009) this technique improves upon SMOTE by carefully sampling minority instances with different weights, based on a computed ”safe level” using the nearest minority neighbours. By developing synthetic instances more around higher safe levels, this technique achieves better accuracy compared to SMOTE and Borderline-SMOTE.Syakiylla Sayed Daud and Sudirman (2023) shows Safe level SMOTE used in the medical domain, where it is used to balance EEG data that is used to detect anxiety disorders. The research shows that it is significantly improve the accuracy of the predictions.

SMOTE Tomek is one of the other techniques that was developed recently as a variant of SMOTE. In Liu et al. (2018) SMOTE Tomek method has been used to improve the sample imbalance and in the research it shows that it works better in evaluation metrics than using SMOTE itself.

ADASYN Adaptive synthetic sampling approach is another method that can be used as a development of SMOTE techniques. He et al. (2008) introduced the method in its research and the idea of ADASYN technique is to use a weighted distribution for every minority class examples depending on the learning difficulty level, where more synthetic data is developed for the minority class examples that are harder to learn.

SVM-SMOTE is another technique that has been used for resampling using border based SMOTE. This enhances the SMOTE by integrating SVM (Support Vector

Machines) into the SMOTE. This focuses on the data near the decision boundary and generate synthetic samples where the classifier is most likely to benefit and increased the ability of the classifier to identify the minority class correctly. Miftahushudur et al. (2023) shows that the SVM-SMOTE has the better MCC (Matthews Correlation Coefficient) Araf et al. (2024b) than the other SMOTE techniques SMOTE, Borderline SMOTE that are compared in the research.

Xu et al. (2021) compares SMOTE, Borderline-SMOTE, ADASYN-SMOTE, ANS-SMOTE Siriseriwan and Sinapiromsaran (2017), MDO-SMOTE, Gaussian-SMOTE (GSM), SOMO, SOI_CJ (SOI), MWMOTE (MWM), K-means-SMOTE (KSM) and KNSMOTE. KNSMOTE is a method used combining K-nearest neighbours and SMOTE used in medical data and compared against other SMOTE techniques using Random forest model shows that KNS shows better performance than other techniques compared. This research has used multiple datasets and multiple SMOTE techniques to analyse the effect of the techniques and run through Random forest model.

The key finding of the literature review is the research done by Sharma and Gosain (2023). This research has used a similar approach to this paper. The research reviews five oversampling methods to address class imbalance: SMOTE, Safe Level SMOTE, SMOTE Tomek Links, Borderline SMOTE1, and Adaptive SMOTE (ADASYN). The performance is then evaluated by Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and K Nearest Neighbour (KNN). The output suggested that in most models SMOTE Tomek Link gives the better results for the evaluation metrics such as accuracy, precision, recall, F1 score, and area under the curve (AUC).

The target of this paper is to extend the study done by Sharma and Gosain (2023) and use more commonly used SMOTE techniques in the medical and other domains and compare the performance against fewer more models than used in the previous research. The focus is to analyse and find out that even when the SMOTE techniques are increased and compared against more models will the findings of the previous researchSharma and Gosain (2023) will stay remain or will it change.

3 Methodology

The Predictive capability of the models are affected by the imbalanced data as shown and discussed above. For that specific reason this study fouses on reducing that error by incorporating SMOTE techniques to balanced the datasets. Figure 2 There will be 8 SMOTE techniques used in this study and the data will be compared using 6 machine learning models.

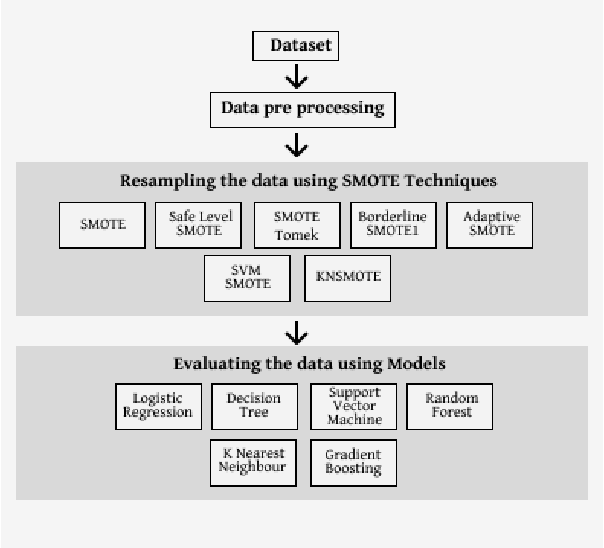

Figure 2: Flow of data

3.1 Data structure and Data preprocessing

Data used for this study was retrieved by Kaggle and the dataset focuses on Cirrhosis patients that were referred to the Mayo Clinic and there are 424 patient recorded in the dataset. This was a dataset focused on a randomized placebo controlled trial of Penicillamine drug.

Dataset has 20 variables that were collected.

Variable Name	Description	Variable Type
ID	Unique identifier	Categorical (Identifier)
N_Days	Number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986	Numeric (Integer)
Status	Status of the patient: C (censored), CL (censored due to liver tx), or D (death)	Categorical (Nominal)
Drug	Type of drug: D-penicillamine or placebo	Categorical (Nominal)
Age	Age in days	Numeric (Integer)
Sex	Gender of the patient: M (male) or F (female)	Categorical (Nominal)
Ascites	Presence of ascites: N (No) or Y (Yes)	Categorical (Nominal)
Hepatomegaly	Presence of hepatomegaly: N (No) or Y (Yes)	Categorical (Nominal)
Spiders	Presence of spiders: N (No) or Y (Yes)	Categorical (Nominal)
Edema	Presence of edema: N (no edema and no diuretic therapy for edema), S (edema present without diuretics, or edema resolved by diuretics), or Y (edema despite diuretic therapy)	Categorical (Ordinal)
Bilirubin	Serum bilirubin in mg/dl	Numeric (Float)
Cholesterol	Serum cholesterol in mg/dl	Numeric (Float)
Albumin	Albumin in gm/dl	Numeric (Float)
Copper	Urine copper in ug/day	Numeric (Float)
Alk_Phos	Alkaline phosphatase in U/liter	Numeric (Float)
SGOT	SGOT in U/ml	Numeric (Float)
Triglycerides	Triglycerides in mg/dl	Numeric (Float)
Platelets	Platelets per cubic ml/1000	Numeric (Integer)
Prothrombin	Prothrombin time in seconds	Numeric (Float)
Stage	Histologic stage of disease (1, 2, 3, or 4)	Categorical (Ordinal)

Table 2: Summary of Variables

Dataset showed missing values present in some of the variables. The figure shows how the missing values are distributed among the variables present. Figure 3

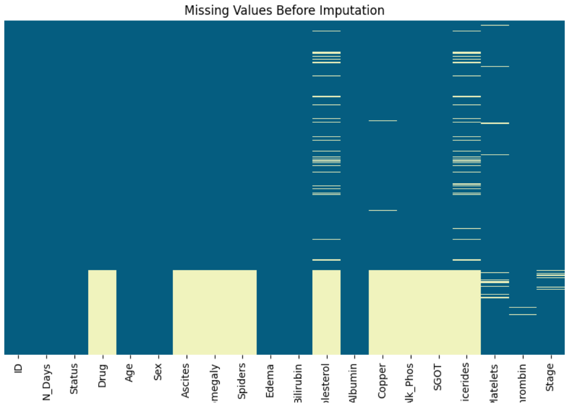

Figure 3: Missing Values Before Imputation

For the missing values few methods were suggested in the initial review of literature and one of the main techniques suggested was the IterativeImputer library from scikit-learn Kawashima et al. (2024) in python. It has showed some potential in predicting missing values Using Bayesian Ridge regression. But since it mostly works in numerical data, the missing values were predicted only for numerical variables. After using IterativeImputer the data distribution shows as follows Figure 4

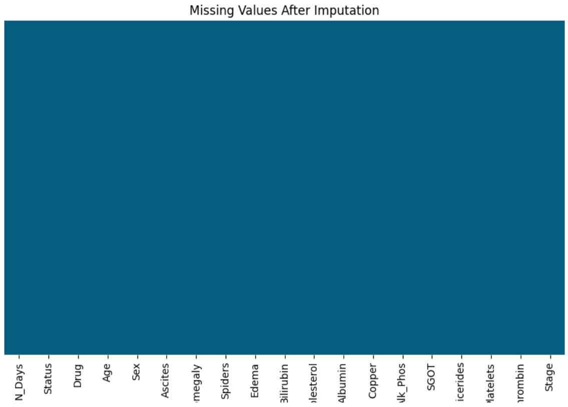

Figure 4: Missing Values After Imputation

After imputation, there are still missing values in the categorical variables. Therefore, logistic regression method was used to fill the missing values for the categorical variables. The approach was to train the logistic model for each categorical variable. Non-null records were used to train the model and predicted the missing values for each variable. This was run for each categorical variable that was missing data. The model was run only using the numerical variables for each categorical variable that was missing data as the dependent variable.

The target variable of the study is the Stage of the Cirrhosis patients. There are four stages presented in the dataset. The four stages are

1. Steatosis (inflammation of the bile duct or liver.)
2. Liver scarring (fibrosis) due to inflammation
3. Cirrhosis
4. Liver failure or advanced liver disease or hepatic failure

The counts in each stage of the data is as follows

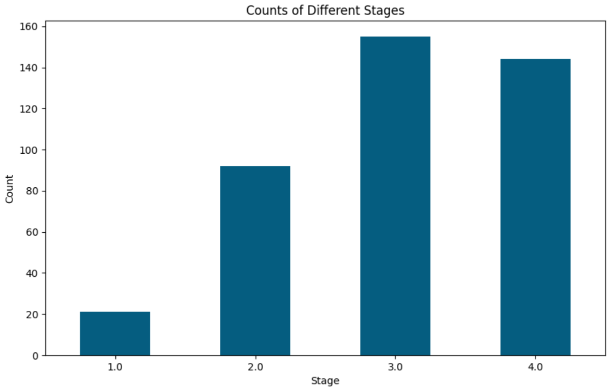

Figure 5: Record counts for each Stage

There is clearly an imbalance of data in the dataset. Therefore, as proposed, SMOTE techniques were used to balance the data for better accuracy.

3.2 SMOTE techniques

3.2.1 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) gives a solution to the class imbalance issue by generating synthetic samples of the minority class. The technique works by picking a random minority class sample (x) and then selecting the k-nearest neighbours (x') of the selected minority class. Then the new sample is created by using the selected sample and the nearest neighbour. Mathematically, it's created by adding a potion of the difference between the sample and the k-nearest neighbour to the sample. Mathematically, the new sample is generated as:

$$\text{New sample} = x + \delta \times (x' - x)$$

where δ is a random number between 0 and 1. This method ensures that the new samples are distributed in the feature space, closely representing the distribution of the minority class, and helping to develop and balance the dataset.

3.2.2 Safe Level SMOTE

Safe Level Smote is a different approach of original SMOTE created to answer some of the issues it had and to improve the technique. This is specially when working with borderline or noisy samples. The theory behind the borderline sample is that generating synthetic samples considering the safety of the original samples that were used. This technique avoids creating minority samples in regions where the minority samples are not well represented or are surrounded by majority class samples. This can lead to overlapping and reduce classification performance.

3.2.3 SMOTE Tomek

SMOTE-Tomek is a hybrid approach that is used by combining the SMOTE technique and the data cleaning capabilities of Tomek. This method has proven in some studies to outperform others SMOTE techniques because of its ability o enhance the quality of training data. First SMOTE is used to created synthetic samples of the minority class and increase the number of minority class samples. Then Tomek links are identified in the dataset. This basically involves finding the pairs of instances from different classes that are the neighbours of each other. Once it identifies, Tomek links are removed from the dataset. This helps to clean the data by eliminating the borderline or noisy samples of data, which leads to clear decision boundaries.

3.2.4 Borderline-SMOTE1

SMOTE does not consider the specific distribution and there is a risk of overlapping classes which can cause poor classification performance, particularly near the decision boundary. To avoid this issue, Borderline SMOTE was developed.

Han et al. (2005) Borderline SMOTE focuses on the instances of the minority class that are near the borderline of the classes. These borderline instances are considered more critical for defining the decision boundary and improving classification performance. Borderline SMOTE generates synthetic samples for the minority class, but only near the borderline instances, thereby creating a more effective separation between classes. Furthermore, it shows that the true positive rates and F values are better compared to basic SMOTE.

3.2.5 Adaptive SMOTE

Adaptive SV-Borderline SMOTE is an advanced variation of the Synthetic Minority Oversampling Technique (SMOTE) designed to handle imbalanced datasets more effectively. It integrates concepts from Support Vector Machines (SVM), adaptive synthetic sampling, and Borderline SMOTE to address the limitations of previous methods. This technique aims to generate synthetic samples near the decision boundary identified by SVM while adapting the sampling strategy based on local density and the difficulty of classification.

3.2.6 SMOTE-RkNN

SMOTE-RkNN (Synthetic Minority Over-sampling Technique using Reverse k-Nearest Neighbors) is an innovative method designed to address the limitations of traditional SMOTE and its variants in handling imbalanced datasets. By incorporating the concept of reverse k-nearest neighbors, SMOTE-RkNN focuses on generating synthetic samples for the minority class with a consideration of the distribution and density of the majority class, thereby enhancing the classification performance.

3.2.7 SVM SMOTE

SVM SMOTE (Support Vector Machine Synthetic Minority Over-sampling Technique) is a sophisticated approach that combines the strengths of Support Vector Machines (SVM) and SMOTE to address the issue of class imbalance in datasets. By leveraging SVM to identify critical regions of the feature space, SVM SMOTE generates synthetic samples that enhance the decision boundary between classes, improving the overall classification performance.

3.2.8 KD-SMOTE

KD-SMOTE (Knowledge Discovery Synthetic Minority Over-sampling Technique) is an advanced variant of the original SMOTE algorithm designed to address class imbalance in datasets by incorporating knowledge discovery principles. By leveraging additional domain knowledge and data-driven insights, KD-SMOTE aims to generate more informative and effective synthetic samples, thereby enhancing the classifier's performance.

3.2.9 KNS SMOTE

KNS SMOTE (K-Nearest SMOTE) is an advanced variant of the SMOTE (Synthetic Minority Over-sampling Technique) algorithm designed to address class imbalance in datasets. The method integrates the principles of SMOTE with K-nearest neighbours (KNN) to create more realistic synthetic samples by focusing on the nearest neighbours of the minority class samples.

There are slight differences between SMOTE and KNS SMOTE. SMOTE randomly selects minority samples without consideration of their safety or local neighbourhood context but in KNS it selects the samples based on the proportion of minority neighbours, making sure that the synthetic samples are generated in safe regions.

3.3 Model Selection

The key point of selecting the majority of the models out of the 5 models is because of Sharma and Gosain (2023) research. This is an extension of the method that the Sharma and Gosain (2023) has followed and there it was tested that Random forest, KNN and SVM has ability to perform under SMOTE techniques. Additionally, Logistic Regression and Gradient Boosting was added to the analysis.

3.3.1 Logistic Regression

Logistic Regression is used in binary classification. This technique is applied in a logistic function to a linear combination of features that is available in the data. Logistic regression is simple, interpretable, and efficient on large datasets, but the only limitation is that it assumes a linear relationship and is limited to binary classification. But it can be extended to Multinomial Logistic Regression, which allows extending the binary classification to more than two categories. Common applications include medical diagnosis and credit scoring.

3.3.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is used for both classification and regression, and in the study it is used for classification. classification is based on the majority class of the nearest neighbours, and it's simple, and it has no connection with the data distribution. But the limitations are the computational power that it takes, and the results can depend on the k metric of the model. k value is the number of nearest neighbours taken into consideration when making a prediction for a new data point.

3.3.3 Support Vector Machines (SVM)

Support Vector Machines (SVM) work by finding the optimal hyperplane that separates classes, with extensions for non-linear classification using kernel functions. SVMs are used for analysis in high-dimensional spaces and robust to overfitting. This method is used for categorization and the reason why it is used in this analysis.

3.3.4 Random Forest

Random Forest used decision trees and the predictions of those are aggregated. This model is suitable because it reduces overfitting, and it can handle large datasets well, and provides feature importance scores which shows the most influential features when making predictions. Applications include medical diagnosis and financial forecasting can be benefited from this model.

3.3.5 Gradient Boosting

Gradient Boosting builds the models sequentially, where it corrects to correct the errors of its predecessor by focusing on the residuals. Residuals are the difference of the actual and the predicted values. This method approach helps to gradually reduce the overall error, but careful tune the model to reduce overfitting and ensure that the model performs well.

4 Design Specification

- **Functional requirements**

  As mentioned above in the methodology section the main focus on this research is to compare a few SMOTE techniques to the cirrhosis data given and compare the performance of the models that was suggested in the study. The performances of the SMOTE techniques will be analysed using the models and the evaluation criteria for the performance will be the Accuracy, Precision, Recall and F1-Score. But later in the result section it will be discussed that the evaluation criteria for this research mainly focused on the Recall metrics mainly for the models.

- **Non Functional Requirements**

  The results of the study will be presented in an interpretable format with the results in the tables that can be used by the professionals to understand and compare the SMOTE techniques. The results will be tested in multiple trials that it the results will be reliable. Finally there is always room for expansion and the design will help to expand the SMOTE techniques and add more SMOTE techniques to compare between the models.

- **Constraints**

  Few constraints of this research are the data reliability because the dataset is taken from the Kaggle and the data can be unreliable or missing. The missing data has been addressed in the research and techniques have been used to fill the missing data. Other constraint will be the processing time of the models. More data or more Models can increase the computational time of the entire analysis, and it can be a technical constraint.

5 Results and Evaluation

5.1 Performance of the Model's with SMOTE inclusion

5.1.1 Results of SMOTE inclusion

Data that was generated using SMOTE was evaluated through 5 models and the evaluation metrics such as precision, recall,f1-score and support were generated.

Table 3: Classification Reports for Models Trained with SMOTE

Logistic Regression				K-Nearest Neighbors			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
1.0	0.06	0.25	0.10	1.0	0.11	0.50	0.18
2.0	0.20	0.21	0.21	2.0	0.23	0.26	0.24
3.0	0.47	0.44	0.45	3.0	0.43	0.31	0.36
4.0	0.78	0.48	0.60	4.0	0.76	0.55	0.64
Support Vector Machine				Random Forest			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00	1.0	0.00	0.00	0.00
2.0	0.26	0.26	0.26	2.0	0.20	0.21	0.21
3.0	0.53	0.50	0.52	3.0	0.46	0.53	0.49
4.0	0.62	0.52	0.57	4.0	0.59	0.45	0.51
Gradient Boosting							
Class	Precision	Recall	F1-Score				
1.0	0.00	0.00	0.00				
2.0	0.19	0.16	0.17				
3.0	0.50	0.62	0.56				
4.0	0.55	0.41	0.47				

In the Table3 you can see that few models have recall value of 0 for Stage 1 and only Logistic Regression and K-Nearest Neighbors have recall values for Stage 1. Overall, comparing the values, it shows that **K-Nearest Neighbors performs better with SMOTE inclusion.**

5.1.2 Results of Borderline SMOTE1 inclusion

Data that was generated using SMOTE was evaluated through 5 models and the evaluation metrics such as precision, recall,f1-score and support were generated.

In Table4 as same as SMOTE in Borderline SMOTE1 we have recall value of 0 for Stage 1 in SVM, Random Forest and Gradient Boosting and additionally Logistic regression shows 0 for recall as well. Only K-Nearest Neighbors have recall values for Stage 1. Overall, comparing the values, it shows that **K-Nearest Neighbors performs better with Borderline SMOTE1 inclusion.**

Table 4: Classification Reports for Models Trained with Borderline SMOTE 1

Logistic Regression				K-Nearest Neighbors			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00	1.0	0.06	0.25	0.10
2.0	0.19	0.21	0.20	2.0	0.12	0.16	0.14
3.0	0.44	0.44	0.44	3.0	0.46	0.38	0.41
4.0	0.75	0.52	0.61	4.0	0.59	0.34	0.43
Support Vector Machine				Random Forest			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00	1.0	0.00	0.00	0.00
2.0	0.19	0.21	0.20	2.0	0.24	0.32	0.27
3.0	0.53	0.50	0.52	3.0	0.44	0.50	0.47
4.0	0.61	0.48	0.54	4.0	0.62	0.45	0.52
Gradient Boosting							
Class	Precision	Recall	F1-Score				
1.0	0.00	0.00	0.00				
2.0	0.24	0.21	0.22				
3.0	0.45	0.62	0.53				
4.0	0.65	0.45	0.53				

5.1.3 Results of Safe Level SMOTE inclusion

Data that was generated using SMOTE was evaluated through 5 models and the evaluation metrics such as precision, recall,f1-score and support were generated.

As same as Borderline SMOTE1 in in Table Table5 Safe Level SMOTE we have recall value of 0 for Stage 1 in Random Forest and Gradient Boosting and SVM, Logistic regression Only K-Nearest Neighbors have recall values for Stage 1. Overall, comparing the values, it shows that **K-Nearest Neighbors performs better with Safe Level SMOTE inclusion as well**

Here's the table updated with the new data from the Safe level SMOTE results:

Table 5: Classification Reports for Models Trained with Safe Level SMOTE Resampling

Logistic Regression			
Class	Precision	Recall	F1-Score
1.0	0.06	0.25	0.10
2.0	0.23	0.26	0.24
3.0	0.38	0.25	0.30
4.0	0.60	0.52	0.56

K-Nearest Neighbors			
Class	Precision	Recall	F1-Score
1.0	0.11	0.50	0.18
2.0	0.26	0.32	0.29
3.0	0.43	0.31	0.36
4.0	0.75	0.52	0.61

Support Vector Machine			
Class	Precision	Recall	F1-Score
1.0	0.08	0.25	0.12
2.0	0.19	0.21	0.20
3.0	0.52	0.53	0.52
4.0	0.59	0.34	0.43

Random Forest			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.21	0.26	0.23
3.0	0.45	0.53	0.49
4.0	0.60	0.41	0.49

Gradient Boosting			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.23	0.32	0.27
3.0	0.48	0.47	0.48
4.0	0.67	0.48	0.56

5.1.4 Results of Adaptive SMOTE ADASYN

Data that was generated using SMOTE was evaluated through 5 models and the evaluation metrics such as precision, recall,f1-score and support were generated

Table 6: Classification Reports for Models Trained with Adaptive SMOTE

Logistic Regression			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.25	0.21	0.23
3.0	0.42	0.62	0.50
4.0	0.78	0.48	0.60

K-Nearest Neighbors			
Class	Precision	Recall	F1-Score
1.0	0.33	0.25	0.29
2.0	0.25	0.32	0.28
3.0	0.49	0.53	0.51
4.0	0.64	0.48	0.55

Support Vector Machine			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.42	0.26	0.32
3.0	0.44	0.69	0.54
4.0	0.64	0.48	0.55

Random Forest			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.30	0.32	0.31
3.0	0.41	0.56	0.47
4.0	0.65	0.45	0.53

Gradient Boosting			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.31	0.26	0.29
3.0	0.45	0.62	0.53
4.0	0.58	0.48	0.53

In Table6 as same as Borderline SMOTE1 and Safe Level SMOTE, in Adaptive SMOTE we have recall value of 0 for Stage 1 in SVM, Random Forest and Gradient Boosting and Logistic regression Only K-Nearest Neighbors have recall values for Stage 1. Overall, comparing the values, it shows that **K-Nearest Neighbors performs better with Adaptive SMOTE inclusion as well**

5.1.5 Results of SMOTE Tomek inclusion

Data that was generated using SMOTE was evaluated through 5 models and the evaluation metrics such as precision, recall,f1-score and support were generated.

Table 7: Classification Reports for Models Trained with SMOTE Tomek

Logistic Regression				K-Nearest Neighbors			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
1.0	0.06	0.25	0.10	1.0	0.16	0.75	0.26
2.0	0.19	0.21	0.20	2.0	0.23	0.32	0.27
3.0	0.39	0.34	0.37	3.0	0.50	0.28	0.36
4.0	0.74	0.48	0.58	4.0	0.67	0.48	0.56
Support Vector Machine				Random Forest			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00	1.0	0.00	0.00	0.00
2.0	0.24	0.26	0.25	2.0	0.21	0.26	0.23
3.0	0.44	0.38	0.41	3.0	0.36	0.38	0.37
4.0	0.64	0.55	0.59	4.0	0.60	0.41	0.49
Gradient Boosting							
Class	Precision	Recall	F1-Score				
1.0	0.00	0.00	0.00				
2.0	0.31	0.26	0.29				
3.0	0.44	0.53	0.48				
4.0	0.67	0.48	0.56				

Table 7 shows that as same as Borderline SMOTE1 in SMOTE Tomek, we have recall value of 0 for Stage 1 in SVM, Random Forest and Gradient Boosting. Only K-Nearest Neighbors and Logistic regression have recall values for Stage 1. Additionally, it is visible that in SMOTE Tomek the recall value for Stage 1 is very high compared to previous SMOTE techniques we analysed in previous steps. Overall, comparing the values, it shows that **K-Nearest Neighbors performs better with SMOTE Tomek inclusion as well**

5.1.6 Results of SMOTE Rknn inclusion

Data that was generated using SMOTE was evaluated through 5 models and the evaluation metrics such as precision, recall,f1-score and support were generated

Table 8: Classification Reports for Models Trained with SMOTE-RkNN

Logistic Regression				K-Nearest Neighbors			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00	1.0	0.10	0.50	0.17
2.0	0.21	0.21	0.21	2.0	0.20	0.26	0.23
3.0	0.40	0.44	0.42	3.0	0.46	0.34	0.39
4.0	0.71	0.34	0.47	4.0	0.73	0.38	0.50
Support Vector Machine				Random Forest			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00	1.0	0.00	0.00	0.00
2.0	0.24	0.21	0.22	2.0	0.31	0.26	0.29
3.0	0.47	0.56	0.51	3.0	0.42	0.56	0.48
4.0	0.67	0.41	0.51	4.0	0.59	0.34	0.43
Gradient Boosting							
Class	Precision	Recall	F1-Score				
1.0	0.09	0.25	0.13				
2.0	0.23	0.16	0.19				
3.0	0.47	0.56	0.51				
4.0	0.68	0.52	0.59				

As same as Borderline SMOTE1 in SMOTE RkNN. Additionally to Logistic Regression and K-Nearest Neighbors, we can see that Gradient Boosting has recall value for Stage 1. But in Table8 if you consider the overall four stages, **K-Nearest Neighbors performs better with SMOTE Tomek inclusion as well**

5.1.7 Results of SVM SMOTE inclusion

Data that was generated using SMOTE was evaluated through 5 models and the evaluation metrics such as precision, recall,f1-score and support were generated.

In Table9 as same as Borderline SMOTE1 in SVM SMOTE, only K-Nearest Neighbors shows recall values for stage 1. So without any debate, **K-Nearest Neighbors performs better with SVM SMOTE inclusion as well**

Table 9: Classification Reports for Models Trained with SVM-SMOTE

Logistic Regression			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.17	0.21	0.19
3.0	0.42	0.44	0.43
4.0	0.78	0.48	0.60

K-Nearest Neighbors			
Class	Precision	Recall	F1-Score
1.0	0.08	0.25	0.12
2.0	0.12	0.16	0.14
3.0	0.44	0.34	0.39
4.0	0.65	0.52	0.58

Support Vector Machine			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.19	0.21	0.20
3.0	0.53	0.56	0.55
4.0	0.64	0.48	0.55

Random Forest			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.22	0.26	0.24
3.0	0.43	0.50	0.46
4.0	0.55	0.41	0.47

Gradient Boosting			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.20	0.16	0.18
3.0	0.42	0.62	0.50
4.0	0.63	0.41	0.50

5.1.8 Results of KD SMOTE inclusion

Data that was generated using SMOTE was evaluated through 5 models and the evaluation metrics such as precision, recall,f1-score and support were generated

Table 10: Classification Reports for Models Trained with KD-SMOTE

Logistic Regression			
Class	Precision	Recall	F1-Score
1.0	0.06	0.25	0.10
2.0	0.20	0.21	0.21
3.0	0.47	0.44	0.45
4.0	0.78	0.48	0.60

K-Nearest Neighbors			
Class	Precision	Recall	F1-Score
1.0	0.11	0.50	0.18
2.0	0.23	0.26	0.24
3.0	0.43	0.31	0.36
4.0	0.76	0.55	0.64

Support Vector Machine			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.26	0.26	0.26
3.0	0.53	0.50	0.52
4.0	0.62	0.52	0.57

Random Forest			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.20	0.21	0.21
3.0	0.44	0.47	0.45
4.0	0.55	0.41	0.47

Gradient Boosting			
Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00
2.0	0.19	0.16	0.17
3.0	0.50	0.62	0.56
4.0	0.55	0.41	0.47

In Table10 that shows the results for KD SMOTE, Logistic Regression and K-Nearest Neighbors has recall value for Stage 1. But if you consider the overall four stages, **K-Nearest Neighbors performs better with KD SMOTE inclusion.**

5.1.9 Results of KNS SMOTE inclusion

Data that was generated using SMOTE was evaluated through 5 models and the evaluation metrics such as precision, recall,f1-score and support were generated.

Table 11: Classification Reports for Models Trained with KNS-SMOTE

Logistic Regression				K-Nearest Neighbors			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
1.0	0.06	0.25	0.10	1.0	0.11	0.50	0.18
2.0	0.20	0.21	0.21	2.0	0.23	0.26	0.24
3.0	0.47	0.44	0.45	3.0	0.43	0.31	0.36
4.0	0.78	0.48	0.60	4.0	0.76	0.55	0.64
Support Vector Machine				Random Forest			
Class	Precision	Recall	F1-Score	Class	Precision	Recall	F1-Score
1.0	0.00	0.00	0.00	1.0	0.00	0.00	0.00
2.0	0.26	0.26	0.26	2.0	0.19	0.21	0.20
3.0	0.53	0.50	0.52	3.0	0.46	0.50	0.48
4.0	0.62	0.52	0.57	4.0	0.55	0.41	0.47
Gradient Boosting							
Class	Precision	Recall	F1-Score				
1.0	0.00	0.00	0.00				
2.0	0.19	0.16	0.17				
3.0	0.50	0.62	0.56				
4.0	0.55	0.41	0.47				

Table11 shows in KNS SMOTE, Logistic Regression and K-Nearest Neighbors has recall value for Stage 1. But if you consider the overall four stages, **K-Nearest Neighbors performs better with KD SMOTE inclusion.**

5.2 Comparison of the Performance

Table 12: Accuracy of Different Models Trained with Various SMOTE Techniques

SMOTE Technique	**LR**	**KNN**	**SVM**	**RF**	**GB**
SMOTE	0.39	0.39	0.43	0.40	0.42
Borderline SMOTE 1	0.39	0.31	0.40	0.42	0.44
SLS	0.43	0.39	0.44	0.45	0.42
SMOTE Tomek	0.36	0.38	0.39	0.35	0.43
SMOTE-RkNN	0.33	0.35	0.40	0.39	0.44
SVM-SMOTE	0.38	0.36	0.43	0.39	0.42
KD-SMOTE	0.39	0.39	0.43	0.37	0.42
KNS-SMOTE	0.39	0.39	0.43	0.38	0.42
Adaptive SMOTE	0.45	0.45	0.49	0.44	0.46

LR, KNN, SVM, RF and GB refers to Logistic Regression, K Nearest Neighbours, Support Vector machine, Random forests and Gradient boosting respectively. Comparing between the SMOTE techniques, the most important metric to check is the accuracy of the models. It is visible in the Table 12 that the Gradient boosting models have higher accuracy compared to other models as an average. But as an individual, Adaptive SVM with SVM has shown the highest accuracy. But as discussed in the part a of results most of the models had no recall values for Stage 1 and that shows that the models don't show a good predictive ability in all stages. What was consistent in every SMOTE technique was that the K-Nearest Neighbors showed recall values for all the stages, and the accuracy of the K-Nearest Neighbors is comparatively high. Therefore, as a comparison, it is safe to say that K-Nearest Neighbors with Adaptive SMOTE performs better at predicting all the stages in Cirrhosis dataset and overall Adoptive SMOTE is recorded the higher Accuracy throughout the 5 models that were used to evaluate the data.

6 Conclusion and Future Work

he Cirrhosis dataset had two main categorical variable which are the Stages and the Status of the patients and for this research the Stages of the Patients was chosen as the target variable. While exploring the data there was an imbalance of data which lately identified that is affecting the analysis. Therefore, SMOTE techniques were used to balance the data and the target of the research was to identify which SMOTE technique works better than others. This was evaluated using 5 different models. Even though most of the model and SMOTE technology combinations showed good accuracy, it failed to perform well in recalling the stages. The basic idea of the research is to predict stages, and what is the point of having a higher accuracy overall if the recall value of one stage is low? Therefore, even though having a low accuracy, Adaptive SMOTE technique showed better performance than other SMOTE techniques and specifically Adaptive SMOTE with K-Nearest Neighbors showed the better metrics accuracy and the recall of the stages combined. It is further realized that the Stage 1 having 0 recall value in some models because of the Stage 1 mostly getting classified as stage 3.

But there is room for improvement here. The accuracy is not very high in any dataset that was used SMOTE technique. Also, most of the models had lower recall values for the stages. This dataset has very few variables that could be analysed to predict the Stages variable. This can be a result of Stage 1 getting classified as stage 3 because there is not enough data to make the distinction difference. In future work, the dataset can be developed be adding more variables into the dataset. This research was mainly focused on evaluating the performance of the SMOTE techniques, but in the future it is possible to add techniques such as Extensive Hyperparameter Tuning and additional Boosting Algorithms to manipulate the dataset and see the performance of the models.

References

Araf, I., Idri, A. and Chairi, I. (2024a). Cost-sensitive learning for imbalanced medical data: a review, *Artificial Intelligence Review* **57**(4): 80.
URL: *https://doi.org/10.1007/s10462-023-10652-8*

Araf, I., Idri, A. and Chairi, I. (2024b). Cost-sensitive learning for imbalanced medical data: a review, *Artificial Intelligence Review* **57**(4): 80.
URL: *https://doi.org/10.1007/s10462-023-10652-8*

Bowyer, K. W., Chawla, N. V., Hall, L. O. and Kegelmeyer, W. P. (2011). SMOTE: synthetic minority over-sampling technique, *CoRR* **abs/1106.1813**.
URL: *http://arxiv.org/abs/1106.1813*

Bunkhumpornpat, C., Sinapiromsaran, K. and Lursinsap, C. (2009). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, *in* T. Theeramunkong, B. Kijsirikul, N. Cercone and T.-B. Ho (eds), *Advances in Knowledge Discovery and Data Mining*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 475–482.

Fernández, A., Garcia, S., Herrera, F. and Chawla, N. V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, *Journal of artificial intelligence research* **61**: 863–905.

Graber, M. L., Franklin, N. and Gordon, R. (2005). Diagnostic Error in Internal Medicine, *Archives of Internal Medicine* **165**(13): 1493–1499.
URL: *https://doi.org/10.1001/archinte.165.13.1493*

Han, H., Wang, W.-Y. and Mao, B.-H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning, *in* D.-S. Huang, X.-P. Zhang and G.-B. Huang (eds), *Advances in Intelligent Computing*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 878–887.

He, H., Bai, Y., Garcia, E. A. and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning, *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328.

Kawashima, A., Furukawa, T., Imaizumi, T., Morohashi, A., Hara, M., Yamada, S., Hama, M., Kawaguchi, A. and Sato, K. (2024). Predictive models for palliative care needs of advanced cancer patients receiving chemotherapy, *Journal of Pain and Symptom Management* **67**(4): 306–316.e6.
URL: *https://www.sciencedirect.com/science/article/pii/S0885392424000113*

Liu, C., Wu, J., Mirador, L., Song, Y. and Hou, W. (2018). Classifying dna methylation imbalance data in cancer risk prediction using smote and tomek link methods, *in* Q. Zhou, Q. Miao, H. Wang, W. Xie, Y. Wang and Z. Lu (eds), *Data Science*, Springer Singapore, Singapore, pp. 1–9.

Miftahushudur, T., Sahin, H. M., Grieve, B. and Yin, H. (2023). Enhanced svm-smote with cluster consistency for imbalanced data classification, *in* P. Quaresma, D. Camacho, H. Yin, T. Gonçalves, V. Julian and A. J. Tallón-Ballesteros (eds), *Intelligent Data Engineering and Automated Learning – IDEAL 2023*, Springer Nature Switzerland, Cham, pp. 431–441.

Roy, D., Roy, A. and Roy, U. (2024). *Learning from Imbalanced Data in Healthcare: State-of-the-Art and Research Challenges*, Springer Nature Singapore, Singapore, pp. 19–32.
URL: *https://doi.org/10.1007/978-981-99-8853-2*$_2$

Schuppan, D. and Afdhal, N. H. (2008). Liver cirrhosis, *The Lancet* **371**(9615): 838–851.
URL: *https://www.sciencedirect.com/science/article/pii/S0140673608603839*

Sharma, H. and Gosain, A. (2023). Oversampling methods to handle the class imbalance problem: A review, *in* K. K. Patel, K. C. Santosh, A. Patel and A. Ghosh (eds), *Soft Computing and Its Engineering Applications*, Springer Nature Switzerland, Cham, pp. 96–110.

Siriseriwan, W. and Sinapiromsaran, K. (2017). Adaptive neighbor synthetic minority oversampling technique under 1nn outcast handling., *Songklanakarin Journal of Science & Technology* **39**(5).

Syakiylla Sayed Daud, S. and Sudirman, R. (2023). Safe-level smote method for handling the class imbalanced problem in electroencephalography dataset of adult anxious state, *Biomedical Signal Processing and Control* **83**: 1–12.

Xu, Z., Shen, D., Nie, T., Kou, Y., Yin, N. and Han, X. (2021). A cluster-based oversampling algorithm combining smote and k-means for imbalanced medical data, *Information Sciences* **572**: 574–589.
URL: *https://www.sciencedirect.com/science/article/pii/S0020025521001985*