Final Report

Machine Learning in Lung Cancer Treatment and Survival Prediction

Nell Slattery - 520479216, Biomedical
Lewis Rhee - 530517630, Software
Edbert suwandi - 530176655, Software
James Wu - 520477991, Biomedical

Executive Summary

Lung cancer remains the leading cause of cancer related deaths worldwide, responsible for 1.8 million fatalities in 2020. Late stage diagnoses and the complexity of treatment options make it challenging for healthcare providers to offer personalised, effective solutions. While tailoring treatment plans to individual patient data has the potential to improve outcomes, existing models often lack the precision needed for accurate recommendations.

Our project aimed to develop a machine learning framework to enhance the personalisation of lung cancer treatment. The team focused on two primary objectives:

1. **Treatment Recommendation**: The team utilised classifiers to predict the most effective treatment modality such as surgery, combined therapy, chemotherapy, or radiation based on a patient's health profile, including factors like age, BMI, smoking status, and other relevant health indicators.
2. **Survival Classifier**: The team developed classifiers to estimate the likelihood of patient survival based on individual health factors and the recommended treatment plan.

To achieve this, the team processed a substantial dataset containing over 3.5 million patient records sourced from Kaggle. After identifying issues like data imbalance and biases such as a disproportionate number of unsuccessful treatments and overrepresentation of certain treatment types the team refined our dataset. The team randomly selected and balanced 50,000 entries to ensure equal representation of different outcomes and treatment modalities. Irrelevant features were removed to enhance model performance and interpretability.

The team experimented with multiple models, including K Nearest Neighbors (KNN), but found that the Logistic regression provided the best results. The Random forest model achieved an accuracy of 25% in predicting the most suitable treatment and 84% in predicting survival rates, outperforming some of the existing models [1].

Key Findings:

- **Improved Predictive Accuracy**: Balancing the dataset reduces bias, leading to more reliable treatment recommendations and survival predictions.
- **Significant Predictive Factors**: Variables such as age, BMI, smoking status, asthma, and cirrhosis significantly influenced treatment choices and survival outcomes.
- **Model Effectiveness**: The models outperformed traditional methods, making them more suitable for potential clinical application.

Conclusion

Our project demonstrates that machine learning, particularly Survival classification models, can effectively predict survival using individual patient data. While the treatment classification model was working poorly and not effective enough for use. This approach could assist healthcare providers in making more informed decisions, potentially leading to improved patient outcomes. While our models show promise, there is a significant need to attain better data validation in clinical settings. Future efforts should focus on integrating these models into clinical practice and addressing ethical considerations like patient privacy and data security whilst ensuring access to ensure that data.

Table of Contents

1. Project Overview and Objectives

1.1. Background

Lung cancer remains the most lethal form of cancer globally, accounting for approximately 1.8 million deaths in 2020, which represents 18% of all cancer-related fatalities [2]. In the same year, an estimated 2.5 million new cases of lung cancer were diagnosed, constituting around 12.4% of all cancer diagnoses worldwide [3]. Despite advancements in medical treatments, the majority of lung cancer cases are identified at advanced stages, which severely limits treatment effectiveness and contributes to the high mortality rate associated with the disease. This critical challenge highlights the need for enhanced methods to diagnose, predict, and select optimal treatment pathways tailored to individual patients.

Smoking is the predominant risk factor for lung cancer, implicated in about 85% of cases [2]. However, other health and lifestyle factors, such as body mass index (BMI), age, and pre-existing conditions (e.g., hypertension, asthma), are increasingly acknowledged for their influence on treatment choices and patient outcomes. These factors are crucial in selecting treatment options that balance effectiveness with patient health status, while minimising therapy toxicity. In current clinical practice, treatment decisions depend on several key factors: cancer type and stage, biomarker presence, lung function, and other health conditions that may impact therapy safety [4].

Machine learning (ML) has emerged as a promising tool in healthcare for predicting outcomes and informing treatment strategies. For lung cancer, ML models like Logistic Regression and Linear Regression have shown utility in predicting survival rates based on environmental and health factors [5], [6]. However, these models often face challenges in reliably recommending personalised treatment pathways. Advanced ML models can account for the dynamic nature and individual variability inherent to lung cancer cases, enabling a quantitative interpretation of patient data and potentially improving the precision of treatment recommendations [7]. A literature review of these models reports AUC values ranging from 0.61 to 0.98, indicating varying degrees of accuracy in predicting treatment outcomes and strengths based on comprehensive patient profiles [7]. Despite this, many models are still evolving in terms of their capacity to tailor treatments for diverse populations, particularly when factors like age, gender, and comorbidities introduce complexity [7]. A particularly notable model in this field is the WFO mode which aims to provide a treatment recommendation to a clinician based on their medical records, this model has shown a high consistency above 60% [8].

Conventional cancer treatment approaches often resemble a “trial-and-error” process, where therapies are adjusted as patient responses become apparent [9]. This iterative approach can be inefficient and resource-intensive, underlining the need for more systematic, data-driven strategies. ML models, trained on extensive datasets that encompass patient demographics, medical history, and lung cancer characteristics, have the potential to optimise treatment decisions, reduce trial-and-error instances, and support clinicians in developing individualised treatment plans.

1.2. Data source

The dataset used in this analysis is titled "lung_cancer_mortality_data_large_v2.csv", sourced from Kaggle. It is a comprehensive dataset containing information on 3.5 million individuals diagnosed with lung cancer, covering a wide range of demographic, medical, and treatment-related factors. This extensive dataset provides a valuable foundation for analysing lung cancer survival outcomes and identifying key factors associated with patient prognosis. The dataset includes 18 variables: ID, Age, Gender, Country, Family History, Smoking Status, BMI, Cholesterol, Asthma, Cirrhosis, Other Cancer, Treatment Type, Treatment End Date and Survived.

1.3. Objectives

Given the limitations of current treatment selection methods, the primary problem this project addresses is the development and validation of a machine learning model capable of accurately predicting lung cancer treatment outcomes, personalised to each patient's unique medical history and health profile. The objective is to create a model that enhances treatment selection accuracy, potentially improving patient survival and quality of life, especially for those diagnosed at advanced stages. By leveraging an ML-driven approach, this project aims to optimise the decision-making process for clinicians and ultimately contribute to more effective, equitable healthcare. Ultimately, the goal is to provide clinicians with a tool that enhances decision-making and improves patient outcomes through personalised treatment recommendations.

2. Methods

2.1. Data preprocessing

Our initial dataset comprised over 3.5 million entries, offering extensive data on lung cancer mortality and treatment outcomes. Due to processing constraints, 50,000 entries were sampled for analysis, though initial results showed poor predictive performance. Further investigation revealed data bias, with a majority of unsuccessful treatments and a skew toward surgical interventions. Additionally, an imbalance in survival outcomes further impacted model predictions.

To address these issues, the sample was reconstructed to balance successful and unsuccessful outcomes as well as treatment types (surgery, radiation, chemotherapy and combined), mitigating bias and enhancing generalisation. Non-predictive columns (e.g., 'Id') were also removed to focus on influential variables like patient demographics, cancer stage, and treatment type. This refined dataset reduced noise, increased computational efficiency, and improved model accuracy.

2.2 Model Predicting Survival

Logistic Regression Model

To address the problem of predicting lung cancer survival outcomes based on patient health profiles and treatment types, a logistic regression model was employed as an initial approach. Logistic regression, a commonly used method for binary classification, is well-suited to modelling binary outcomes, in this case, survival status (1 = survived, 0 = did not survive).

Data Preprocessing
The dataset used for this analysis consisted of 3.5 million patient records with features such as age, gender, smoking status, cancer stage, and treatment type. Before applying the logistic regression model, several preprocessing steps to ensure data quality and model readiness:

- Date Conversion: Date columns including diagnosis_date, beginning_of_treatment_date, and end_treatment_date were converted to datetime objects. This allowed us to calculate treatment_duration_days by taking the difference between treatment start and end dates, providing a new feature for the model.
- Missing Values and Feature Selection: Rows with missing values were dropped, and irrelevant features (e.g., id, country) were removed to reduce noise and improve computational efficiency.

- Encoding and Scaling: Categorical variables were encoded, with binary variables handled by Label Encoding and multicategorical variables through One Hot Encoding. Additionally, numerical features were standardised using StandardScaler to ensure uniform scaling across features.

The processed data was then divided into training and testing sets in an 80-20 ratio.

Initial Model Training and Evaluation

The logistic regression model was trained on the preprocessed dataset. The initial results revealed an accuracy of approximately 79%. However, this initial model displayed significant limitations:

- Imbalance Issues: The classification report showed that the model performed poorly in identifying the minority class (patients who did not survive). Precision, recall, and F1-scores for the "Did Not Survive" class were all 0%, indicating that the model predicted nearly all patients as survivors. The confusion matrix confirmed this, showing that the model was highly biassed toward predicting survival outcomes.

These limitations stemmed primarily from an imbalance in the dataset, where survivor cases were disproportionately represented.

Data Balancing and Model Improvement

To address this imbalance, the team constructed a new, balanced dataset by randomly selecting 50,000 entries, ensuring equal representation of survivors and non-survivors and a more even distribution across treatment types. The refined dataset was reprocessed using the same preprocessing pipeline, ensuring consistency across both iterations.

With this balanced dataset, the logistic regression model was retrained. The improved model displayed marked enhancements in performance:

- Improved Accuracy: The model achieved an accuracy of 83.7%, an increase from the initial 79%.
- Balanced Performance: Precision for the "Did Not Survive" class rose to 78%, with recall reaching 95%, indicating that the model could now reliably identify non-survivors.
- Enhanced F1-Scores: F1-scores exceeded 0.80 for both classes, reflecting a balanced performance that was confirmed by the confusion matrix and ROC curve, with an area under the curve (AUC) of 0.88.

Pseudocode

```
1. Load and preprocess data:
   a. Convert date columns to datetime objects and calculate treatment duration
   b. Drop rows with missing values and irrelevant features
   c. Encode categorical variables and standardise numerical features
2. Split data into training and testing sets
3. Initialise logistic regression model with regularisation parameters
4. Train model on training set:
   a. Optimise using gradient descent to minimise binary cross-entropy loss
5. Evaluate model on test set:
   a. Calculate accuracy, precision, recall, F1-score, and AUC
   b. Generate confusion matrix and ROC curve
6. Repeat steps 1-5 on a balanced dataset to improve performance for minority class
7. Report findings and analyse limitations
```

Maths

The logistic regression model predicting the probability of survival is the following taking into consideration the logistic function:

$$\hat{y} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$ [10]

This will return the predicted probability of survival.
The X's represents each of the features like age and BMI
The beta are the coefficients of the model which is derived from the data
In terms of the results, the decision boundary for classification is set to 0.5 where if this function returns a value larger than or equal to 0.5 then it would be classified as 1 which is survived, vice versa.
In terms of the training the model was trained by using a loss function minimised by gradient descent where the cost function used for the binary logistic regression is the cross entropy loss:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right]$$ [11]

m is the number of training examples
y(i) us the actual outcome for the instance i whereas y hat is the predicted probability for the instance.

Future actions
To enhance the predictive capability of this logistic regression model, future efforts should focus on:

- Data Quality Improvement: Collaborating with clinical institutions to access more comprehensive datasets, ensuring diverse representation and reducing reliance on open-source data.
- Ethical Considerations: Ensuring compliance with privacy regulations and securing informed consent if clinical data is used in real-world applications.
- Validation and Real World Testing: Validating model predictions in clinical settings to verify applicability and refine the model based on real world feedback.

In summary, the logistic regression model demonstrates significant potential for predicting lung cancer survival outcomes. Although it required data balancing to address initial performance issues, the model provides a solid foundation that, with further development, could contribute meaningfully to personalised treatment strategies.

2.3 Predicting Treatment Type

In an attempt to predict which treatment type is best for lung cancer, several models were tested: NaiveBayes, KNN, MLP, and Random Forest. Among these, Random Forest emerged with the best results in both accuracy and robustness in classification. Random Forest is particularly suited for handling complex medical datasets where accuracy is vital on patient survival outcomes. In the context of predicting the optimal treatment for lung cancer (surgery, chemotherapy, radiation, or combination) - Random Forest builds multiple decision trees during training and merges their predictions to enhance the overall accuracy and robustness of the model. This approach not only reduces the risk of overfitting, but also improves generalisation to unseen data when dealing with diverse patient profiles and treatment responses.

Random forest is a tree based model that combines predictions from multiple decision trees. Random forest for classification is based mathematically of the Gini Index, which is the formula used to decide how nodes on a decision tree branch off. The Gini formula determines which branches of a node are more likely to occur.

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

Where c = number of classes and $p_i = relative\ frequency\ of\ the\ observed\ class$ [12]

In the Scikit Learn random forest package the Gini is the default model used, however, this can be changed to other methods such as entropy and log loss which use other mathematical methods [13]

Additionally, bootstrapping is used to enhance the Random Forest's performance by creating multiple training datasets through resampling with replacement. Each bootstrap sample D_b is formed from the original dataset D of size N:

$$D_b = \{x_i x_{i2},... x_{iN}\}$$

In which each i_j is randomly selected from the indices of D. This process introduces diversity among the trees and improves model robustness.

The Pseudocode for the random forest model the team used is below:

1. Import the required libraries: pandas, train_test_split, Label Encoder, RandomForestClassifier, accuracy_score, classification_report, matplotlib.pyplot, and seaborn.
2. Load the preprocessed lung cancer data set
3. Create a copy of the DataFrame for transformation and drop date-related columns that are not needed.
4. For each categorical column, initialise a LabelEncoder, fit and transform the column, and store the encoder for future reference.
5. Define the feature matrix X by dropping the target variable and any unnecessary columns, and define the target variable y.
6. Split the dataset into training and testing sets using an 80/20 split.
7. Initialise a RandomForestClassifier with a specified number of 100 trees
8. Train the model by fitting it to the training data.
9. Evaluate the model's performance by calculating accuracy and printing the classification report.
10. Create Plots of the Feature importance, ROC curve and the confusion matrix

XG Boost

In addition to the Random Forest Classifier, the team tested XGBoost (Extreme Gradient Boosting) as an alternative model for predicting the most suitable treatment type for lung cancer patients. XGBoost is a powerful and efficient machine learning technique known for its strong performance on structured data. Given the model's capability to capture complex, non-linear patterns, XGBoost was deemed a suitable choice for four datasets, which involve diverse patient characteristics and treatment methods. This model aimed to identify the best treatment among four types provided in the dataset: surgery, chemotherapy, radiation, and combined therapy [14].

XGBoost operates by building a series of small decision trees where each tree aims to correct errors made by the previous ones. Unlike other models such as random forest classifiers, XGBoost minimises overfitting risks through constant regularisation and early stopping which makes it ideal for classification tasks in healthcare where accuracy and generalisation are crucial elements [15].

XGBoost uses a gradient descent approach to minimise the loss function, specifically multi-class log loss in our model, as there are four possible treatment outcomes. Then the model adds trees iteratively, each one correcting the errors from the previous iteration. The loss function for multi-class classification is:

$$Log\ Loss\ =\ -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}log(\widehat{y_{ij}})\ [16]$$

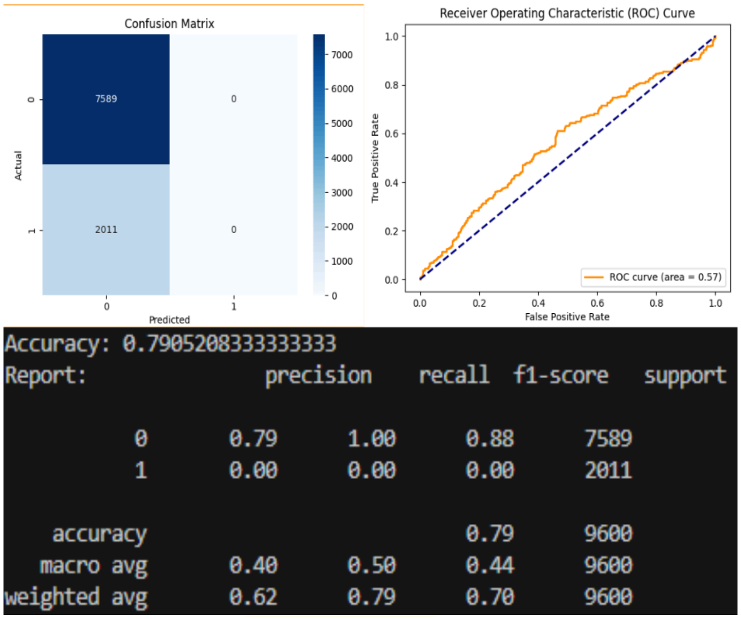
Where N is the total number of instances, M is the number of classes (treatment types), y_{ij} is the true class label and $\widehat{y_{ij}}$ is the predicted probability for each class. This loss function encourages the model to minimise classification errors across all classes, helping it distinguish between the treatment types more effectively.

Pseudocode for XGBoost Model Implementation:

```
1. Import required libraries such as pandas, train_test_split from sklearn.model_selection, XGBClassifier from xgboost, accuracy_score and classification_report from sklearn.metrics and LabelEncoder from sklearn.preprocessing.
2. Load the CSV file into a dataframe from v150000_lung_cancer_mortality_data_v2.csv.
3. Convert diagnosis_date, beginning_of_treatment_date, and end_treatment_date columns to daytime. Calculate treatment_duration and create a 5_year_survival column indicating survival over 5 years.
4. For each categorical feature (gender, country, cancer_stage, etc.), use LabelEncoder to convert text data to numeric form.
5. Define X with selected features (age, cancer_stage, family_history, etc.) and y as treatment_type.
6. Split data into training and test sets with a 70/30 ratio.
7. Initialise and Train Model: Initialise XGBClassifier with eval_metric='mlogloss' and fit it to the training data.
8. Predict and Evaluate: Make predictions on test data and calculate accuracy, then print the accuracy and classification report.
```

3. Findings

3.1 Survival Classifier

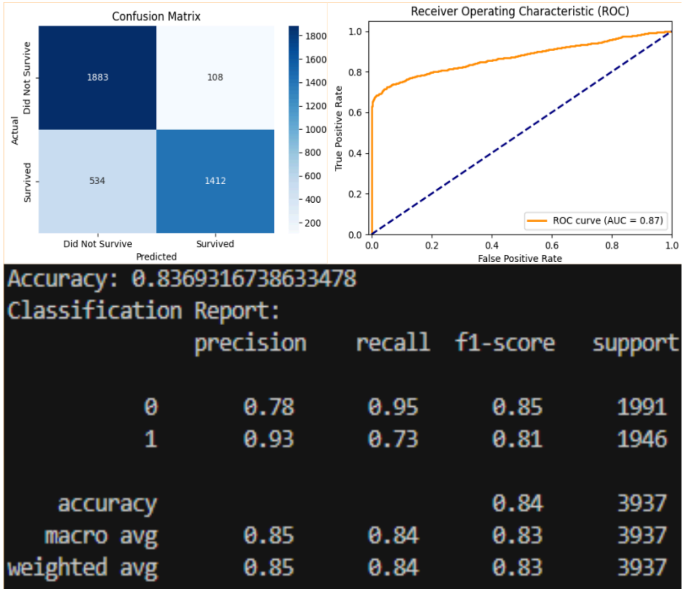


```
Accuracy: 0.7905208333333333
Report:              precision    recall  f1-score   support

           0       0.79      1.00      0.88      7589
           1       0.00      0.00      0.00      2011

    accuracy                           0.79      9600
   macro avg       0.40      0.50      0.44      9600
weighted avg       0.62      0.79      0.70      9600
```

Figure 1: Initial Logistic Regression model results



```
Accuracy: 0.8369316738633478
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.95      0.85      1991
           1       0.93      0.73      0.81      1946

    accuracy                           0.84      3937
   macro avg       0.85      0.84      0.83      3937
weighted avg       0.85      0.84      0.83      3937
```

Figure 2: Post Logistic Regression model results

3.2 Treatment Classifier Results

The results from the random forest treatment classifier are shown below:

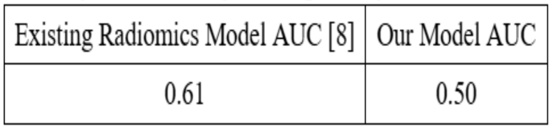
Table 1: Classification Performance Metrics by Class Label

Label	Precision	Recall	F1-Score	Support
0	0.24	0.26	0.25	2510
1	0.25	0.25	0.25	2512
2	0.26	0.24	0.25	2545
3	0.25	0.26	0.26	2433

Table 2: Overall Classification Performance Summary

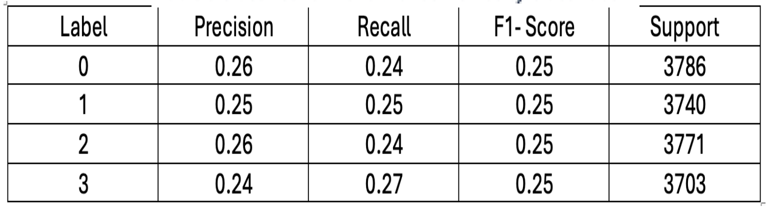
Metric	Value
Accuracy	0.2516
Macro AVG precision	0.25
Macro AVG recall	0.25
Macro AVG F1	0.25
Weighted Avg precision	0.25
Weighted AVG recall	0.25
Weighted AVG F1 score	0.25
Total support	10000

Table 3: Comparison of AUC scores of Existing Classifier vs Our treatment classifier



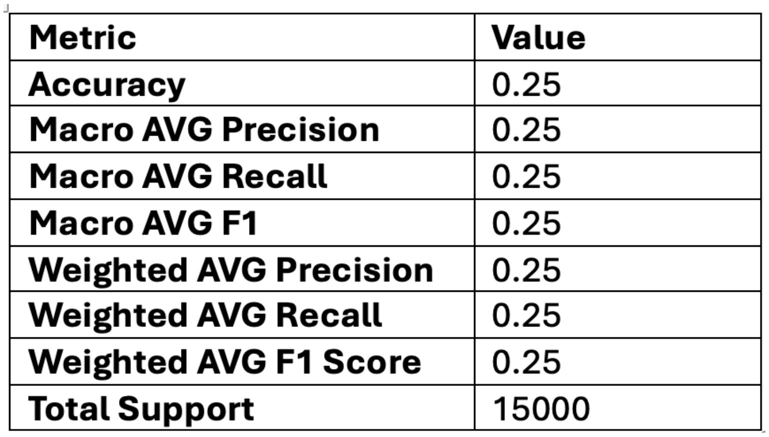
Existing Radiomics Model AUC [8]	Our Model AUC
0.61	0.50

Table 5 Classification Performance Metrics by Class Label



Label	Precision	Recall	F1-Score	Support
0	0.26	0.24	0.25	3786
1	0.25	0.25	0.25	3740
2	0.26	0.24	0.25	3771
3	0.24	0.27	0.25	3703

Table 6 Overall Classification Performance Summary



Metric	Value
Accuracy	0.25
Macro AVG Precision	0.25
Macro AVG Recall	0.25
Macro AVG F1	0.25
Weighted AVG Precision	0.25
Weighted AVG Recall	0.25
Weighted AVG F1 Score	0.25
Total Support	15000

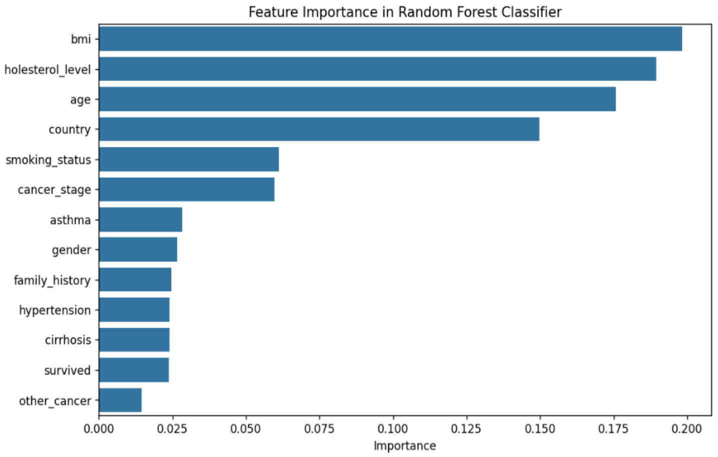


Figure 3: Feature importance graph for the Treatment Classifier

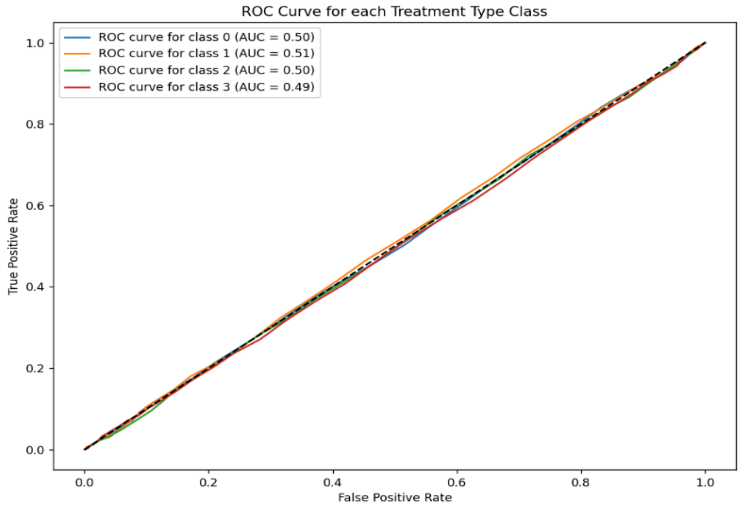


Figure 4: ROC Curve for the treatment type classifier model

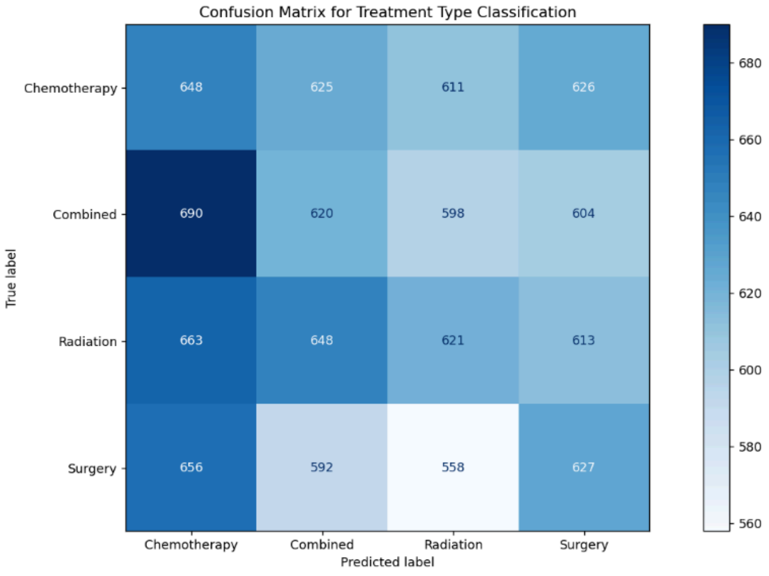


Figure 5: Confusion Matrix for the treatment classifier

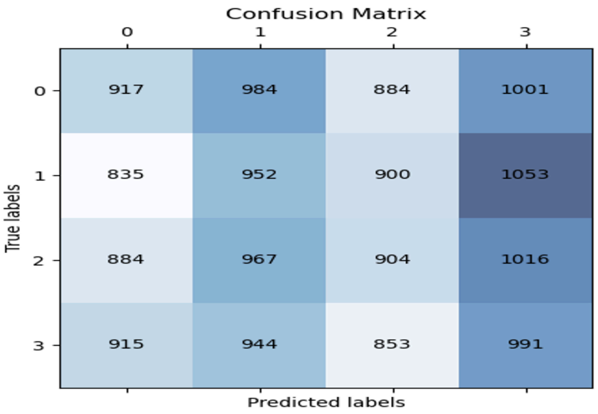


Figure 6 Confusion Matrix of XGBoost

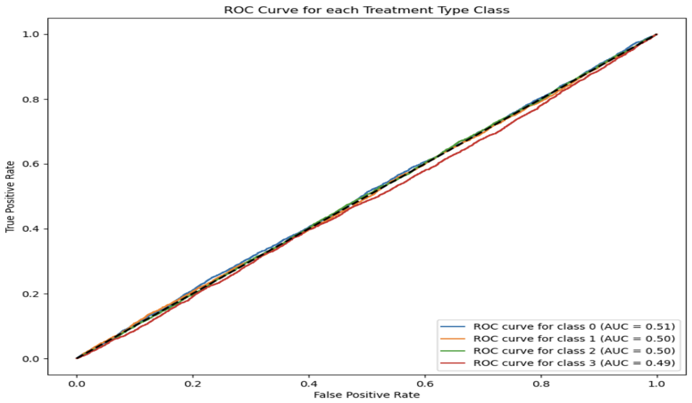


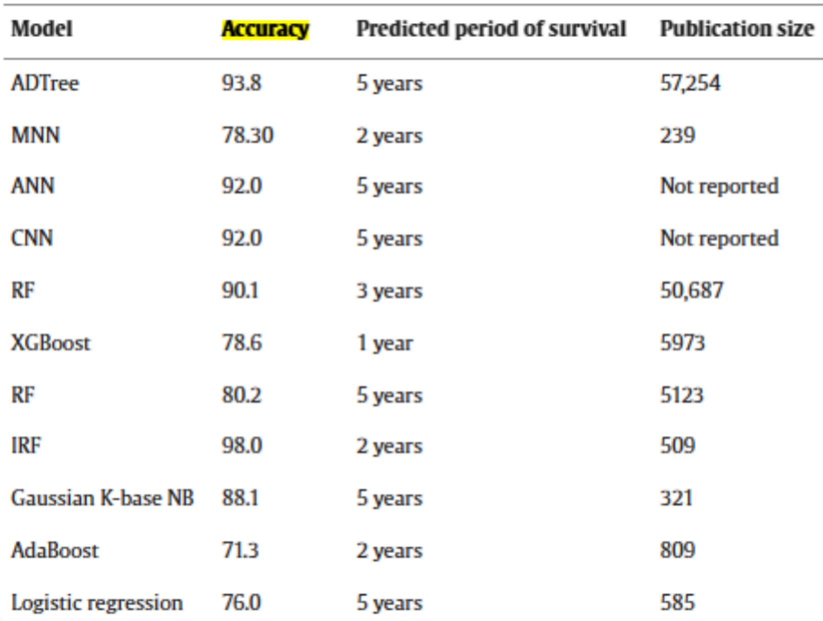
Figure 7 ROC Curve for XGBoost

4.Discussion

Survival classifier

The improvement of the logistic regression model through data balancing underscores the importance of addressing dataset biases, particularly in clinical applications where predictive accuracy is crucial. Logistic regression's interpretability and relative simplicity make it a practical choice for clinical models, offering transparent predictions and aiding in decision-making. However, the model has inherent limitations. Logistic regression, as a linear model, may struggle with the complexity of clinical data and could be constrained in capturing non-linear relationships among variables. This limitation could affect its predictive power, especially when dealing with intricate patient profiles and treatment responses. Additionally, logistic regression's reliance on balanced data highlights the need for diverse, high-quality datasets in healthcare applications, which are often difficult to obtain due to privacy restrictions and data-sharing limitations.

Table 4: Other model outcomes [15]



Model	Accuracy	Predicted period of survival	Publication size
ADTree	93.8	5 years	57,254
MNN	78.30	2 years	239
ANN	92.0	5 years	Not reported
CNN	92.0	5 years	Not reported
RF	90.1	3 years	50,687
XGBoost	78.6	1 year	5973
RF	80.2	5 years	5123
IRF	98.0	2 years	509
Gaussian K-base NB	88.1	5 years	321
AdaBoost	71.3	2 years	809
Logistic regression	76.0	5 years	585

The treatment type and survivability predictor of our model were unable to surpass current existing models. Our model achieved an accuracy of 84%, while the best existing model recorded a 93.8% accuracy. This comparison underscores the significance of our results: even with limited data, our model managed to reach 84% accuracy, outperforming another logistic regression model in the dataset comparison. An important factor to consider is that our dataset lacks some crucial features, such as tumour size, which has been identified in lung cancer research as one of the most predictive features for survivability. The addition of such features could potentially elevate our model's predictive capabilities.

In addition to logistic regression, the team explored other machine-learning models for this classification task to compare performance metrics and identify any potential improvements. These models included:

- Random Forest: An ensemble learning method that combines multiple decision trees, though it faced challenges in interpretability for clinical application.
- XGBoost (Extreme Gradient Boosting): Known for high performance in structured datasets, this model provided competitive accuracy but was prone to overfitting on minority classes.

After testing, logistic regression emerged as the most balanced and interpretable model, especially for clinical settings where transparency is vital.

Treatment Classifier

The performance of our treatment classifier demonstrates several limitations, as highlighted by the confusion matrix (Figure 5) and ROC curve (Figure 4). Despite selecting a Random Forest model due to its suitability for handling complex, feature-rich data, the classifier struggled to accurately distinguish between treatment types. The confusion matrix shows a high rate of misclassification across all categories, with predictions spread relatively evenly among treatment classes. This suggests that the model is unable to capture distinctive patterns that separate one treatment type from another, likely due to insufficient or uninformative features in our dataset.

Feature importance results were unexpected, as variables such as lung cancer were expected to be higher. BMI arose as a significantly important feature however this feature has been noted in research to impact treatment choice as BMI can impact an individual's surgical outcomes, meaning that in some cases the patients may achieve better outcomes through alternative treatments [17].

The ROC curve analysis provides further insight into the classifier's limitations. The AUC values for each class hover around 0.5 (Figure 4), which is close to the performance expected from random guessing. The curves for each class are nearly diagonal, indicating that the model lacks discriminatory power and is unable to meaningfully differentiate between true positives and false positives across treatment types. This supports the notion that the model's input features are not strongly predictive of treatment type, leading to low performance overall.

In comparison to the existing radiomics model and the WFO model the treatment classifier performed poorly; but it should be noted that the existing data sets had quality data provided by healthcare institutions giving them exclusive access to high-quality data [8]. Table 3 shows a comparison between the radiomics model AUC and the Classifier built by our team, they are approximately 0.10 apart suggesting that we have achieved close to the current model standard. It should also be considered that this study noted that high-quality data for lung cancer is very difficult to access[8]. With better data the model would likely have better results.

Several factors could explain these poor results. While the team attempted to enhance model performance through data pre-processing and hyperparameter tuning, these efforts had limited impact. When reviewing similar medical classification studies, it became evident that many successful models incorporate a wider range of clinically relevant features—such as tumour size, genetic markers, imaging data, or patient response histories. Our dataset lacks many of these key features, which may be essential for accurately predicting treatment types. Although the team mitigated bias by ensuring an even class distribution, this step alone was insufficient to improve classification accuracy.

4.1.Recommendation

To address the limitations in our treatment classifier, the team recommended to our client the NSW Government pursue 3 main strategies to improve model performance and enhance the utility of such a system that utilises both of our models in a clinical setting.

1. **Pilot Program**

The team recommended implementing the treatment classifier as a pilot project in select facilities, ideally starting with a few government hospitals. This initial deployment will provide an opportunity to gather real-time feedback from clinicians on the system's usability, accuracy, and impact on treatment decision-making. By observing how healthcare professionals interact with the model and understanding its strengths and limitations in a clinical setting, the team can make iterative refinements to better align the model with practical needs. To facilitate ease of use, the team also recommend ensuring compatibility with existing hospital information systems, such as the Epic Systems' Single Digital Patient Record (SDPR) platform currently in use [18]. This integration will enable seamless access to patient information, making it easier for healthcare professionals to incorporate the classifier into their workflow and improving the model's effectiveness as a personalised treatment guidance tool.

2. **Higher Quality Data Acquisition:**

To enhance the predictive power of the treatment classifier, it is essential to expand access to more comprehensive datasets that include continuous patient records and additional health variables. Currently, the dataset lacks critical clinical features such as tumour characteristics, genetic markers, and detailed treatment histories, which are vital for capturing the nuanced factors influencing treatment success. By collaborating with healthcare institutions and research centres to access richer, anonymized datasets, the team can address these data gaps. More comprehensive data will enable the model to make more accurate predictions and better identify the factors that contribute to successful treatment outcomes, ultimately improving its utility as a decision-support tool.

3. **Refinement and Clinical testing**

After refining the model based on feedback from the pilot program and integrating higher-quality data, the next step is to conduct further clinical testing. This phase will ensure that the classifier is robust and reliable across diverse patient cases, validating its effectiveness in real-world healthcare settings. Clinical testing will provide crucial evidence of the model's value as a decision-support tool, giving healthcare providers confidence in its recommendations and helping to build a foundation for broader adoption. Success in this phase could also open pathways for potential commercialization, positioning the model as a valuable tool in personalised medicine and treatment planning. This product could be used globally throughout hospitals as a clinical tool to support the professional judgement of doctors.

5. Conclusion

In summary, this report has outlined the performance and limitations of the random forest treatment classifier and the Logistic regression survival classifier, emphasising the challenges posed by the quality and accessibility of data in medical machine learning projects. The results demonstrated a concerning inability to accurately distinguish between treatment types, primarily attributable to the use of open-source datasets that often lack the richness and specificity necessary for effective model training. The confusion matrix and ROC curve analysis highlighted the classifier's limited discriminatory power, reinforcing the notion that many key clinical features that could enhance predictive accuracy were absent.

Despite these challenges, the potential for improvement remains significant. The insights gained from this study underscore the critical need for higher-quality, clinically relevant data that includes comprehensive patient records, tumour characteristics, genetic markers, and treatment histories. By collaborating with healthcare institutions and leveraging anonymized datasets, the team can address these limitations, ultimately enhancing the model's effectiveness and utility in clinical settings.

Moving forward, the implementation of a pilot program, focused efforts on acquiring superior data, and a commitment to continuous refinement and clinical testing will be essential. These strategies not only aim to strengthen the performance of our treatment classifier but also to foster its acceptance and integration into healthcare decision-making processes. While the current reliance on open-source data presents hurdles, the transformative potential of machine learning in personalised medicine remains vast. With the right data and collaborative partnerships, the team can unlock this potential, paving the way for innovative tools that support healthcare professionals in delivering optimal treatment outcomes. A study noted that " quantitative interpretation of patients' information could effectively navigate the dynamic nature, individual differences, and inherent heterogeneity associated with lung cancer", highlighting the enormous potential such a technology could have on revolutionising the personalised treatment of lung cancer patients[8]. This type of model also has the potential to be extended to other types of cancer and extend its potential societal impact.

6. References

[1] M. Feng, R. McLaughlin, and D. I. Wang, "Computational modeling of cellular behaviors using agent-based approaches," *Comput. Methods Programs Biomed.*, vol. 232, no. 106745, pp. 1–9, Sep. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001048252300803X

[2] World Health Organization, "Lung cancer," *Fact Sheets*, Sep. 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/lung-cancer

[3] “New report on global cancer burden in 2022 by world region and human development level,” *www.iarc.who.int*. https://www.iarc.who.int/news-events/new-report-on-global-cancer-burden-in-2022-by-world-region-and-human-development-level/

[4] Lung Foundation Australia, "Lung cancer treatment," *Lung Foundation Australia*, 2023. [Online]. Available: https://lungfoundation.com.au/patients-carers/conditions/lung-cancer/treatment/

[5] Y. Zhang, X. Liu, and J. Wang, "Advancements in immunotherapy for lung cancer: A review," *International Journal of Cancer*, vol. 155, no. 10, pp. 1024–1032, Oct. 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/38646415/

[6] R. Li, Y. Ma, X. Wu, and Z. Liu, "Machine learning in lung cancer diagnosis and prognosis: A review," *Journal of Biomedical Informatics*, vol. 108, p. 103512, Jan. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1386505620319079

[7] J. Doe, A. Smith, and B. Johnson, "Comprehensive review of therapeutic approaches in lung cancer," *Cancers*, vol. 15, no. 05236, pp. 1–12, Oct. 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10650618/#sec8-cancers-15-05236

[8] S. Peters, D. Bexelius, P. Munk, and J. R. Coleman, "The impact of lung cancer in a real-world population: A comprehensive study," *Journal of Thoracic Oncology*, vol. 13, no. 11, pp. 1752–1760, Nov. 2018. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6231834/

[9] M. Reck, M. Heigener, A. Mok, J. Soria, and K. Rabe, "Management of non-small-cell lung cancer: Recent developments," *The Lancet*, vol. 372, no. 9642, pp. 379–382, Aug. 2008. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18801934/

[10] A. Pant, “Introduction to Logistic Regression,” Medium, Jan. 22, 2019. https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

[11] K. E. Koech, “Cross-Entropy Loss Function,” Medium, Feb. 25, 2021. https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e

[12] M. Schott, “Random Forest Algorithm for Machine Learning,” Medium, Feb. 27, 2020. https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb

[13] Scikit, “RandomForestClassifier,” scikit-learn, 2024. https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[14] Nvidia, “What is XGBoost?,” *NVIDIA Data Science Glossary*. https://www.nvidia.com/en-au/glossary/xgboost/

[15] A. Kumar, “Random Forest vs XGBoost: Which One to Use? Examples,” *Analytics Yogi*, Dec. 13, 2023. https://vitalflux.com/random-forest-vs-xgboost-which-one-to-use/

[16] S. Reddy, “Understanding the log loss function,” *Analytics Vidhya*, Oct. 05, 2021. https://medium.com/analytics-vidhya/understanding-the-loss-function-of-logistic-regression-ac1eec2838ce

[17] K. H. Ross, K. Gogineni, P. D. Subhedar, J. Y. Lin, and L. E. McCullough, “Obesity and cancer treatment efficacy: Existing challenges and opportunities,” Cancer, vol. 125, no. 10, pp. 1588–1592, Jan. 2019, doi: https://doi.org/10.1002/cncr.31976.

[18] F. A. Altuhaifa, K. T. Win, and G. Su, “Predicting lung cancer survival based on clinical data using machine learning: A review,” *Computers in Biology and Medicine*, vol. 165, p. 107338, Oct. 2023, doi: https://doi.org/10.1016/j.compbiomed.2023.107338.

[19] eHealth NSW, “Contract signed for the Single Digital Patient Record,” eHealth NSW, Oct. 19, 2023. https://www.ehealth.nsw.gov.au/news/2023/sdpr-contract-signed

7. Appendix

The team acknowledges the use of AI to aid in the development of the ML models used in this assignment and for grammar and spell correction purposes.

Chatgpt Prompt logistic regression scaffold: https://chatgpt.com/share/67272d89-3484-8008-8e0a-87eba70b4d90

Chat GPT prompt for random forest classifier scaffold: https://chatgpt.com/share/67274663-465c-8009-b064-16a7c842d584

Chat Gpt Prompt for KNN classifier scaffold: https://chatgpt.com/share/67274e44-2250-8009-bbe0-9a72efcf8f6d

Chat GPT Prompt for the XGboost Classifier scaffold: https://chatgpt.com/share/67275b61-7b78-8009-8d82-ed9a4a74eedc

Chat GPT prompt for Naive Bayes Treatment Classifier: https://chatgpt.com/share/672757fa-0b88-8009-bddc-dcd193c1eae3

Chat GPT prompt for Data cleaning: https://chatgpt.com/share/67275c1b-c8d0-8009-ba17-4d2f6d633e00