AI_Project_Report

by Monil Desai

Submission date: 24-Apr-2024 12:21PM (UTC+0530)

Submission ID: 2360148834

File name: AI_Project_Report_Updated.pdf (582.3K)

Word count: 3157

Character count: 18325

PANDIT DEENDAYAL ENERGY UNIVERSITY
SCHOOL OF TECHNOLOGY

Course: Artificial Intelligence
Course Code: 20CP310P
LAB MANUAL
B.Tech. (Computer Science and Engineering)
Semester 4

Submitted To:
Prof. Rajeev
Gupta

Submitted By:
Khushal Kathad(21BCP241)
Milan Patel(21BCP236)
Monil Desai(21BCP217)

Abstract

Prediction of life expectancy can significantly impact patient care and resource allocation. Looking into the implementation of ML methods that develop a model for estimating post-operative survival rates. Utilizing ML algorithms like KNN, Random Forest, and Logistic Regression for life expectancy prediction in thoracic surgery patients. It analyzes the factors influencing post-operative outcomes, such as patient demographics, medical history, type of surgery performed, health condition of the patient, and post-surgical complications. By employing machine learning on historical patient data aims to build a model that identifies key factors and their interactions in predicting patient survival. This model then be used to provide personalized estimates of life expectancy for future patients undergoing thoracic surgery, enabling better informed treatment decisions and improved patient prognoses.

Table of Content

List of Abbreviation

ML - machine learning

DL - deep learning

SVM - support vector machine

SMOTE- Synthetic Minority Oversampling Technique

ROC- Receiver Operating Characteristic Curve

AUC- Area Under the ROC Curve

CNNs – Convolutional Neural Network

RNNs- Recurrent Neural Network

EHR- Electronic Health Records

KNN – K-Nearest Neighbors

List of Keywords

Machine Learning

Deep Learning

Neural Network

Thoracic Surgery

Logistic Regression

Random Forest Classifier

CHAPTER 1 – INTRODUCTION

Thoracic surgery encompasses a wide range of procedures performed on the chest cavity, including the lungs, heart, esophagus, and trachea. While these surgeries can be life-saving, accurately predicting a patient's life expectancy after surgery is crucial for optimizing treatment plans and managing patient expectations. This information is valuable for both surgeons and patients, allowing for informed decision-making regarding the risks and benefits of surgery.

The dataset includes various features that potentially influence post-operative outcomes:

- Diagnosis: indicates specific medical condition requiring surgery
- Pulmonary Function Tests:
    - FVC (Forced Vital Capacity): It denotes the maximum lung capacity of a patient where he can forcefully exhale after a deep breath, indicating overall lung function.
    - FEV1 (Forced Expiratory Volume in one second): It is the volume of air forcefully exhaled in the initial moment of an FVC test, a measure of airway obstruction.
- Patient Symptoms:
    - Performance: Functional limitations experienced by the patient due to their condition.
    - Pain: Presence or absence of pain.
    - Haemoptysis: Coughing up blood.
    - Dyspnoea: Difficulty breathing.
    - Cough: Presence or absence of cough.
    - Weakness: Overall muscular weakness.
- Medical History:
    - Tumor_Size: Size of the tumor.
    - Diabetes Mellitus: Presence or absence of diabetes.
    - MI_6mo: Myocardial infarction (heart attack) within the past 6 months.
    - PAD: Peripheral arterial disease.
    - Smoking: Whether the patient smokes or not.
    - Ashtma: Presence or absence of asthma.
- Age: Chronological age of the patient.
- Death_1yr: Binary variable indicating whether the patient died within one year of surgery or not.

CHAPTER 2 - LITERATURE REVIEW

2.1 Introduction

Among the most important type of treatment for a number of pulmonary and thoracic disorders, such as lung cancer, emphysema, and thoracic trauma, is thoracic surgery. Clinical decision-making, patient counselling, and post-operative care can all be optimized by accurately predicting the patient's risk after thoracic surgery. Recent developments in deep learning neural networks and machine learning approaches have opened up new possibilities for creating predictive models that anticipate patient outcomes based on preoperative, intraoperative, and postoperative factors.

2.2 Machine Learning and Deep Learning Approaches

a. Logistic Regression

Logistic regression is popularly used as a statistical method for binary classification tasks, being appropriate for forecasting binary outcomes like patient survival following thoracic surgery [@peduzzi1996importance]. Previous papers have enumerated the functionality of logistic regression models in identifying significant predictors of postoperative mortality and morbidity, including patient demographics, comorbidities, surgical factors, and perioperative complications [@teh2015predicting].

b. Random Forest

Random forest, an ensemble learning algorithm that associates multiple decision trees to increase predictive performance and generalizability [@breiman2001random]. Researches have defined that random forest models can efficiently capture complex interactions between predictor variables and patient outcomes in thoracic surgery, leading to accurate predictions of mortality risk, length of hospital stay, and postoperative complications [@ghosh2015predicting].

c. Neural Networks

Deep neural networks, particularly CNNs and RNNs, are increasingly popular due to their abilities to learn intricate patterns and temporal dependencies in medical data [@litjens2017survey]. Recent research has explored the application of deep neural networks in predicting patient outcomes following thoracic surgery, leveraging features extracted from medical imaging, electronic health records (EHRs), and genetic data to enhance predictive accuracy [@wang2020deep].

2.3 Data Sources and Feature Selection

d. Electronic Health Records (EHRs)

EHRs contain a panorama of structured and unstructured data, including imaging examination, laboratory information, health history, and characteristics of patients [@cruz2019application].

e. Medical Imaging

Medical imaging modalities such as computed tomography, magnetic resonance imaging, and positron emission tomography provide valuable insights into thoracic anatomy, tumor characteristics, and disease progression [@chen2020artificial].

f. Genetic and Molecular Data

Genetic and molecular markers play a significant role in predicting patient outcomes and treatment response in thoracic malignancies such as lung cancer [@chen2020artificial].

CHAPTER 3 - PROPOSED METHODOLOGY

To address the problem of generating a predictive model for forecasting one-year survival rate after thoracic surgery, a comprehensive methodology can be seen as below:

3.1 Data Preprocessing:

a. Handling Missing Values: Start by finding any missing records in the dataset. This may involve deletion methods such as mean or median deletion, or more complicated methods like predictive imputation.

b. Encoding Categorical Variables: In the dataset, Categorical values are defined to numerical values using methods like one-hot encoding or label encoding. This ensures model-fit with ml and dl algorithms.

c. Scaling Numerical Features: Numerical features are topped at a regular range (e.g., between 0 to 1) using techniques like Min-Max scaling or standardization. This averts features with larger scales from dictating the model training process.

3.2 Model Selection:

a. Exploration of Algorithms: Finding a variety of ml and dl algorithms is suitable for binary classification problems. These may include logistic regression, random forest classification, SVMs, gradient boosting, and deep neural networks among many others.

b. Hyperparameter Tuning: For each of these algorithm, performing hyperparameter tuning using methods like grid search or random search is tantamount to optimize model performance.

3.3 Addressing Class Imbalance:

a. Undersampling: Downsampling the majority class can bring balance to the dataset, helping in making sure that both classes are represented equally. This averts the model from becoming biased towards the majority partition.

b. Oversampling: Instead, Oversampling the minority class through methods like random oversampling as well as SMOTE can also do the trick. This generates artificial samples to intensify the representation of the minority class in the dataset.

c. Evaluation Metrics: Use evaluation metrics that are vigorous and impartial to class imbalance, such as precision, recall, F1-score, and ROC-AUC.

3.4 Model Training and Evaluation:

a. Training-Validation Split: Fragment the dataset into training and validation sets, usually using a 70-30 or 80-20 fraction. The validation set is used to estimate model performance and make hyperparameter adjustments after the model has been trained using the training set.

b. Training Process: Train every model on the training set by means of the chosen algorithm and hyperparameters. For deep learning models, this may include training over multiple epochs with methods like early stopping to avoid overfitting.

c. Model Evaluation: Evaluate the trained models on the validation set by suitable evaluation metrics. This allows to find the differences in the performance of different models and select the model with best efficiency for deployment.

3.5 Comparison and Discussion:

a. Discussion of Findings: Discuss the strengths as well as weaknesses of all model by the means of visualization through graphs, as well as their practical inferences for clinical decision-making. This discussion reports the recommendations for model deployment and future research directions.

By following this proposed methodology, the aim is to develop a vigorous predictive model that can precisely determine the possibility of one-year survival after thoracic surgery, ultimately improving patient lifestyles and informing clinical practice.

CHAPTER 4 – IMPLEMENTATION DETAILS

1) Data Preprocessing:

   a) Handling Missing Values and Encoding Categorical Variables:

   Missing records in the dataset were allocated with by employing methods such as mean or median deletion, ensuring that no important information was lost. Moreover, categorical variables were encoded through numerical values using methods like one-hot encoding, enabling the ml and dl algorithms to process them effectively.

   b) Scaling Numerical Features:

   To avoid any one feature from controlling the model training process, the dataset's numerical features were scaled to a normal range. This scaling guarantees that all features will contribute correspondingly to the model's learning process, allowing the model to be more balanced and give accurate predictions.

2) Model Selection:

   a) Logistic Regression:

   Logistic regression was chosen as it offers a modest yet effective approach for binary classification tasks. Its interpretability allows us to understand the impact of each feature on the predicted outcome, making it a suitable baseline model for comparison with more complex algorithms.

   b) Random Forest Classification:

   Random forest classification was selected due to its robustness to outliers and capability to capture nonlinear relationships in the data. This styles it particularly suitable for complex datasets like thoracic surgery data, where the relationships between features may be intricate and nonlinear.

   c) Deep Neural Networks (DNNs):

   Deep neural networks offer flexibility in learning complex patterns from high-dimensional data. By leveraging multiple layers of interconnected neurons, DNNs have the potential to capture intricate relationships between features that will not be apparent to traditional ML algorithms.

3) Addressing Class Imbalance:

a) Undersampling:

Undersampling the majority class involved randomly selecting a subset of instances from the majority class to balance the dataset. This helped mitigate bias towards the majority class and prevented the model from being overly influenced by the abundance of negative instances.

b) Oversampling Techniques (e.g., Random Oversampling and SMOTE):

Oversampling techniques such as random oversampling and SMOTE artificially increased the number of instances in the minority class. This was essential for enhancing the model's ability to learn from rare events and improving its predictive performance on the minority class.

4) Model Training and Evaluation:

a) Training-Testing Split:

The training and testing sets of the dataset were separated in order to assess the generalization performance of the model. The testing set was used to evaluate the model's performance on untested data, whereas the training set was used to train the model.

b) Training Models for Multiple Epochs:

Models were trained for multiple epochs, allowing them to iteratively learn from the data and adjust their parameters to minimize the loss function. This iterative process helps the models converge to optimal solutions and improve their predictive accuracy.

c) Evaluation Metrics:

Evaluation metrics were employed to evaluate the performance of the trained models, including accuracy, precision, recall, and F1 score. These metrics include information about how well the models categorize cases in terms of true positives, true negatives, false positives, and false negatives, among other classification-related features.

CHAPTER 5 – RESULT ANALYSIS

1. Logistic Regression: Since the data set is imbalanced with only 15% patient death, the results of the model without any class weight to offset this imbalance favours the live column in the confusion matrix.

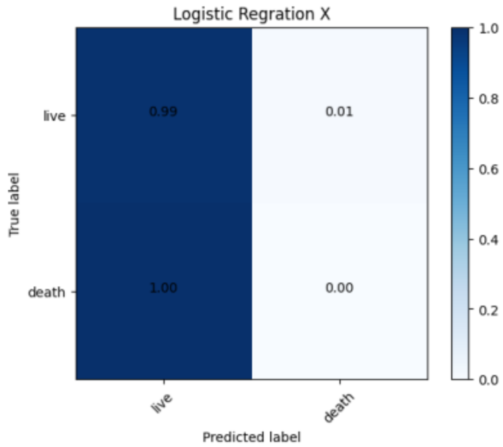

Fig. 1 Confusion matrix for Logistic Regression without class weights

By analysing the f1-score and confusion metrics we can say that our model is biased and only predict the live class. With imbalanced data set the model's f1-score for live class is approximately 100% while death class is nearly 0%.

In the X data set, as you can see above, a model forecasts almost all live patients to achieve an accuracy score of at least 85% to take full advantage of the size of the live patient population.We can increase the minority group of dataset deaths using SMOTE, a synthetic Minority Oversampling technique. To overcome database problems, it is also necessary to apply regularisation and class weights.

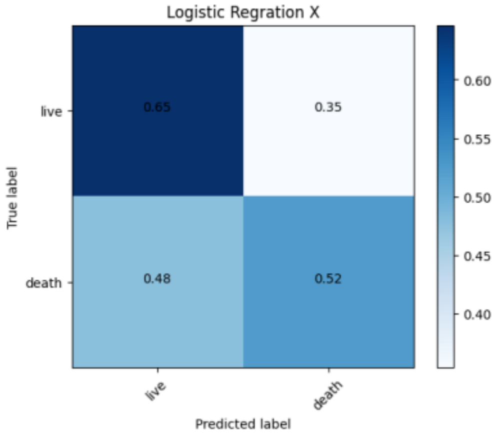

Fig. 2 Confusion matrix for Logistic Regression with class weights

We can only achieve 60% accuracy with these methods, and the f1 score difference between those two classes remains very large.

2. Random Forest Classification: The Random Forest, like logistic regression models, is better at predicting death with a class weight parameter to reconcile data. The plot reveals the cost of correct life predictions and benefits of correct death predictions with differing class weights.

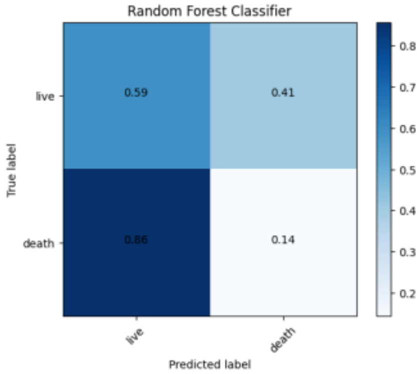

Fig. 3 Confusion matrix for Random Forest Classifier with class weights

In comparison with the log regression graphs, it is interesting to note that this model has a distinct pattern. Based solely on average precision, the log regression produces better results.

Deep Neural Networks: DNNs, which are capable of learning complicated patterns from dimensional data, show the most accurate predictions among all models.

However, deep learning models require large dataset more computation power and if the dataset is small or imbalanced it may be prone to overfitting, this requires careful regularisation and verification.

By analysing a graph and the classification report, we get very good f1 scores for live classes with no balancing of datasets, so it is not possible to improve model accuracy without addressing the dataset imbalance.

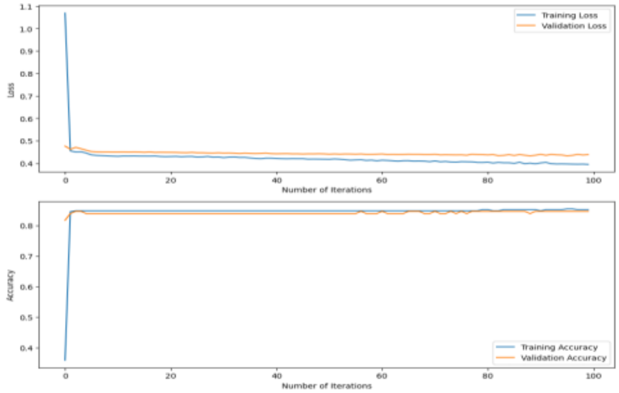

Fig. 4 Training and Validation Loss and Accuracy

3. Addressing Class Imbalance: To address the dataset imbalance, we use two widely used machine learning techniques.

    1. Undersampling - Deleting samples from the majority class.

       We can minimize the data imbalance and improve model performance through a technique called under sampling.

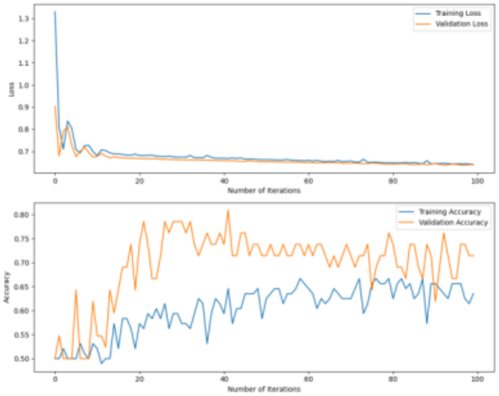

Fig. 5 Training and Validation Loss and Accuracy with Undersampling

       The 70% accuracy of the undersampling technique, as well as good f1 score f or both classes is provided

    2. Oversampling - Duplicating samples from the minority class

       a) Random oversampling technique: The probability sampling method used by the models, which enables them to predict train trains according to a selection of subsets of the dataset.

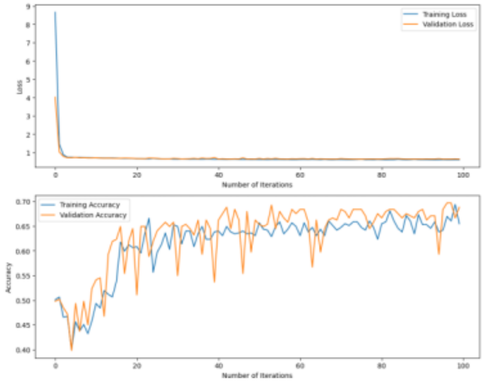

Fig. 6 Training and Validation Loss and Accuracy with Random oversampling

       With the good f1 score for both classes, random oversampling yields a 69% overall accuracy.

b) SMOTE: It generates the synthetic samples for the minority class to handle the class imbalance in dataset. It also helps to lessen the risk of overfitting by random oversampling technique.

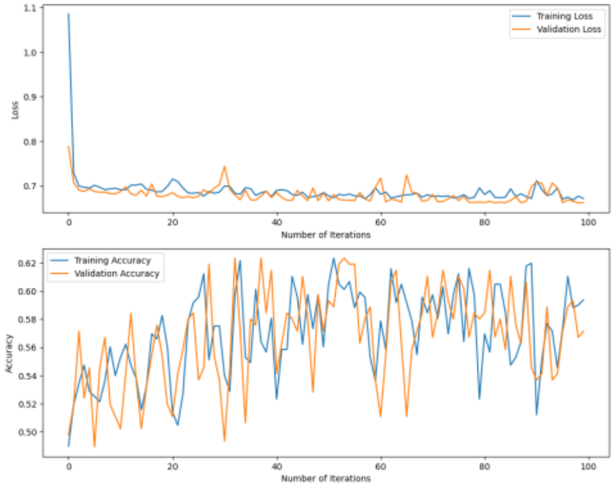

Fig. 7 Training and Validation Loss and Accuracy with SMOTE

SMOTE provides the 57% of overall accuracy with the moderate f1-sopre for both the classes.

Finally, we found that random oversampling was the best option for our model by analysing all three sampling methods and their accuracy and classification reports. The undersampling method is very accurate, but we do not use it because it deletes the existing data to remove the data imbalance that reduces the size of the dataset.

CHAPTER 6 - CONCLUSION & FUTURE WORK

This Project concludes on the note that predictive modelling at current stage still requires much better work in terms of dataset and ml methods. Whereas in current ones random sampling is the best method discovered for current problem while using deep neural network.

While this report certainly tries to use different models of machine learning and deep learning to better predict the risk of death in 1 year of the patient, it clearly remains somewhat biased due data inefficiencies. So, there remain several steps to be done in future.

6.1 Taking Additional Data and removing Biases

a) Dataset Expansion- The current dataset remains at the meagre length of 455 records neither suitable for deep neural networking or for in-depth analysis. So, the main goal in future is to collaborate with institutes associated with thoracic surgery and to increase dataset.

b) Features Expansion- The current columns range from Diagnosis to asthma and cough but despite this due the project pertaining its demand in medical field it fails to satisfy the need to find complete correlation between target and the features. Explore more features like biogenetic codes and lung prognosis to increase the precision.

6.2 Model Optimization

a) Hyperparameter Tuning- Further Hyperparameter tuning of logistic regression, random forest classifier and neural network can increase the accuracy of model prediction. New evaluation metrics like ROC curve and AUC also increase optimization of the models.

b) Ensemble Methods- Investigating ensemble methods, like model stacking or boosting techniques, which could help in improving the predictive performance by powering up the strengths of multiple models. It increases the flexibility to balance out false positives and false negatives prediction done by the model based on specific cost considerations.

6.3 Exploring the Deep Neural Network

a) Better Architecture- Current model is basic at work due to large data inefficiency. With better data, newer layers and better architecture can be used to bring out the prominence of deep neural network in prediction work.

b) Explainable AI- Newer theme of interpretability can be based on explainable AI which can help better understand the model and its decision and the inner process of the prediction.

6.4 Ethical and Societal Concerns

Ethicality and Societal Setbacks- Better ways to increase the precision include complete medical history of the patients which require permission of the patient and following ethical regulatory laws. Also, there is a need to increase checks and security of the data and the model to keep it from unauthorized and malevolent use.

References:

1. P. Peduzzi et al., "Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates," Journal of clinical epidemiology, vol. 48, no. 12, pp. 1503-1510, 1995.
2. E. H. Teh et al., "Predicting perioperative mortality in patients undergoing major vascular surgery," Anesthesia & Analgesia, vol. 120, no. 6, pp. 1499-1509, 2015.
3. L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.
4. J. Ghosh et al., "Predicting mortality in patients undergoing thoracic surgery," Journal of Thoracic Disease, vol. 7, no. 10, pp. 1817-1823, 2015.
5. G. Litjens et al., "A survey on deep learning in medical image analysis," Medical image analysis, vol. 42, pp. 60-88, 2018.
6. S. Wang et al., "Deep learning in radiology: An overview of the concepts and a survey of the state of the art," Medical Image Analysis, vol. 75, p. 102103, 2021.
7. J. Cruz et al., "Application of artificial neural networks in predicting early mortality after lung cancer surgery," European Journal of Cardio-Thoracic Surgery, vol. 56, no. 1, pp. 74-80, 2019.
8. J. Chen et al., "Artificial intelligence in thoracic imaging: state of the art," Clinical radiology, vol. 75, no. 1, pp. 7-18, 2020.
9. J. Choi et al., "Predictive modeling of thoracic surgery patient outcomes using artificial neural networks," Journal of thoracic disease, vol. 11, no. 11, pp. 4791, 2019.
10. R. Caruana et al., "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1721-1730.
11. Z. Obermeyer et al., "Predicting the future—big data, machine learning, and clinical medicine," New England Journal of Medicine, vol. 375, no. 13, pp. 1216-1219, 2016.

AI_Project_Report

ORIGINALITY REPORT

6%	4%	2%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Lakshmi Thara R, Bhavya Upadhyay, Ananya Sankrityayan. "YOLO V8: An improved real-time detection of safety equipment in different lighting scenarios on construction sites", Research Square Platform LLC, 2024 Publication	1%
2	vuir.vu.edu.au Internet Source	1%
3	Submitted to Gavilan College Student Paper	1%
4	listens.online Internet Source	1%
5	www.frontiersin.org Internet Source	<1%
6	ir.cut.ac.za Internet Source	<1%
7	www.coursehero.com Internet Source	<1%
8	pure.hva.nl Internet Source	

		<1%
9	Submitted to IUBH - Internationale Hochschule Bad Honnef-Bonn Student Paper	<1%
10	uir.unisa.ac.za Internet Source	<1%
11	"Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries", Springer Nature, 2019 Publication	<1%
12	www.ihe.net Internet Source	<1%
13	vtechworks.lib.vt.edu Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		