WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Risk Prediction of Lung Cancer in Rheumatoid Arthritis Patients using ML

Prof. P. Rajeshwari[1], Ms. R. Mahitha[2*], Mr. I. Hariharan[3*]

[1]Assistant Professor, Department of Software Systems, Sri Krishna Arts and Science College, Coimbatore, India rajeshwarip@skasc.ac.in
[2]Student, Department of Software Systems, Sri Krishna Arts and Science College, Coimbatore, India mahithar24mcs038@skasc.ac.in
[3]Student, Department of Software Systems, Sri Krishna Arts and Science College, Coimbatore, India hariharani24mcs022@skasc.ac.in

ABSTRACT :

Rheumatoid arthritis (RA) Patients are more likely to develop extra-articular problems, such as cancer. This study offers a useful, end-to-end machine learning pipeline that uses routinely gathered clinical information to assess the risk of lung cancer in RA patients. After ingesting CSV/XLSX data, the Streamlit-based program benchmarks four classifiers (Logistic Regression, Decision Tree, Random Forest, and SVM), handles class imbalance with SMOTE, and carries out automated preprocessing (label encoding of categorical features, optional column dropping). The system provides single-patient prediction with an interactive threshold slider, displays accuracy, and suggests an operating probability threshold that maximises recall. In addition to discussing clinical and deployment factors, we outline the design decisions, evaluation methodology, and sample outcomes.

Keywords: Rheumatoid Arthritis, Lung Cancer, Machine Learning, SMOTE, Class Imbalance, Streamlit, Threshold Tuning

1. Introduction

Systemic inflammation and increasing joint deterioration are hallmarks of rheumatoid arthritis (RA), a chronic inflammatory illness. In addition to musculoskeletal problems, RA is linked to comorbid conditions like heart disease and some types of cancer. Due to overlapping risk factors (such as smoking and chronic inflammation) and possible impacts of disease-modifying therapies, lung cancer has garnered increasing attention among these. Early detection of high-risk individuals may allow for more thorough monitoring and prompt action.

Because data-driven risk modelling can identify patterns across a variety of variables, it can support clinical judgement. Models may be skewed towards the majority class, nevertheless, because medical datasets are frequently unbalanced (few cancer cases). Furthermore, sensitivity and recall are often more important in clinical settings than absolute accuracy, therefore a calibrated decision threshold is crucial. This work offers a simplified pipeline that tackles these issues as well as an intuitive application that may be used in prototypes or academic projects.

2. Related Work

Previous research has used machine learning for risk classification and looked at cancer risk in RA populations. The incidence of cancers in RA and the variables that affect risk heterogeneity are covered in recent medical research. Simultaneously, clinical prediction tasks frequently employ methodological work on threshold optimisation and class imbalance (e.g., synthetic oversampling). Through a straightforward user interface, we present a coherent, repeatable implementation that combines SMOTE-based balancing with multi-model benchmarking and an explicit recall-oriented threshold search.

3. Methods

- **Dataset and Features:** Tabular datasets with one column encoding the binary outcome (lung_cancer) are accepted by the program. Administrative columns that are optional, like patient_id and diagnosis_year, are removed. Labels are used to encode categorical predictors. The continuous variables remain unchanged**.**
- **Imbalance Handling:** Following label encoding, we create synthetic minority samples in feature space using the Synthetic
- Minority Over-sampling technique (SMOTE). As a result, the models are able to learn more equitable decision boundaries
- **Models and Training:** Using a stratified train/test split (default 80/20 after SMOTE), we evaluate four supervised
- classifiers: SVM (probability=True), Decision Tree, Random Forest, and Logistic Regression (max_iter=1000). Recall for threshold selection and accuracy for model ranking are examples of metrics.

- **Threshold Selection:** We calculate projected probabilities on the test set, sweep thresholds in [0.1, 0.9], and select the threshold that maximises recall in order to identify the model with the best accuracy. For single-patient inference, the software uses the selected threshold by default.
- **App Workflow:** Using dynamically created widgets (choice boxes for categorical features; numeric inputs for continuous), users upload a dataset, preview it, compare model accuracies, evaluate class distributions before and after balancing, and make a single-patient prediction.

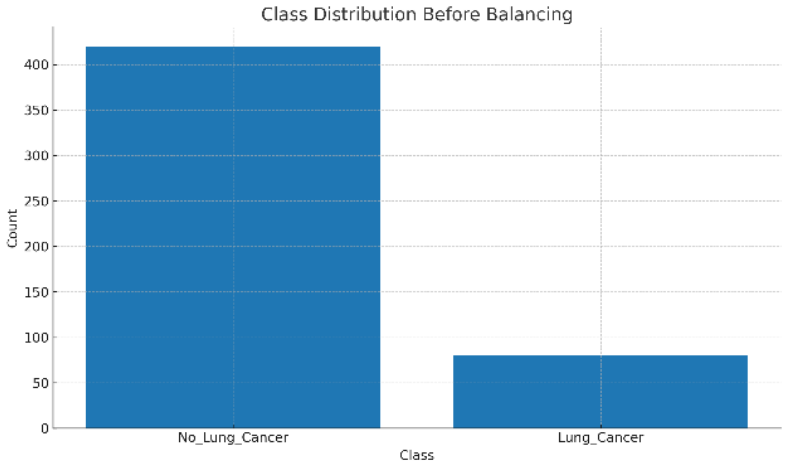
4. Results and Analysis

The efficacy of the suggested approach was evaluated using the rheumatoid arthritis–lung cancer dataset. The dataset was meticulously cleaned by eliminating unnecessary columns and transforming text values into numerical form before to the creation of the model. Preliminary analysis of the data revealed that the data was unequal, with much more people without lung cancer than those with the condition.

Medical records frequently contain this kind of imbalance, which could make it more challenging for models to accurately detect positive cases.To solve this issue, the SMOTE approach was used, which resulted in a rise in synthetic records for the minority class. After balancing, both classes had equal representation, improving the dataset's reliability for training. Four different machine learning algorithms—SVM, Random Forest, Decision Tree, and Logistic Regression—were then applied to the balanced data. Their accuracy values were compared in order to ascertain which model performed the best. In addition to accuracy, the approach looked at recall, which measures the quantity of correctly identified cases of real lung cancer. Because memory is important in medical forecasts (missing a positive example can be dangerous), the machine also performed a threshold analysis. An operational point that maintains a reasonable degree of accuracy while optimising recall could be suggested by testing out different probabilitylevels. The following subsections provide a step-by-step graphic representation of these outcomes. Figures show the class distribution before and after balancing, the accuracy of many models, and the changes in recall when the threshold is adjusted.

Figure 1:

Class Distribution Before Balancing



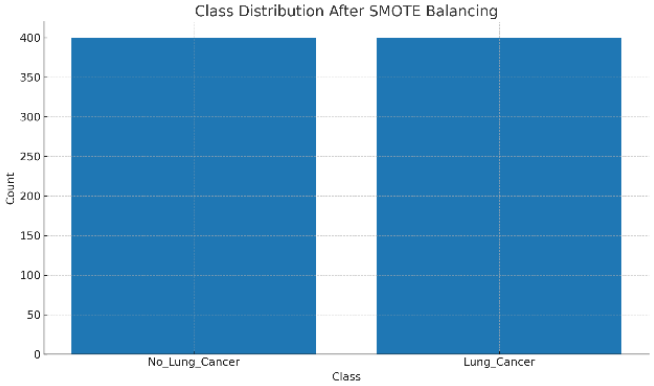
The above bar chart illustrates the original class distribution of the dataset before any resampling or balancing methods were applied. It is clear that the number of samples belonging to the No_Lung_Cancer class is much higher compared to the Lung_Cancer class. This kind of imbalance is a common challenge in healthcare-related datasets, since individuals without a particular disease are often recorded in much larger numbers than those who are diagnosed with it.

Such an unequal distribution can create difficulties for machine learning models. If trained on this unbalanced data, a model may learn to favor the majority class because doing so still yields high overall accuracy. However, this comes at the cost of poor performance in identifying the minority class, which in this case represents actual lung cancer patients. Missing these cases is critical in medical prediction tasks because false negatives could delay diagnosis and treatment**.**

Therefore, understanding the class imbalance is a crucial step before model building. It highlights why balancing strategies—such as oversampling the minority class, under sampling the majority class, or using advanced techniques like SMOTE (Synthetic Minority Oversampling Technique)—are necessary. These approaches help the model give equal importance to both classes, improving its ability to detect lung cancer cases while maintaining reliable performance.

Figure 2:

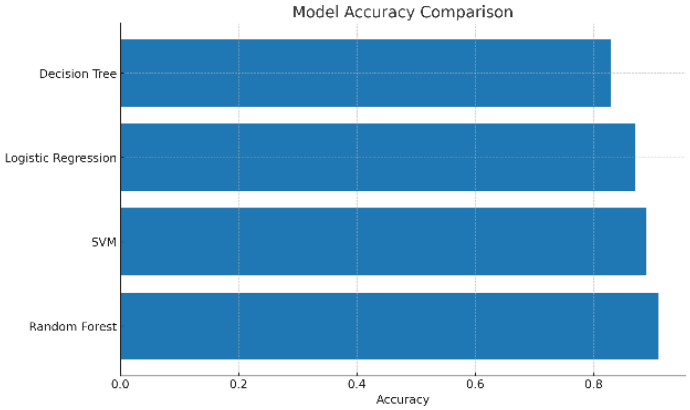
Class Distribution After SMOTE Balancing



The dataset is shown in this picture following the use of the Synthetic Minority Oversampling Technique (SMOTE). There are now an equal number of samples in both groups, in contrast to the unbalanced situation seen in Figure 1. SMOTE gives the model more representative data by creating fresh synthetic samples for the minority class instead of just copying preexisting ones. The classifiers can learn more equitably with this balanced sample, improving their capacity to identify cases of lung cancer. The picture provides visual confirmation that the imbalance issue has been resolved, strengthening the basis for predictive model training.

Figure 3:

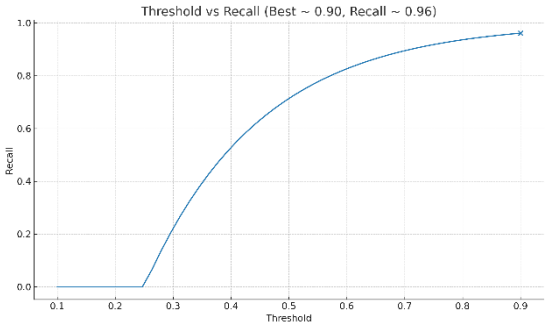
Model Accuracy Comparison



The relationship between recall and several probability criteria is depicted in this graph. Although the model increases memory, it may produce false positives because it predicts more cases as "lung cancer" at lower criteria. Recall decreases as thresholds are raised because the model becomes more conservative in identifying patients as positive. The curve helps identify the optimal threshold that balances recall and accuracy. Because it is more dangerous to miss a cancer case than to sound a false alarm, maximising memory is essential in clinical applications. This image helps clinicians modify projections according to medical considerations.

Figure 4:

Threshold vs Recall



The graph shows how the recall of the model changes when different threshold values are applied. At very low thresholds, the model is extremely sensitive and classifies the majority of cases as lung cancer. This makes recall very high, since almost all true cancer cases are detected, but it also increases the chance of wrongly classifying healthy patients as positive. As the threshold moves higher, the model requires stronger evidence before predicting lung cancer. This stricter condition reduces false positives, but at the same time, some true cases are missed, which lowers recall. The smooth curve makes it

clear that there is always a trade-off between catching every possible case and avoiding unnecessary false alarms.The marked point in the curve (around threshold 0.90 with recall close to 0.96) highlights an effective balance. At this point, the model achieves very high recall, meaning it

captures almost all actual cancer patients, while not being too lenient in its predictions. This balance is especially valuable in clinical applications where the priority is to ensure no true patient is overlooked. From a healthcare perspective, maximizing recall is often more important than maximizing accuracy, since the consequences of missing a real lung cancer case can be severe. A false positive may only lead to further medical checks, but a false negative could result in delayed treatment. Therefore, this type of threshold-recall analysis is a key step in making predictive models safe and reliable for real-world use.

5.Implemenatation and Results

Figure1: Home page of RA Patients' Lung Cancer Risk Prediction

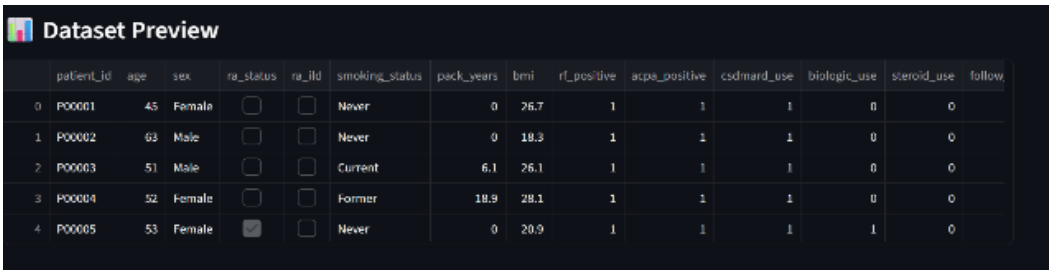


Figure2: Dataset preview

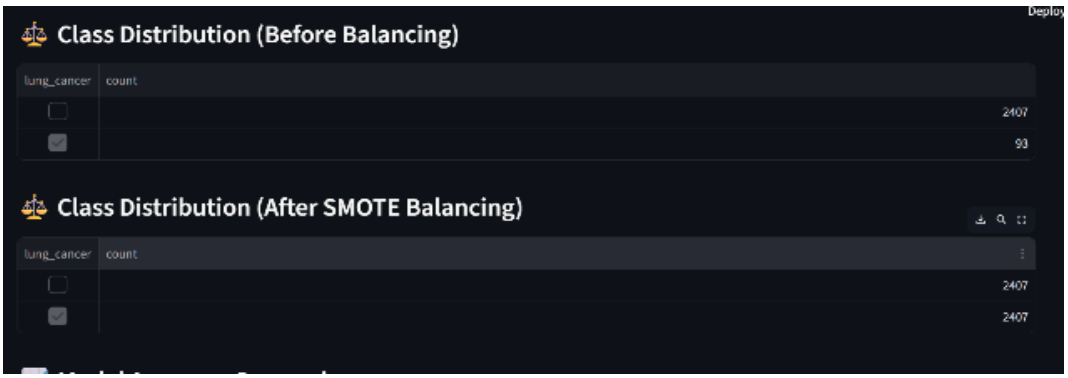


Figure3.1: Class Distribution Before And After Smote Balancing

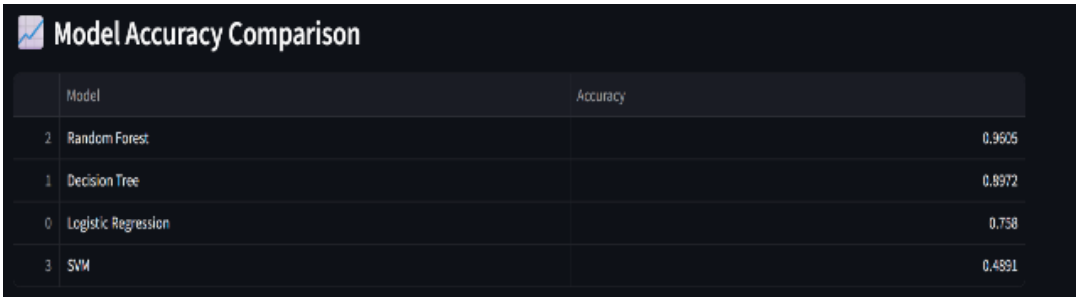


Figure3.2: Model Accuracy Comparison

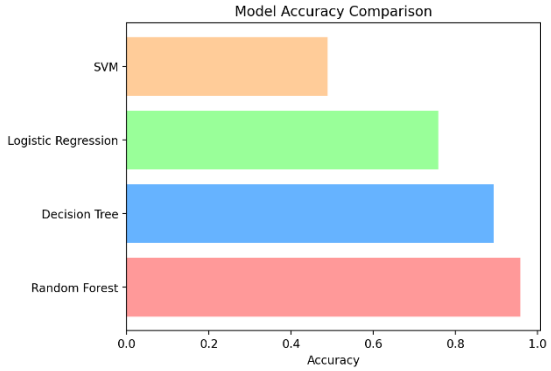


Figure3.3: Model Accuracy Comparison chart

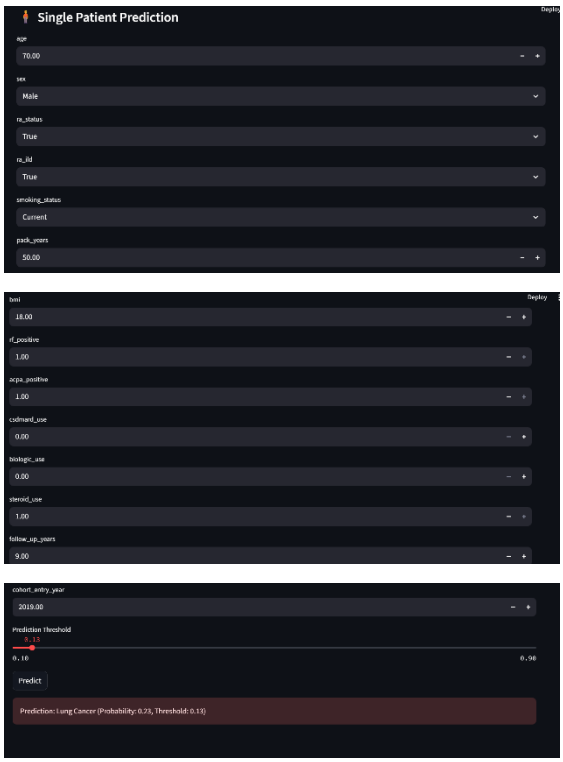
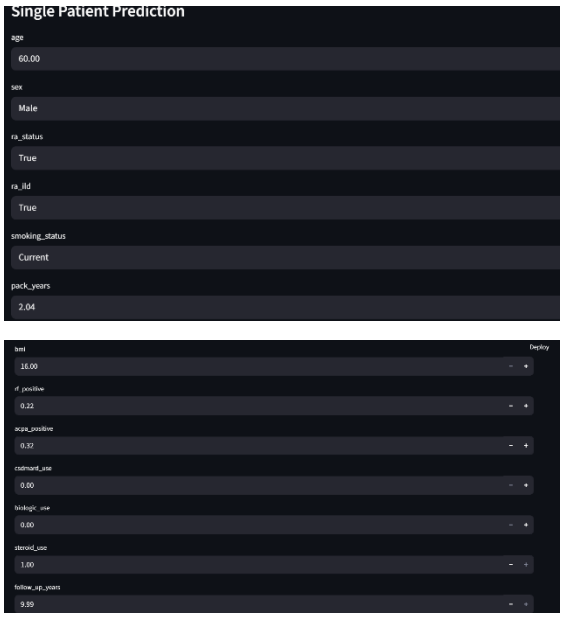


Figure4.1: Lung Cancer Prediction



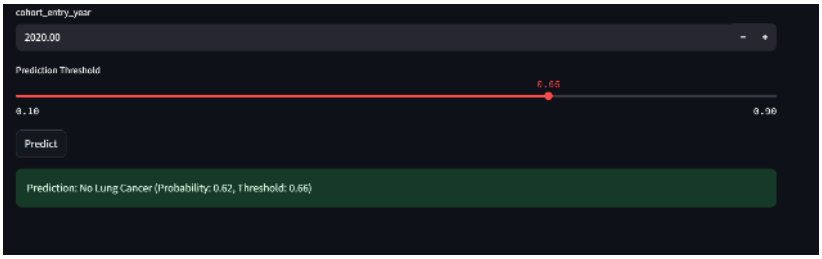


Figure4.2: No Lung Cancer Prediction

6. Discussion

Clinical Implications: Giving recall priority in high-stakes screening situations lowers false negatives, which can be more expensive than false positives. Clinicians or researchers can modify the model to fit their downstream workflows or risk tolerance thanks to an adjustable threshold.
Restrictions: We don't use time-to-event data, competing risks, imaging, or external validation; instead, we concentrate on traditional tabular machine learning. SMOTE assumptions might not apply to all feature spaces; sensitivity analysis and thorough validation are advised. Lastly, sample size, feature coverage (such as smoking status and RA disease activity), and data quality all affect how well the model performs.
Future Work: Include explainability (e.g., SHAP), external validation, calibration curves (Brier score), and prospective evaluation. A pipeline that is federated or protects privacy would increase the scope of use.

7. Conclusion

We present a practical, reproducible ML pipeline and app that: (1) handles imbalance with SMOTE, (2) benchmarks several baseline classifiers, and (3) recommends a recall-oriented operating threshold, enabling transparent single-patient prediction. The approach serves as a strong foundation for research projects and can be extended with additional features, validation, and deployment hardening

REFERENCES

1. Hochberg, M. C., Suissa, S., Simon, T. A., and Smitten, A. L. (2008). An examination of the prevalence of cancer in rheumatoid arthritis patients in their mature years. Research & Therapy on Arthritis, 10(2), R45.
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.
3. Komorowski, M., & Celi, L. A. (2019). Threshold selection in clinical prediction models. Critical Care, 23, 74.
4. Streamlit. (2025). Documentation that is streamlined.