SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

The Effect of Balanced and Imbalanced Dataset for Comparing Machine Learning Models in Cancer Survival Prediction with Poverty Status Data

Michelle Tan[1], Stephanie Chua[1,*], Puteri Nor Ellyza Nohuddin[2]

[1] Faculty of Computer Science and Information Technology, University Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[2] Higher Colleges of Technology, Sharjah Women's College, 79799 Abu Dhabi, United Arab Emirates

ABSTRACT

This study focuses on the performance of machine learning algorithms on balanced and imbalanced datasets on cancer survival prediction with poverty status data. The intricate relationship between cancer survival and poverty was examined, addressing the pressing concern of cancer's substantial impact on mortality rates and the role of socioeconomic status in exacerbating disparities. Despite extensive examinations of the link between cancer mortality and socioeconomic status, little attention has been directed towards cancer survival rooted in poverty. Moreover, prevailing comparative studies typically focus on singular cancer types, leaving a void in comprehensive insights. This study seeks to bridge this gap by employing machine learning algorithms to predict cancer survival, leveraging data from a dataset extracted from SEER STAT. Five machine learning algorithms, namely, Support Vector Machine, Random Forest, Logistic Regression, Decision Tree, and Naïve Bayes were compared in their performances using balanced and imbalanced data with data from those above and below the poverty line. This study delved into class-balancing techniques to mitigate biases arising from imbalanced data, particularly in the context of poverty. The result showed that Support Vector Machine, Random Forest, Logistic Regression, and Naïve Bayes demonstrated stable and excellent performance in dealing with both balanced and imbalanced datasets. However, the performance of the Decision Tree was less satisfactory in this context.

Keywords:
Imbalanced dataset; Machine learning; Cancer survival; Prediction

1. Introduction

Cancer is a complex group of diseases characterized by the uncontrolled growth and spread of abnormal cells in the body. It can occur in virtually any organ or tissue and may manifest in various forms, such as carcinomas, sarcomas, leukemias, and lymphomas. Cancer is a major global health concern, and understanding it is crucial for effective prevention, diagnosis, and treatment. ML (ML) plays a crucial role in cancer research and treatment by harnessing the power of data analysis to improve early detection, treatment personalization, drug discovery, and patient outcomes. It

* *Corresponding author.*
E-mail address: chlstephanie@unimas.my

https://doi.org/10.37934/araset.59.2.232244

complements traditional medical approaches, helping clinicians and researchers make more informed decisions and advancing our understanding of cancer.

Using ML algorithms in the prediction of cancer occurrence and survival had been a popular method to foresee the outcomes, aiding in the decision-making process of healthcare professionals. However, the dataset in the real-world is often large, biased, and noisy. The accuracy obtained may be satisfactory, but the ML algorithms would have a large performance gap between the major class and the minor class. A model's performance was often biased towards the major class.

In certain cases of cancer occurrence, patients are privileged to have a better outcome and higher survival rate due to easier detection of cancer or earlier treatments of cancer. According to Yabroff *et al.,* [1] patients who are underprivileged often do not have the chance to seek treatments that could possibly avoid serious consequences that lead to deaths. The focus of this paper is on poverty, as we noticed a huge gap between the patients divided by the poverty line. It is found that countries with high poverty rate result in a higher rate of mortality for all types of cancer based on Kollman *et al.,* [2]. Figure 1 depicts the rate of mortality based on poverty status. Age-adjusted cancer death rates in Florida from 2014 to 2018 are estimated using a disparity risk ratio and a 95 percent confidence interval.

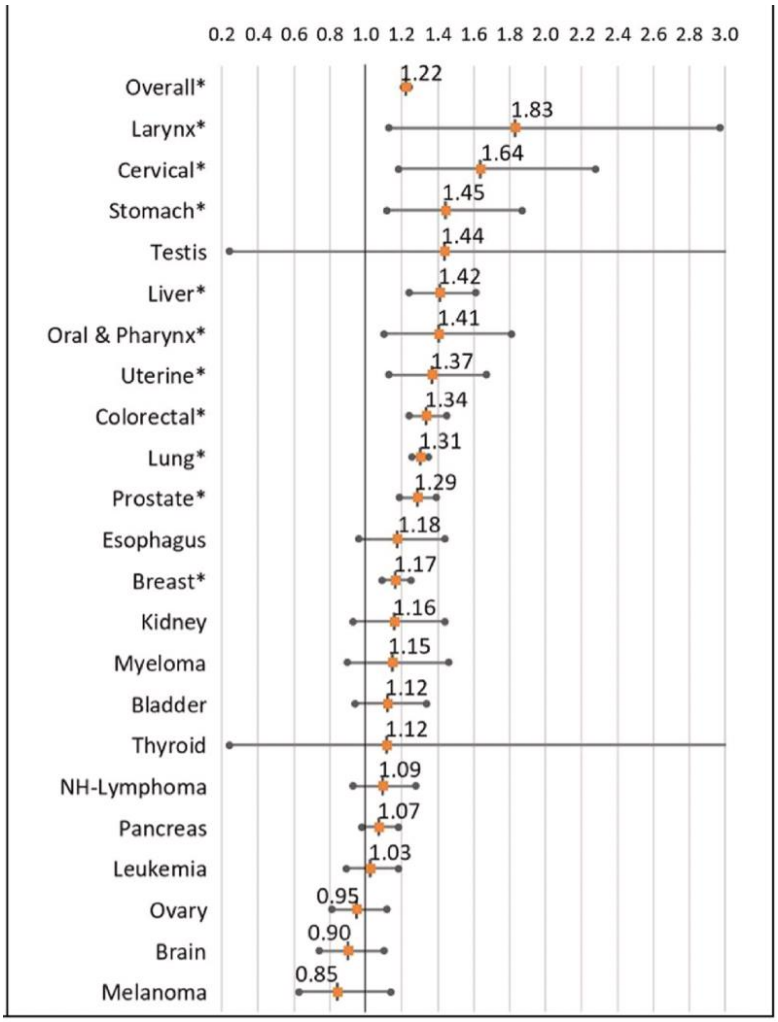

Fig. 1. Disparity of high vs. low poverty mortality rates

For example, larynx cancer demonstrates a high rate of mortality in high-poverty areas, where it is 83% higher than the rate of mortality in low-poverty areas. Besides, cervical cancer also has a high

rate of mortality in high-poverty areas. Poverty issues restrict the reach of an individual towards important resources that help to maintain our health, such as healthcare services and sanitation facilities [3].

To predict cancer survival, these ML algorithms that were widely used in this context were adopted in this research, including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and Naïve Bayes (NB). According to Naji *et al.,* [4] it was found that SVM was the most efficient algorithm when working with the dataset given. However, issues arise when there is uncertainty in the efficiency of algorithms when working with datasets of different characteristics. Besides, there were also insufficient studies on applying ML algorithms to predict cancer survival using poverty data. In this paper, a dataset was extracted from a SEER STAT database entitled 'Incidence- SEER Research Data, 18 Registries, Nov 2020 Sub (2000-2018)', including poverty-related data such as household income. Generally, the dataset includes 441,537 rows and 8 columns. We intend to fill the research gap by identifying the most efficient ML algorithm for predicting cancer survival based on poverty data. The evaluation metrics that will be applied in this study include accuracy, precision, F_1-measure, and recall.

The lack of comprehensive research comparing the performance of ML algorithms on cancer datasets, particularly those categorized by poverty statuses, poses a significant problem. Additionally, there is a dearth of studies examining the performance of ML algorithms that specifically differentiate between poverty and non-poverty groups. Furthermore, the variability of ML algorithm performance across datasets, particularly concerning evaluation metrics, presents another critical concern.

This study aims to achieve three main objectives. Firstly, it seeks to comprehensively evaluate how both imbalanced and balanced cancer datasets impact the performance of various ML models. Secondly, the study aims to compare the predictive capabilities of different ML algorithms concerning the survival prognosis of cancer patients, specifically focusing on distinguishing between poverty and non-poverty groups. Lastly, the research aims to identify the ML algorithm that exhibits the highest level of performance in accurately predicting cancer survival outcomes, with a particular emphasis on the contextual factors of poverty and non-poverty groups. By addressing these objectives, the study aims to contribute valuable insights into the application of ML in the medical domain, particularly in predicting cancer patient survival outcomes based on different socio-economic contexts.

2. Related Works

2.1 Poverty and Cancer Survival

This section discovers the relationship between poverty and the survival of cancer patients, as countries of higher poverty have a higher rate of cancer mortality as discussed in the previous section. Many factors contributed to the correlation between poverty and the survival of cancer patients, such as the failure to obtain treatment, limitations in accessing healthcare benefits, and insufficient knowledge in the domain of health, based on Denny *et al.,* [5].

Yousef *et al.,* [6] suggested that disparities between individuals of various socioeconomic statuses, such as race and employment, occur in obtaining colorectal cancer screening (CRC), where more than 20 million of them fail to obtain CRC. As a result, this causes death cases from conditions that could be avoided.

A study suggested that cervical and breast cancer disease and mortality in Brazil are influenced by socioeconomic status and access to healthcare services, according to Oliveira *et al.,* [7]. Various socioeconomic indicators were considered including per capita income and percentage of poverty.

In terms of mortality, the discrepancy in cervical cancer is a common issue across the world due to social inequalities. Hence, these studies have proven the positive correlation between socioeconomic status and cancer in terms of its occurrence and mortality. These results allow for reflection on the significance of structure and equity in health service access, allowing for a reorientation of public policies aiming at reducing health inequalities and maximizing quick access to high-quality treatments. However, there is no direct evidence whereby poverty is the main issue of cancer mortality.

2.2 Cancer Prediction via ML Models

Predictive models are used for risk estimation and to predict the outcome based on certain information. ML is also a technique that is commonly leveraged by researchers due to its suitability in analysing complicated datasets which results in the uncovering of patterns that may be hindered via traditional approaches. In this section, previous studies that incorporated poverty data in cancer prediction are studied, examining the techniques deployed and the outcomes of the research.

Dong *et al.,* [8] also acknowledged the impact of poverty on the inequality of breast cancer diagnosis in the late stage. The authors utilized the Classification and Regression Tree (CART) in this predictive study. The dataset from 2009 to 2018, including 812,048 patients that have breast cancer in the late stage was applied. Based on the results obtained, a relationship occurs between late-stage breast cancer diagnosis and poverty. The diagnosis is caused by issues such as high poverty, rurality, and high area deprivation.

A study was carried out to investigate cancer disparities due to rurality and poverty in Florida, focusing on 22 types of cancer, based on Hall *et al.,* [9]. The dataset consists of instances from high and low-poverty areas, in both urban and rural counties of Florida. Based on the results, high-poverty areas experience a higher rate of mortality due to cancer, which was 22% higher than areas of low poverty. However, the authors utilized statistical methods instead of ML algorithms in this research, leaving future work for cancer survival prediction using poverty data via ML algorithms.

2.3 Related Works on ML in Cancer Prediction

A review of recent comparative studies on ML in cancer prediction is examined in this section. The comparative studies aim to carry out a comparison in terms of the capability of prediction of the selected ML algorithms that were leveraged by the authors.

Related studies on the comparison of ML models in predicting the survival of cancer is discussed in this section. The work by Gong *et al.,* [10] aims to forecast individuals with esophageal cancer's five-year mortality state. The research investigates the use of ML methods. The Surveillance, Epidemiology and End Results (SEER) Program data were used in this investigation, as they were in this paper. Eight distinct models were applied by the authors for classification. These models comprised support vector machines (SVM), RF, NB, XGBoost, CatBoost, LightGBM, gradient boosting models (GBM), gradient boosting decision trees (GBDT), artificial neural networks (ANN), and gradient boosting models (SVM). Overall, the results point to the predictive power of ML techniques, particularly XGBoost, to accurately forecast the result.

In the study by Haque *et al.,* [11], ML algorithms were investigated for the potential to predict prognostic indicators about survival of breast cancer patients. It comprises female patients who were diagnosed with breast cancer between the years 2006 and 2010 and utilizes an extensive dataset gathered from the SEER Program. The prediction models are built using different ML methods, including gradient boosting (GB), kNN, DT, AdaBoost, LR, RF, voting classifier, and SVM. Based on the

results obtained, RF has the best accuracy of 95% among all the algorithms, whereas LR achieved the lowest accuracy (80.57 percent).

In the next study, Charlton *et al.,* [12] compare five classification models to predict the survival of brain tumour patients longer than a year after diagnosis in a retrospective manner. The ML algorithms applied include Bayesian Rule Lists, LR, RF, Explainable Boosting Machines (EBM), and SVM. These models were trained to forecast one-year survival. Based on the results obtained. RF model reached 78.9% accuracy which is the highest in this case, with macro-f_1 of 0.790, AUROC of 0.878, sensitivity of 0.844, and specificity of 0.734. The performance is followed by SVM achieving 77.7%, LR which achieves 77.5%, and EBM reaching only 77.1%.

Vial *et al.,* [13] compared the ML techniques to foresee the 2-year survival of non-small cell lung cancer (NSCLC) patients. The dataset consists of 422 records, where only 312 were valid and complete. Various ML techniques, including SVM, LR, and NB, are compared to determine the most effective method for survival prediction. The ML techniques implemented include SVM, NB, and LR. These algorithms are compared to find out the most effective algorithm for this context. The findings showed that SVM outperformed the other two algorithms in terms of accuracy. SVM attained an accuracy of 71.18%, hence being the ML algorithm that attain the highest accuracy according to this paper.

The study by Pradeep *et al.,* [14] focuses on the forecast of lung cancer patients' chances of survival using ML methods. To assess patterns connected to lung cancer, they combine SVM, NB, and C4.5 classification trees. The findings demonstrate that when trained on a larger dataset, C4.5 performs better in predicting lung cancer survivorship. It achieved a precision of 82.6% for a dataset of size 2200. In contrast, SVM has the lowest precision for the same dataset size, only achieving 54.9%. Hence, C4.5 seems to be a suitable algorithm for precise predictions that aid in the decision-making process.

A work by Akcay *et al.,* [15] was carried out to investigate the survival and repetition patterns in gastric cancer patients who underwent radiation therapy (RT) using ML techniques. The study analysed data from 75 instances of gastric cancer treated with RT and chemotherapy. Various ML algorithms, including LR, multilayer perceptron, XGBoost, SVM, RF, and Gaussian Naive Bayes (GNB), were used for prediction. The overall survival, hematogenous distant metastases, and peritoneal metastases were predicted using a variety of ML algorithms, including LR, XGBoost, SVM, multilayer perceptron, GNB, and RF. The research discovered that the most effective algorithms for the prediction of overall survival, peritoneal metastases, and distant metastases were GNB, XGBoost, and random forest. The accuracy achieved by GNB was 81% to predict the overall survival, making it the best-performing ML algorithm in this context. In the case of predicting peritoneal metastases, RF has the best performance the others by reaching an accuracy of 97%, whereas XGBoost was the best-performing ML algorithm in predicting distant metastases, achieving an accuracy of 86%.

In the paper by Ganggayah *et al.,* [16] the authors used ML approaches to discover important prognostic factors of breast cancer survival rates. The study made use of a massive dataset ranging from year 1993 to 2016, which included over 8000 cases. A variety of ML techniques were used, such as LR, DT, RF, extreme boost, NN, and SVM. The findings showed that RF method had the best accuracy (82.7%), while the DT approach had the lowest accuracy (79.8%).

In the study of Lynch *et al.,* [17], the authors utilized the SEER database to examine the use of various ML approaches to estimate lung cancer patients' survival times. In the study, the predictive ability of supervised learning techniques including LR, DT, GBM, SVM, and a bespoke ensemble is investigated. The GBM model, which has a Root Mean Square Error (RMSE) score of 15.05, gives the best performance. With an RMSE score of 15.82, the SVM model has the least outstanding performance in this context but still produces a standout result.

Despite the survival of cancer patients, ML algorithms are normally used in the detection, prediction, and diagnosis of cancer too. Much work has been done in the early prediction of cancer focusing on a specific type of cancer. Naji *et al.,* [18] utilized SVM, DT, RF, LR, and kNN in breast cancer prediction and diagnosis. The authors worked on the Wisconsin Breast Cancer Dataset which contains 569 rows. It is found that SVM has the best performance in terms of precision and accuracy of 97.2%. Besides, the ROC curves of each method were examined; the SVM showed the highest area under the ROC curve, an AUC score of 0.96 percent, demonstrating its superior overall performance in comparison to the other classifiers.

In terms of the most frequently used algorithm in recent years, Painuli *et al.,* [19] suggested that SVM has been the most chosen ML algorithm. However, there are multiple disadvantages when using SVM. SVM is less efficient when working with a noisy dataset and it is less suitable to be used for a large dataset. The medical dataset could be large and noisy as it includes many data types. Hence, it is most feasible if multiple algorithms are compared when carrying out research.

2.4 Synthetic Minority Random Oversampling Technique (SMOTE)

As the real-world data are mostly highly imbalanced and biased, the Synthetic Minority Random Oversampling Technique (SMOTE) is widely used in handling imbalanced data when performing predictions using ML algorithms. The equilibrium of classes is preserved with SMOTE. The characteristic of the minority group is analysed at first, then a number that is closest to k is selected based on Gupta *et al.,* [20].

SMOTE is also carried out by Ebrahimy *et al.,* [21] to resolve the issue of class imbalance to enhance the accuracy of prediction using ML algorithms. Other techniques such as ADASYN and ROS were also analysed by the author, where it is found that SMOTE would be a more suitable data balancing technique. A major disadvantage of ADASYN is its nature of producing synthetic records through weights, where those of bigger weights would be generated more frequently. Besides, ROS worked on a random basis when generating synthetic records. The nature of SMOTE working on k-nearest neighbours can offer bigger decision regions for the minor group. From the result obtained, the accuracy of the ML models increased with the use of SMOTE, especially in the minority group.

Moulaie *et al.,* [22] also applied SMOTE in the data-balancing process for the prediction of Covid-19 mortality. It is believed that ML algorithms can predict the survival of Covid-19 patients. However, when dealing with large, noisy, and imbalanced datasets, the performance of ML algorithms may be hindered. In this work, SMOTE is carried out after data pre-processing to overcome the issue of the imbalanced dataset.

3. Methodology

The overview of the research methodology is presented in this section. The main processes that were focused here were data acquisition, feature selection, data pre-processing, SMOTE Upsampling, data mining, and result analysis as shown in Figure 2.

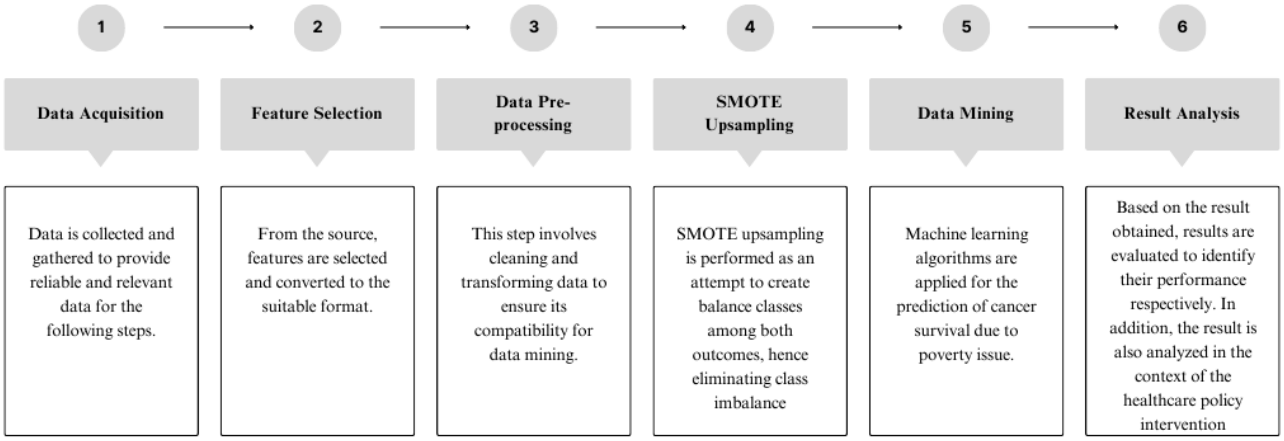

Fig. 2. Data mining process

3.1 Data Acquisition

In this project, the data source for cancer patients in the United States was obtained from the Surveillance, Epidemiology, and End Results (SEER) database that offers cancer-related data for research purposes. Data acquisition was performed via the SEER database, entitled 'Incidence- SEER Research Data, 18 Registries, Nov 2020 Sub (2000-2018)'. The first step was to request access to the SEER database through the SEER Data Access Request form. After receiving the approval via email, we were authorized to access the database. The dataset from 2018 was extracted as it was the most recent year that was included in the dataset, promising its relevance to this research.

3.2 Feature Selection

The SEER*Stat Software provided a platform to perform feature selection. A new case listing session was initialized using the platform as it listed down all the records needed. The features selected were Sex, Year of Diagnosis, Race Recode (W, B, AI, API), Site Recode ICD-O-3/WHO 2008, Median Household Income Inflation Adj to 2019, COD to Site Recode, Vital Status Recode (Study Cutoff Used), and Age Recode (<60,60-69,70+). In this context, only records from 2018 were selected, ensuring the relevance of the dataset. Any unknown records that contained empty columns or unknown columns were also filtered out using the Selection feature in the SEER*STAT software. After executing the session, a dataset with a total of 441,537 rows and 8 columns was acquired. It was then exported in the .xlsx format for data mining purposes. Table 1 summarizes the features that were selected along with their data types.

Table 1
List of features and data type

Feature	Data Type
ID	Integer
Sex	Binomial
Race Recode (W, B, AI, API)	Nominal
Site Recode ICD-O-3/WHO 2008	Nominal
Median Household Income Inflation Adj to 2019	Nominal
COD to Site Recode	Nominal
Vital Status Recode (Study Cutoff Used)	Binomial
Age Recode (<60,60-69,70+)	Nominal

3.3 Data Pre-Processing

The data pre-processing phase ensured that the data set was suitable for the data mining process. The process started in the data acquisition and feature selection phases. First, missing values were handled where it was ensured that the data set contains no empty rows and columns. This step took place during the data acquisition process in the SEER*STAT software, where empty instances were filtered out from the data set retrieved.

Balancing the dataset was a crucial task in this section, as it could potentially increase the performance of the ML models. The original dataset that was sampled was highly imbalanced, where it consisted of more than 4,000 cases that had a positive outcome (alive), but only less than 1,000 cases had a negative outcome (dead). Hence, it influenced the performance of the models, specifically in the prediction of negative cases.

This issue was resolved using SMOTE Upsampling during the data pre-processing process. SMOTE balanced the minority class with the majority class by producing more data for the minority class. It created brand-new synthetic cases amongst already current minority class instances. Table 2 depicts the distribution of dead and alive cases before and after SMOTE Upsampling. The four sets of data were used in the experiments to investigate the effects of the balanced and imbalanced dataset in comparing the ML models for cancer survival prediction with poverty status data.

Table 2
Distribution of cases of datasets

	Poverty Imbalanced Dataset	Poverty Balanced Dataset	Non-Poverty Imbalanced Dataset	Non-Poverty Balanced Dataset
Alive Cases	3,982	3,982	4,357	4,357
Dead Cases	1,018	3,982	643	4,357
Total Cases	5,000	7,964	5,000	8,714

3.4 Data Mining

In this phase, the data mining process was carried out. The datasets were trained and tested via the cross-validation technique. The dataset was divided into k=10 folds to train and test the models. The ML models that were applied were SVM, RF, LR, DT, and NB. In this phase, the experiments were carried out on the four sets of data.

Here, we could assess the effects of data balancing methods, particularly SMOTE, on the effectiveness of the ML models chosen by comparing the results generated from using the datasets. It could also be justified if SMOTE could improve prediction accuracy, precision, recall, and F_1-measure by addressing the class imbalance issue. The data mining process was carried out in parallel using SVM, RF, LR, DT, and NB.

3.5 Result Analysis

Model evaluation was done using accuracy, precision, recall, and F_1-measure. After evaluating the performance of the ML models, result interpretation was carried out to identify the suitability of ML models in the prediction of cancer survival. The evaluation was crucial in shaping healthcare interventions, especially in countries that were facing poverty. The results obtained are discussed in the next section.

4. Results

As mentioned in the previous section, the implementation of ML algorithms was applied to both balanced and imbalanced datasets. This aimed to identify the significance of a balanced dataset in terms of the performance of ML algorithms chosen. On the other hand, we also aimed to compare the ML algorithms selected to perform prediction of cancer survival based on poverty-related data.

Table 3 and Table 4 display the result comparison for both poverty and non-poverty dataset. The performance of ML algorithm for balanced and imbalanced datasets were compared side-by-side. After the dataset was extracted from the SEER database, the proportion of positive (alive) and negative (dead) outcome was highly imbalanced. According to Alam *et al.,* [23], imbalanced dataset often leads to the deterioration of quality of ML algorithms as it results in a biased prediction. Hence, suitable technique must be applied to handle the imbalanced dataset, ensuring that the result does not skew to a certain category. Replication was recommended as one of the methods to balance out the rare class of the imbalanced dataset. In this case, negative cases were generated synthetically via SMOTE.

According to Table 3, the balance of positive and negative classes produced only a slight impact for SVM, RF, LR, and NB models. Only the DT model was affected by the class balancing. After the dataset was balanced using SMOTE Upsampling, the accuracy of the DT model dropped from 87.14% to 49.95%. However, it had a more balanced performance in predicting the outcome of the positive and negative classes, in comparison with its performance for the imbalanced dataset where it was biased towards the positive class.

Table 3
Comparison of ML performances on balanced and imbalanced poverty dataset

	Balanced Poverty Dataset					Imbalanced Poverty Dataset				
	SVM	RF	LR	DT	NB	SVM	RF	LR	DT	NB
Accuracy (%)	99.97	99.96	100.00	49.96	99.97	99.74	99.66	99.70	79.64	99.80
Precision (%)	100.00	100.00	100.00	49.91	100.00	100.00	100.00	100.00	unknown	100.00
Recall (%)	99.95	99.92	100.00	20.00	99.95	98.72	98.33	98.53	0.00	99.02
F_1-measure (%)	99.97	99.96	100.00	28.52	99.97	99.35	99.15	99.24	unknown	99.50

Similar phenomenon was seen in the non-poverty dataset, where the performances of RF, SVM, LR, and NB were remarkably higher than DT as shown in Table 4. The performance of the SVM, LR, and NB models exceeded 97.00% when using the imbalanced dataset. All four ML models experienced a slight increase in performance after SMOTE Upsampling. RF had the most outstanding performance when working with a balanced non-poverty dataset, where 100.00% was achieved for all evaluation metrics. Drastic changes were seen in the DT models before and after handling the imbalanced dataset. The accuracy of the DT model was remarkably high, where it achieved 87.14%. However, the performance was skewed towards the positive class, resulting in a poor performance in predicting the negative class. Like the poverty dataset, the performance of the DT model was also less satisfactory compared to the other ML models.

Generally, the performance of ML algorithms was significantly better on the balanced dataset compared to the imbalanced dataset. Besides, the stability of models also improved after handling the imbalanced dataset. Based on the result above, the RF model benefited the most from SMOTE Upsampling. Besides, the performance of the LR model was the most outstanding model among all ML models for the balanced poverty dataset, while the RF model had the best performance among the ML models for the balanced non-poverty dataset.

Table 4
Comparison of ML performances on balanced and imbalanced non-poverty dataset

	Balanced Non-Poverty Dataset					Imbalanced Non-Poverty Dataset				
	SVM	RF	LR	DT	NB	SVM	RF	LR	DT	NB
Accuracy (%)	99.99	100.00	99.99	49.95	99.99	99.60	99.22	99.60	87.14	99.78
Precision (%)	100.00	100.00	100.00	49.94	100.00	99.83	100.00	99.83	unknown	100.00
Recall (%)	99.98	100.00	99.98	40.00	99.98	97.06	93.94	97.06	0.00	98.29
F_1-measure (%)	99.99	100.00	99.99	44.38	99.99	98.41	96.85	98.41	unknown	99.13

It was found that the ML models potentially impacted the healthcare sector with their ability to predict the survival of cancer patients. First, it facilitates healthcare professionals in making clinical decisions. The vast amount of healthcare data enables the ML algorithms to learn the patterns, and ultimately provide reliable predictions in various kinds of healthcare issues, including cancer, based on Javaid *et al.,* [24]. As the dataset obtained could be large, noisy, and biased, it was crucial to select an ML algorithm that could cater to these issues.

From the experiment that was carried out above, it was found that SVM, LR, RF, and NB acquires better performances as compared to DT. As the SVM, LR, RF, and NB models demonstrated little to no difference in the performance between poverty and non-poverty datasets, it was suggested that these ML models could be applied in developed, developing, and non-developed countries as there was no significant effect on the performance due to demographic issues. Hence, the models could support healthcare professionals in making clinical decisions regardless of the demographic status of patients.

In most countries that are facing poverty, having resource allocation is crucial. There are unfair disparities in healthcare distribution between those facing poverty and those who do not. Hence, new initiatives and policies that neglect the importance of considering poverty groups might unintentionally worsen population disparities if they are not implemented with care. Considering the underprivileged group was of utmost importance in making healthcare policies. The aid of ML algorithms in predicting the survival of cancer patients is important in resolving distribution issues by improving the treatment of underprivileged groups. The limited healthcare resources must be distributed effectively. Healthcare professionals might prioritize the patients with higher rates of mortality for more intense treatments and support by using ML-based predictions.

The stability and high performance of the LR, SVM, RF and NB models can also be applied in various contexts other than the prediction of cancer survival based on poverty dataset. Healthcare professionals can create individualized treatments based on patient's risk profiles and economic backgrounds by utilizing the prediction potential of LR, SVM, RF, and NB. Healthcare professionals may identify people from low-income backgrounds who are not likely to survive cancer by using reliable and accurate prediction algorithms. This makes it possible to focus screening efforts and guarantee that high-risk patients receive earlier and more frequent tests, which results in early identification and better treatment outcomes.

5. Conclusions

Wrapping up this paper, it revolves around the performance of ML algorithm on balanced and imbalanced datasets, which is further split into poverty and non-poverty dataset as the patients are grouped above and below the poverty line. The data balancing technique used and how it affects the prediction of the cancer patients' survival based on poverty data was studied. From our studies, it was identified that the balancing of dataset is not a significant contributor towards the performance of the selected ML algorithms. The performances of LR, RF, SVM, and NB had a slight increase upon

balancing the dataset. On the other hand, DT demonstrated a relatively better performance when the dataset was balanced. It could predict the minor class after the class distribution was balanced out. Besides, it was found that there was no significance between the performance of ML algorithms on poverty and non-poverty datasets. LR, RF, SVM, and NB were relatively stable algorithms with satisfactory performances when working with all four datasets, indicating their versatility to work with real-world datasets that could be large, noisy, and biased. It also indicated that the ML models could be adopted to predict cancer survival regardless of the demographic data. Besides, it is also applicable to developed, developing, and non-developed countries, where there is a disparity of demographic data. Based on the results obtained, LR, RF, SVM, and NB would be a good fit to work with real-world datasets in the prediction of cancer survival regardless of the demographic data. Ultimately, it is hoped that this experiment could aid decision-makers in enhancing healthcare policies to reduce disparities between the poverty and non-poverty groups.

There were significant limitations in this experiment. First, overfitting may occur in the process of data mining. As the balanced datasets were achieved through SMOTE, similar records were synthetically generated. Hence, it may affect the actual performance of the ML algorithms in the real-world context. Certain ML algorithms, such as LR, are also prone to overfitting due to their linearly related assumption between dependent and independent data. Next, there were limited classes that indicated the poverty status of cancer patients. In our context, the household income was the only indicator to identify if the patient belonged to the category of poverty. Indicators such as asset ownership and demographic statuses were not considered in this study. Hence, this would be an area of improvement for future work. In addition, we worked on datasets where the size was relatively smaller compared to the real-world dataset due to limitations in computation power. This would potentially hinder the performance of certain ML algorithms. Future work is expected to cover these issues.

Acknowledgment

The authors would like to acknowledge the Faculty of Computer Science and Information Technology (FCSIT), Universiti Malaysia Sarawak (UNIMAS) for the research opportunity and financial assistance.

References

[1] Yabroff, K. Robin, Ted Gansler, Richard C. Wender, Kevin J. Cullen, and Otis W. Brawley. "Minimizing the burden of cancer in the United States: Goals for a high-performing health care system." *CA: a cancer journal for clinicians* 69, no. 3 (2019): 166-183.

[2] Kollman, John, and Holly L. Sobotka. "Peer Reviewed: Poverty and Cancer Disparities in Ohio." *Preventing chronic disease* 15 (2018).

[3] Önal, Ayşe Emel, ed. *Healthcare Access-New Threats, New Approaches: New Threats, New Approaches*. BoD–Books on Demand, 2023.

[4]

References

[4] Naji, Mohammed Amine, Sanaa El Filali, Kawtar Aarika, EL Habib Benlahmar, Rachida Ait Abdelouhahid, and Olivier Debauche. "Machine learning algorithms for breast cancer prediction and diagnosis." Procedia Computer Science 191 (2021): 487-492. https://doi.org/10.1016/j.procs.2021.07.062

[5] Denny, Lynette, Ahmedin Jemal, Mary Schubauer-Berigan, Farhad Islami, Nadia Vilahur, Miranda Fidler, Diana Sarfati, Isabelle Soerjomataram, Catherine de Martel, and Salvatore Vaccarella. "Social inequalities in cancer risk factors and health-care access." (2021).

[6] Yousef, Mohammad A., Niteesh Sundaram, Swadha Das Guru, Burt Cagir, Robert Behm, Michael J. Georgetson, and Matthew Lincoln. "TThe Influence of Demographic, Racial, and Socioeconomic Factors on Colorectal Cancer Screening at a Mid- Atlantic Tertiary Rural Healthcare Centre." In Gastroenterology, vol. 162, no. 7, (2022): S699-S700. https://doi.org/10.1016/S0016-5085(22)61636-6

[7] de Oliveira, Nayara Priscila Dantas, Camila Alves dos Santos Siqueira, Kálya Yasmine Nunes de Lima, Marianna de Camargo Cancela, and Dyego Leandro Bezerra de Souza. "Association of cervical and breast cancer mortality with socioeconomic indicators and availability of health services." Cancer Epidemiology 64 (2020): 101660. https://doi.org/10.1016/j.canep.2019.101660

[8] Dong, Weichuan, Wyatt P. Bensken, Uriel Kim, Johnie Rose, Nathan A. Berger, and Siran M. Koroukian. "Phenotype discovery and geographic disparities of late-stage breast cancer diagnosis across US Counties: a machine learning approach." Cancer Epidemiology, Biomarkers & Prevention 31, no. 1 (2022): 66-76. https://doi.org/10.1158/1055-9965.EPI-21-0838

[9] Hall, Jaclyn M., Sarah M. Szurek, Heedeok Cho, Yi Guo, Michael S. Gutter, Georges E. Khalil, Jonathan D. Licht, and Elizabeth A. Shenkman. "Cancer disparities related to poverty and rurality for 22 top cancers in Florida." Preventive Medicine Reports 29 (2022): 101922. https://doi.org/10.1016/j.pmedr.2022.101922

[10] Gong, Xian, Bin Zheng, Guobing Xu, Hao Chen, and Chun Chen. "Application of machine learning approaches to predict the 5-year survival status of patients with esophageal cancer." Journal of Thoracic Disease 13, no. 11 (2021): 6240. https://doi.org/10.21037/jtd-21-1107

[11] Haque, Mohammad Nazmul, Tahia Tazin, Mohammad Monirujjaman Khan, Shahla Faisal, Sobhee Md Ibraheem, Haneen Algethami, and Faris A. Almalki. "Predicting characteristics associated with breast cancer survival using multiple machine learning approaches." Computational and Mathematical Methods in Medicine 2022 (2022). https://doi.org/10.1155/2022/1249692

[12] Charlton, Colleen E., Michael TC Poon, Paul M. Brennan, and Jacques D. Fleuriot. "Development of prediction models for one-year brain tumour survival using machine learning: a comparison of accuracy and interpretability." Computer methods and programs in biomedicine 233 (2023): 107482. https://doi.org/10.1016/j.cmpb.2023.107482

[13] Vial, Alanna, David Stirling, Matthew Field, Montserrat Ros, Christian Ritz, Martin Carolan, Lois Holloway, and Alexis A. Miller. "A comparative study of machine learning techniques for the improved prediction of nsclc survival analysis." In 2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC), pp. 1-2. IEEE, 2018. https://doi.org/10.1109/NSSMIC.2018.8824269

[14] Pradeep, K. R., and N. C. Naveen. "Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4. 5 and Naive Bayes algorithms for healthcare analytics." Procedia computer science 132 (2018): 412-420. https://doi.org/10.1016/j.procs.2018.05.162

[15] Akcay, Melek, Durmus Etiz, and Ozer Celik. "Prediction of survival and recurrence patterns by machine learning in gastric cancer cases undergoing radiation therapy and chemotherapy." Advances in radiation oncology 5, no. 6 (2020): 1179-1187. https://doi.org/10.1016/j.adro.2020.07.007

[16] Ganggayah, Mogana Darshini, Nur Aishah Taib, Yip Cheng Har, Pietro Lio, and Sarinder Kaur Dhillon. "Predicting factors for survival of breast cancer patients using machine learning techniques." BMC medical informatics and decision making 19 (2019): 1-17. https://doi.org/10.1186/s12911-019-0801-4

[17] Lynch, Chip M., and Joshua D. BehnazAbdollahi. "Fuqua, Alexandra R. de Carlo, James A. Bartholomai, Rayeanne N. Balgemann, Victor H. van Berkel, Hermann B. Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques." International Journal of Medical Informatics 108 (2017): 1-8. https://doi.org/10.1016/j.ijmedinf.2017.09.013

[18] Naji, Mohammed Amine, Sanaa El Filali, Kawtar Aarika, EL Habib Benlahmar, Rachida Ait Abdelouhahid, and Olivier Debauche. "Machine learning algorithms for breast cancer prediction and diagnosis." Procedia Computer Science 191 (2021): 487-492. https://doi.org/10.1016/j.procs.2021.07.062

[19] Painuli, Deepak, and Suyash Bhardwaj. "Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review." Computers in Biology and Medicine 146 (2022): 105580. https://doi.org/10.1016/j.compbiomed.2022.105580

[20] Gupta, Palak, Anmol Varshney, Mohammad Rafeek Khan, Rafeeq Ahmed, Mohammed Shuaib, and Shadab Alam. "Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques." Procedia Computer Science 218 (2023): 2575-2584. https://doi.org/10.1016/j.procs.2023.01.231

[21] Ebrahimy, Hamid, Babak Mirbagheri, Ali Akbar Matkan, and Mohsen Azadbakht. "Effectiveness of the integration of data balancing techniques and tree-based ensemble machine learning algorithms for spatially-explicit land cover accuracy prediction." Remote Sensing Applications: Society and Environment 27 (2022): 100785. https://doi.org.remotexs.unimas.my/10.1016/j.rsase.2022.100785

[22] Moulaei, Khadijeh, Mostafa Shanbehzadeh, Zahra Mohammadi-Taghiabad, and Hadi Kazemi-Arpanahi. "Comparing machine learning algorithms for predicting COVID-19 mortality." BMC medical informatics and decision making 22, no. 1 (2022): 1-12. https://doi.org/10.1186/s12911-021-01742-0

[23] Alam, Talha Mahboob, Kamran Shaukat, Waseem Ahmad Khan, Ibrahim A. Hameed, Latifah Abd Almuqren, Muhammad Ahsan Raza, Memoona Aslam, and Suhuai Luo. "An efficient deep learning-based skin cancer classifier for an imbalanced dataset." Diagnostics 12, no. 9 (2022): 2115. https://doi.org/10.3390/diagnostics12092115

[24] Javaid, Mohd, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. "Significance of machine learning in healthcare: Features, pillars and applications." International Journal of Intelligent Networks 3 (2022): 58-73. https://doi.org/10.1016/j.ijin.2022.05.002