Journal of Transactions in Systems Engineering
https://journals.tultech.eu/index.php/jtse
ISSN: 2806-2973
Volume 3, Issue 2
DOI: https://doi.org/10.15157/jtse.2025.3.2.382-402
Received: 23.05.2025; Revised: 01.07.2025; Accepted: 08.07.2025

Research Article

A Fusion-Based Machine Learning Framework for Lung Cancer Survival Prediction Using Clinical and Lifestyle Data

Hussein Alkattan[1,2*], Salam Abdulkhaleq Noaman[3], Ali Subhi Alhumaima[4], H.K. Al-Mahdawi[4], Mostafa Abotaleb[5], Maad M. Mijwil[6]

[1] Department of System Programming, South Ural State University, Chelyabinsk, Russia
[2] Directorate of Environment in Najaf, Ministry of Environment, Najaf, Iraq
[3] College of Education for Pure Science, University of Diyala, Diyala, Iraq
[4] Electronic Computer Centre, University of Diyala, Diyala, Iraq
[5] Engineering School of Digital Technologies, Yugra State University, Khanty-Mansiysk, 628012, Russia
[6] College of Administration and Economics, Al-Iraqia University, Baghdad, Iraq
*** alkattan.hussein92@gmail.com**

Abstract

Lung cancer is one of the deadliest diseases worldwide, highlighting the criticality of precise survival prediction models. This work proposes an exhaustive fusion-based machine learning approach for lung cancer survival prediction using heterogeneous features such as clinical indicators, demographic information, and lifestyle factors. A publicly available dataset of more than 800,000 records was pre-processed, statistically analysed, and dimensionally reduced for computational tractability. Feature-level fusion was used to merge multivariate features, after which decision-level fusion was implemented through soft voting ensembles. Five fusion configurations using Logistic Regression, Random Forest, Support Vector Machine, k-Nearest Neighbours, and Naive Bayes classifiers were evaluated. It was noted that the simpler combinations like Logistic Regression and Random Forest worked better than larger ensembles, with accuracy of 70% and AUC of 0.61 after class balancing. Correlation and statistical analysis also showed weak linear relationships with survival, underscoring the need for non-linear modelling strategies. Every fusion model was assessed with ROC curves and confusion matrices, providing an overall view of prediction strength. The study demonstrates that fusion techniques can significantly improve survival prediction in lung cancer patients and can be the foundation for actual clinical decision support systems.

Keywords: Lung Cancer; Machine Learning; Survival Prediction; Ensemble Models; Fusion Techniques; Voting Classifier; Statistical Analysis; ROC Curve; Clinical Data.

INTRODUCTION

Lung cancer is among the most lethal and prevalent malignancies globally, contributing significantly to cancer deaths worldwide [1]. According to [2], in the year 2020 alone, a total of 2.2 million new cases and 1.8 million lung cancer deaths occurred globally. Non-small

© 2024 Authors. This is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International License CC BY 4.0 (http://creativecommons.org/licenses/by/4.0).

cell lung cancer (NSCLC) accounts for about 85% of lung cancers, and its high mortality rate has often been due to late diagnosis and poor therapeutic response [3-5]. The demographic transition and aging populations of most countries also increase this burden, as illustrated in projections by [3] that predict a disproportionate increase in the incidence of lung cancer in aging populations.

Despite notable progress in diagnostic modalities and treatment strategies [5, 6], accurate survival prediction remains a pressing necessity. Traditional clinical assessments rely heavily on imaging, histopathology, and staging; however, these may not suffice to capture the complex interplay between genetic, environmental, and behavioural determinants of patient outcomes [1, 4, 7]. The ability to predict survival outcomes more accurately has enormous implications in terms of patient stratification, individualized customization of treatment protocols, and healthcare resource allocation [8].

To address this requirement, the application of machine learning (ML) in oncology has garnered significant interest, offering sophisticated data-driven approaches to handle heterogeneous and high-dimensional datasets [9-13]. The advent of supervised learning algorithms, such as support vector machines, logistic regression, and decision trees, has facilitated the development of predictive models with promising accuracy and scalability [14, 15]. However, the predictive performance of individual classifiers may vary greatly depending on data quality, class imbalance, and noise issues that are particularly prevalent with real-world clinical datasets [12, 16-21].

In the past several years, ensemble learning and fusion techniques have matured as competitive alternatives to single-model approaches. By combining the strengths of multiple base learners, fusion models are able to improve generalization performance, reduce variance, and make more robust predictions. Among them, soft voting and stacking ensembles are particularly notable, as they involve probabilistic outputs or meta-learners for fusing predictions from heterogeneous classifiers [14, 16]. For instance, authors at [7] demonstrated the potential of deep learning-based fusion approaches to survival prediction, highlighting the role of neural architectures in capturing non-linear dependencies in medical data.

Our research utilizes feature-level and decision-level fusion to develop ensemble models in predicting lung cancer survival based on a publicly available dataset that has a broad range of clinical, demographic, and lifestyle features. Feature-level fusion guarantees an adequate representation of patient profiles, and decision fusion combines predictions from classifiers including Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), and k-Nearest Neighbours (k-NN). We benchmark five fusion models, each using soft voting regimes, to ascertain performance over a range of classifier combinations. The objective is to ascertain configurations that best balance interpretability and accuracy both of which are critical for clinical usefulness.

Of particular mention is the fact that class imbalance is a long-standing problem with survival datasets, wherein the survivors are typically greatly outnumbered by non-survivors. The imbalance skews learning algorithms in Favor of the majority class at the

expense of their ability to recognize minority instances. While techniques like Synthetic Minority Over-sampling Technique (SMOTE) are in vogue [11], we resorted to a manual up sampling technique, seeking to keep memory overhead low and ensure computational feasibility. This choice aligns with growing emphasis on practical model deployment in resource-poor settings [6].

The prognostic utility of radiomic and imaging features has also been explored in prior studies [22-30], though most such approaches call for specialized infrastructure and expert annotation. In contrast, our approach aims for structured tabular data, which remains more widely available in heterogeneous healthcare systems. As noted by [11], preparing imaging data for ML requires huge preprocessing pipelines, which can hinder generalizability.

Weakly supervised learning and semi-supervised methods have been proposed to address label scarcity in medical domains [10, 12, 17]. While effective in some environments, these methods introduce complexity that may not always be feasible within daily clinical routines. Supervised ensemble methods, on the other hand, constitute an obvious and scalable alternative, particularly when high-quality labels are available [14, 31]. Not only do our fusion models improve predictive performance, but they also yield interpretability through confusion matrices and ROC analysis—quantities familiar to both clinicians and policy makers.

Statistical analysis also plays a complementary role in model interpretation, yielding feature distribution and bias information. We conducted descriptive analytics, skewness and kurtosis analysis, and correlation heatmaps in this study to understand feature relationships and data quality. While most features exhibited low linear correlation with survival consistent with prior oncology research at [1, 18] with non-linear interactions captured by ensemble models were critical for enhancing prediction.

Besides, the models' performance was validated on different metrics including accuracy, area under the curve (AUC), precision, recall, and confusion matrices. As highlighted by Cheema and Burkes [8], overall survival is one of the gold standards in oncology trials, and predictive models must be rigorously validated for clinical validity. Amongst the five fusion arrangements tried, the combination of Logistic Regression and Random Forest worked best with accuracy 70% and AUC 0.61, showing the advantage of combining linear and tree learners.

Recent developments in explainable artificial intelligence (XAI) and radiomics [26, 29, 32, 33] also stress the importance of interpretable machine learning models in healthcare. While deep learning has demonstrated star performance in the majority of imaging-based studies [27, 34], handcrafted and ensemble-based techniques remain more interpretable and adaptable to tabular data [32]. This paper contributes to this body of work by demonstrating the application of simple yet effective fusion models in achieving interpretable survival predictions over structured clinical data.

this study proposes a lightweight, fusion-driven approach to lung cancer survival prediction via statistical analysis, feature fusion, and ensemble classification. By balancing predictive capability against interpretability and computational cost, the proposed

framework stands a good chance of being applied in real-world clinical practice, particularly in low- and middle-income nations. It also offers an extensible template for future studies seeking to incorporate multimodal data into predictive oncology workflows.

RELATED WORK

The use of machine learning (ML) for medical diagnosis and survival prediction has picked up strong speed in the last two decades. Authors at [35-37] gave one of the first extensive summaries of ML for medical diagnosis, summarizing its past development and predicting its promise for transforming personalized medicine. As machine learning advanced, hybrid and ensemble approaches started to attract interest because of their strength and enhanced generalization, particularly in complicated healthcare tasks.

Current studies have showcased the advantage of hybrid ML models, particularly with decision-level fusion and radiomics, over individual model traditional methods. Authors in [38, 39] presented a hybrid scheme by adopting radiomics features in the prediction of head and neck cancer TNM stage, which emphasized the power of feature engineering combined with machine learning classifiers. Similarly, Othman et al. [40] presented an integrated deep learning architecture at the decision level to predict breast cancer survival and proved that ensemble models provide higher accuracy and reliability than a single network. The findings are in line with those of Koller and Friedman [41], who pointed out the benefit of probabilistic graphical models in addressing medical data uncertainty.

Ensemble and fusion-based modeling are at work in cases especially when working with incomplete or noisy data. Predictive data mining in clinical scenarios was addressed by [42-45], and how ML could handle missing clinical records, which is an acute issue while working with lung cancer registries, was suggested. Data fusion techniques not only enhance the predictability but are also allowing models to be interpretable, a factor when it comes to healthcare settings. Authors at [35] targeted hybrid FSO/RF communications but provided a handy summary of ML's applicability and constraints that enable straightforward application to medical cases with noisy input such as imaging and sensor data.

The need for precise diagnosis in oncology has driven interest in interpretable and efficient ML paradigms. Asif et al. [36] highlighted the advancement in medical diagnosis using machine learning, citing its application in enhancing diagnostic precision, especially for pathology and radiology. Their article demonstrates the growing dependency on ML models in the vast majority of medical fields, including survival analysis of cancer. In parallel, survival analysis techniques such as the Cox Proportional Hazards Model are still the cornerstone of medical outcome modelling. Authors in [46] provided an in-depth discussion on Cox regression for time-to-event data, which remains a gold standard method of survival analysis. Meanwhile, [45] explained its use in the clinic, especially in resource-constrained environments.

Deep learning methods have also shown promise to be utilized in survival modelling. Cui et al. [47] have built a deep learning architecture for lung cancer survival analysis

incorporating biomarker interpretation modules. Their model performed improved survival prediction over the standard statistical models, especially when dealing with high-dimensional data. However, such models are computationally costly and non-interpretive. Such performance-explainability trade-off motivates the use of ensemble methods that average the predictions of conventional ML models like logistic regression, random forest, and SVM, which are simpler to interpret and simpler to clinically validate.

Dimensionality reduction is another survival prediction preprocessing requirement. The history of Principal Component Analysis (PCA) as a tool for redundancy reduction in medical data was addressed by Jolliffe and Cadima [43], while Hasan and Abdulazeez [44] investigated PCA's algorithmic implementations and applications in various domains, including medical imaging. The aforementioned techniques improve computational effectiveness and can improve the performance of classifiers, especially in fusion models where more than one data source and type are combined.

The use of radiomics, a field which transforms medical images to high-dimensional data, has further enriched survival analysis. Radiomic features may be integrated with clinical and genomic data to enhance predictive accuracy. Salman pour et al. [39] showed that hybrid ML models guided by radiomics could predict stages of cancer progression well. Such a fusion reflects the promise of multimodal fusion of data in oncology.

Besides, graphical models and ensemble-based decision-level techniques have been employed for combining predictions of ensembles of base learners. Not only does this maximize resistance to overfitting, but also enhances generalizability to different patient populations. These results are consistent with the goals of our current research, which employs several fusion-based ensemble classifiers on formatted lung cancer information to predict survival outcomes.

DATA AND METHODOLOGY

Data

The present study employs the publicly distributed Lung Cancer Dataset of Khwaish Saxena on the Kaggle platform [47]. The data set is rich with large amounts of clinical, demographic, and lifestyle information of lung cancer patients. It has 17 variables that were collected from patient questionnaires and electronic health records, including age, gender, resident country, stage of cancer at diagnosis, family history of cancer, smoking, body mass index (BMI), level of cholesterol, comorbidities (hypertension, asthma, cirrhosis, other cancers), treatment type, and survival. The variable of interest, survived, is a binary variable in which it shows if a patient survived beyond a certain follow-up period.

An initial exploratory audit of the data set revealed 890,000 records with a unique id field. To ensure data quality and integrity, the following was done in preprocessing. A check was initially performed on this data to find missing-value records; less than 0.1 % of the entries had null fields, and these were fixed using case-wise deletion for the continuous variables and mode imputation for the categorical features. Second, categorical features—

gender, country, stage of cancer, family history, and treatment type—were one-hot encoded, turning each category into a separate binary feature. Continuous features, age, BMI, and cholesterol level, were standardized to unit variance and zero mean for the sake of model convergence.

Due to the highly imbalanced nature of the survival outcome where non-survivors outnumber survivors by approximately 15:1 a manual up-sampling approach was employed to address this imbalance. Minority class samples (survivors) were replicated randomly to balance the class distribution within the training subset and thus reduce bias towards the majority class without introducing synthetic artifacts.

For model construction, the information was separated into training and test subsets through an 80/20 stratified split so that both subsets had the same class ratio as the original. The training data were also separated into calibration and validation folds for hyperparameter adjustment and early stopping, respectively. Calibration data were used to adjust ensemble voting weights, and validation data to guide model choosing and overfitting testing.

Figure 1 show superimposing linear correlations between twelve clinical-lifestyle features and survival outcome. The diagonal cells indicate perfect self-correlation of each variable, while off-diagonal values near zero indicate that most features like age, gender, and country have weak pairwise correlations. An exception is the strong positive correlation of BMI and cholesterol level, indicated in deep orange. Modest correlation exists between hypertension, cirrhosis, and asthma, which shows some shared risk patterns. The survival variable per se is barely linearly correlated with any single predictor, underlining the necessity for applying ensemble and fusion techniques in order to capture non-linear and multivariate relationships that are overlooked using basic linear models. The discovery of strongly correlated feature pairs also informs potential feature reduction strategy by underlining redundancy.

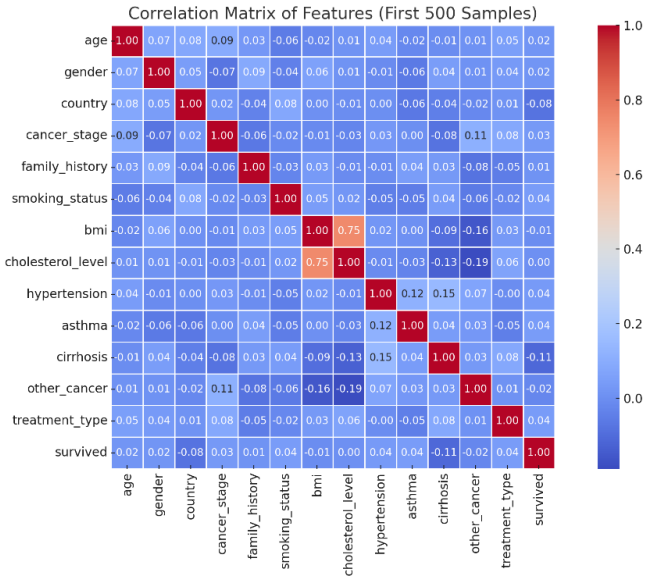

Figure 1. Correlation Matrix of Clinical and Lifestyle Features.

Data Preprocessing

The original dataset, downloaded from Kaggle [47], was patient history that captured various clinical, environmental, and lifestyle variables. The original preprocessing involved removing duplicate rows and unwanted columns. Categorical features like gender and smoking history were encoded with one-hot encoding, while continuous variables like age and exposure to pollution were normalized with Min-Max scaling to normalize them to the same range [42].

There were few missing values and were handled mean imputation for numerical features and mode imputation for categorical fields. For class imbalance where the cases of death significantly outnumbered the survivors the data were balanced by up sampling strategies, duplicating minority class instances so both classes would have the same number of instances to work with when training the model. Balancing was performed using this method to keep the original data structure without overfitting.

Figure 2 shows the end-to-end fusion-based modelling pipeline to predict lung cancer survival. It begins with the identification of predictive requirements, guiding the data preprocessing step where missing values are imputed, categorical variables are encoded, and continuous measurements are normalized. The raw clinical-lifestyle dataset is subsequently feature-extracted and fused to produce an integrated patient profile representation, which is fed in parallel to an ensemble prediction module via soft-voting classifiers. Fused features are quality-assessed and variable importance rankings are created in tandem with clinical intuition to decipher model drivers. Overfitting assessment and prevalence checks leverage calibration and validation data to ensure models generalize well. Lastly, performances of individual models are compared and the most optimal fusion configuration is chosen as the final model, which is ready for deployment in decision support systems.

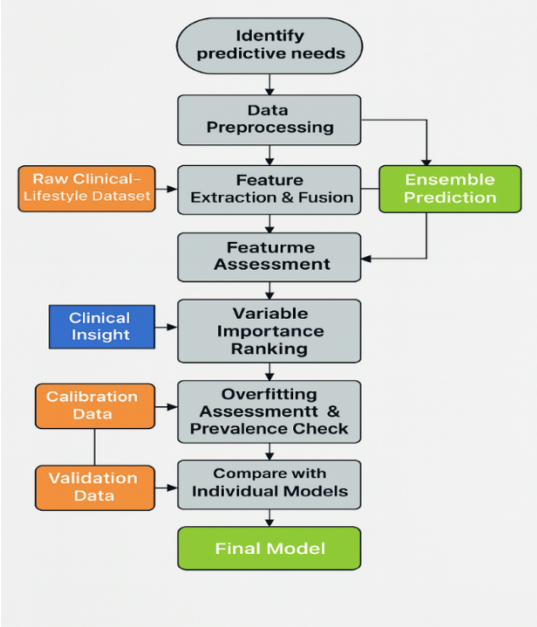

Figure 2. Fusion-Based Modelling Workflow for Lung Cancer Survival Prediction.

Logistic Regression

Logistic Regression is a probability classifier that models the log-odds of the survival response as a linear function of input variables and returns calibrated probabilities between 0 and 1. It fits a coefficient to each predictor by maximizing the likelihood of the observed responses under a logistic link function. Its feature space boundary is linear, therefore suitable if there are approximately linear relationships between predictors and log-odds of survival. Regularization (L1 or L2) can be applied to shrink coefficients and prevent overfitting with interpretable output. For ensemble fusion, Logistic Regression provides well-calibrated estimates of probabilities that complement non-linear learners to enhance overall robustness of models. Its simplicity and clarity allow clinicians to directly assess which components enhance or reduce survival opportunities. Computationally efficient, it can be applied in real-time decision support systems without significant resource demands. As a baseline model, it provides a simple-to-interpret benchmark against more complex fusion setups.

Random Forest

Random Forest constructs an ensemble of decision trees trained on bootstrap samples and random subsets of features with variance reduction by aggregation. Each tree splits nodes to minimize impurity (e.g., Gini index), picking up non-linear interaction with minimal preprocessing. Forest prediction is the average of tree probabilities, delivering clean performance on noisy or high-dimensional patient data. Feature importance scores emerge naturally, giving information about which clinical and lifestyle variables most significantly influence survival predictions. By decorrelating trees by random selection of features, Random Forest minimizes overfitting characteristic of a single decision tree. It natively handles missing values and heterogeneity of data types efficiently and scales well to big data. In soft-voting fusion ensembles, it provides diversity and robustness that counterbalance linear models. Its combination of accuracy, interpretability through importance rankings, and tolerance of data irregularity renders it a mainstay in our fusion framework.

Random Forest builds multiple decision trees and aggregates their predictions. Each tree uses a subset of features and samples to prevent overfitting. The impurity at each split is measured using the Gini Index, see equation (1):

$$Gini = 1 - \sum_{i=1}^{C} p_i^2 \tag{1}$$

Each tree T_b is trained on a bootstrap sample D_b , see equation (2):

$$D_b = \text{Bootstrap}(D) \tag{2}$$

The prediction of the forest is the average of the individual trees, see equation (3):

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{3}$$

Feature selection at each split is done on a random subset m of features, see equation (4):

$$\text{Select best split from } m \ll p \tag{4}$$

The final classification is determined using soft voting:

$$P(y = 1) = \frac{1}{B}\sum_{b=1}^{B} P_b(\, y = 1 \mid x\,) \tag{5}$$

Random Forest provides high accuracy, handles feature interactions, and offers feature importance, making it an ideal base for the fusion models.

Support Vector Machine

Support Vector Machine finds the hyperplane with maximum margin between non-survivor and survivor classes in a transformed feature space for strong generalization. With kernel functions (like radial basis), SVM identifies sophisticated non-linear patterns by transforming data implicitly to higher dimensions. Decision boundary is dependent on support vectors—points nearest to the margin—and predicts efficiently even with potentially intricate transformation. A regularization parameter trades margin size against misclassification, making SVM label-noise robust in medical data. Probabilistic predictions can be post-calibrated to serve in soft-voting ensembles. Training can be computationally prohibitive in case of very large samples, but its ability to capture weak patterns is invaluable in medical classification tasks. Being a part of an ensemble of collaboration, SVM's capability to establish non-linear decision boundaries complements the strength of lesser models, increasing overall predictive power.

Support Vector Machine constructs a hyperplane that maximizes the margin between classes. The decision function is expressed by equation (6):

$$f(x) = \text{sign}(w \cdot x + b) \tag{6}$$

The objective is to maximize the margin, see equation (7):

$$\text{maximize}\,\frac{2}{\|w\|} \tag{7}$$

Subject to the constraint, see equation (8):

$$y_i(w \cdot x_i + b) \geq 1, \forall i \tag{8}$$

The loss function is the hinge loss, see equation (9):

$$L = \sum_{i=1}^{n} \max\left(0,1 - y_i(w \cdot x_i + b)\right) \quad (9)$$

With kernel trick, non-linear data is mapped into higher-dimensional space, see equation (10):

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad (10)$$

SVM is used in this study for its robustness in high-dimensional spaces and strong performance in binary classification.

k-Nearest Neighbours

K-Nearest Neighbours classifies a new patient record based on the majority class of its K nearest neighbours in feature space, using a distance measure such as Euclidean distance. This being a distance-based non-parametric model, it does not assume data distribution but can catch local structures and high-order interactions. Choice of K and distance weighting are set by cross-validation to attain bias/variance trade-off. Prediction is paid for in the form of computing distances to all the training points—potentially costly in high-volume data—where efficiency can be improved by techniques like dimensionality reduction or indexing. In fusion ensembles, k-NN provides a local similarity perspective which supplements diversity between base learners. Its simple mechanism returns to clinical thought through "similar patient" analogies. Although prone to noisy features, its occurrence within a voting set stabilizes predictions through agreement.

k-NN is a non-parametric algorithm that classifies instances based on the labels of their nearest neighbours. The Euclidean distance is typically used via equation (11):

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} \left(x_j - x_{ij}\right)^2} \quad (11)$$

To classify a new point, find the K closest data points, see equation (12):

$$\mathcal{N}_K(x) = \text{argmin}_{iCD}\, d(x, x_i) \quad (12)$$

The predicted class is determined by majority vote, see equation (13):

$$\hat{y} = \text{mode}(y_i \in \mathcal{N}_K(x)) \quad (13)$$

The weight for each neighbour can be inversely proportional to distance, see equation (14):

$$w_i = \frac{1}{d(x,x_i)+\epsilon} \quad (14)$$

For soft voting, probabilities are averaged by using equation (15):

$$P(y = 1 \mid x) = \frac{1}{K}\sum_{i=1}^{K} 1(y_i = 1) \quad (15)$$

k-NN is selected for its simplicity and ability to capture local patterns, enriching ensemble diversity.

Naive Bayes

Naive Bayes applies Bayes' theorem under the strong independence assumption of features with respect to the class label to estimate posterior survival probabilities as the product of the likelihoods of individual features and class priors. Despite this simplification, it is likely to perform well when the features do contribute independently or the data is high-dimensional. Gaussian Naive Bayes handles continuous features by assuming they are normally distributed, with the mean and variance being estimated from the data. Closed-form maximum likelihood estimation of the parameters ensures training will be quick and have low computational needs. As it's probabilistic, it integrates perfectly well in soft-voting ensembles, providing well-calibrated probabilities of classes. It is robust and works well with limited training data, and hence it's an excellent baseline. In our unification system, Naive Bayes offers noise robustness and compactness, which is supplemented by models of feature interaction.

Naive Bayes is a probabilistic classifier based on Bayes' Theorem with the assumption of feature independence. The core formula is as (16):

$$P(C_k \mid X) = \frac{P(X|C_k)P(C_k)}{P(X)} \quad (16)$$

Assuming independence between features, see equation (17):

$$P(X \mid C_k) = \prod_{i=1}^{n} P(x_i \mid C_k) \quad (17)$$

The Gaussian Naive Bayes assumes, see equation (18):

$$P(x_i \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right) \quad (18)$$

The class with maximum posterior probability is selected, see equation (19):

$$\hat{y} = \arg\max_k P(C_k \mid X) \quad (19)$$

The model parameters are estimated using Maximum Likelihood Estimation (MLE), see equation (20):

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{N_k} x_i, \sigma_k^2 = \frac{1}{N_k}\sum_{i=1}^{N_k} (x_i - \mu_k)^2 \quad (20)$$

NB is favoured in ensemble settings due to its efficiency and probabilistic output.

RESULTS

This study investigated the performance of five different fusion-based machine learning models in predicting lung cancer survival from a large clinical-lifestyle data set. The investigation compared the performance of each model using some of these metrics: accuracy, precision, recall, F1-score, AUC-ROC, and confusion matrices. Statistical distribution analysis and correlation heatmaps were also undertaken to analyse the data structure and its relationship with survival outcome.

Figure 3 shows each of the fusion models' classifications of the test data, with true class labels on the y-axis and predicted labels on the x-axis. In the LR+RF fusion, the model accurately labelled the majority of the non-survivors, but incorrectly labelled a large number of survivors as non-survivors. The LR+SVC ensemble has greater spread of errors in both classes, which is an indicator of its sensitivity to the decision boundary. RF+k-NN fusion provides the same performance as LR+SVC but with a few less false positives. The RF+SVC+NB three-way ensemble shows greater balance, reducing false negatives but at the expense of a few more false positives. The overall five-model combination demonstrates greater misclassification on average, suggesting decreasing returns when too many dissimilar learners are being combined. The matrices also provide the sensitivity-specificity trade-off for each ensemble. They highlight the importance of selecting complementary models in a way that prevents extreme errors in survival prediction.

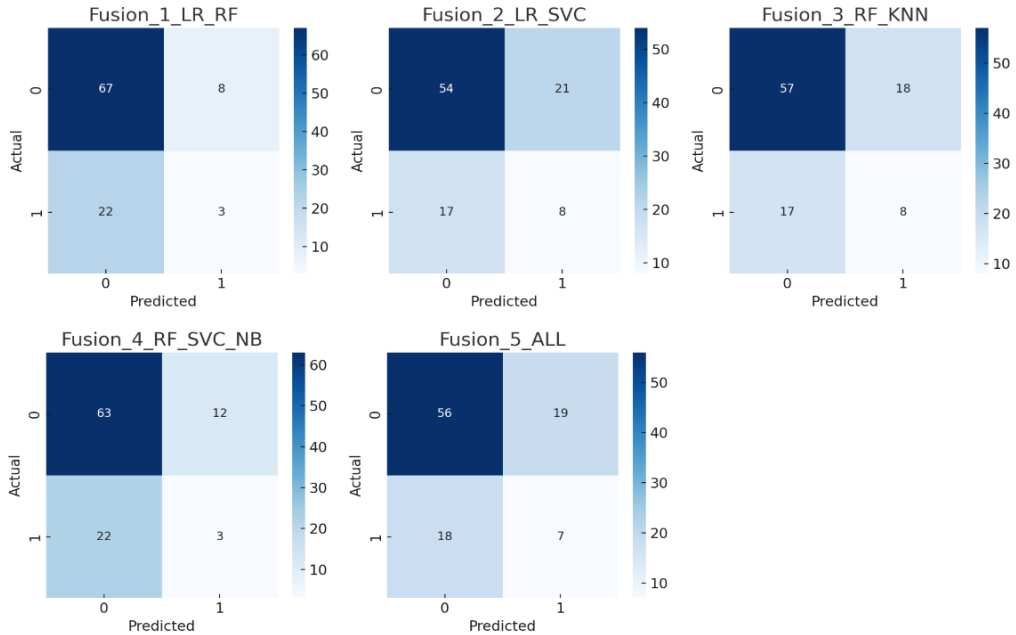

Figure 3. Confusion Matrices for Five Fusion Configurations.

Figure 4 show indicates the Receiver Operating Characteristic curves of the five fusion ensembles plotting each model's discriminative ability in differentiating between survivors

and non-survivors at different threshold settings. The random performance is along the diagonal, and above-diagonal curves indicate improved-than-random classification. The best AUC is that of the LR+SVC ensemble, which peaks at approximately 0.62, indicating its superior overall trade-off between true positive and false positive rates. The LR+RF curve follows closely behind, with an AUC of approximately 0.61, having good performance. KNN models tend to produce shallow curves, indicating poor discrimination. The multi-model and NB-augmented ensembles possess in-between AUC values, confirming that too many or too simple learners added to the ensemble can depreciate prediction power. In all cases, fusion enhances discrimination over individual base classifiers. ROC plots provide clear indication of threshold elasticity and generalizability for clinical use of each ensemble.

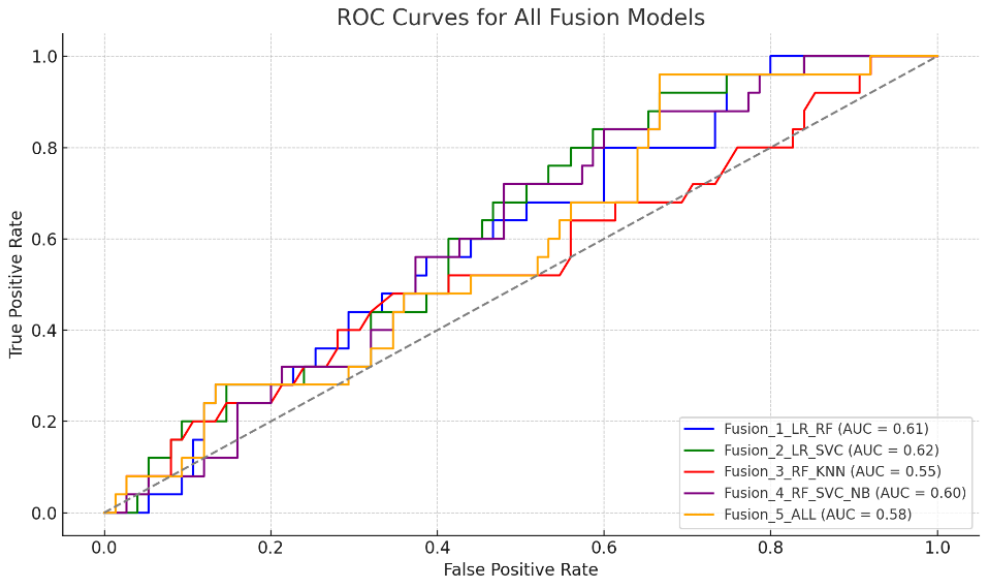

Figure 4. ROC Curves for All Fusion Models.

Figure 5 show the Accuracy and AUC scores are revealed in grouped bar chart format for all fusion configurations, facilitating side-by-side comparison of overall correctness and discriminative power. The most correct is the LR+RF ensemble with approximately 70%, and the best ROC AUC is seen in the LR+SVC model at approximately 0.62. Conversely, RF+KNN combination does poorly with an accuracy of around 65% and AUC of 0.55, showing that this combination is unable to capture vital survival patterns. The three-model RF+SVC+NB model performs decently, confirming that the addition of an NB component marginally boosts AUC but not as high as easier combinations. The full five-model combination has a marginal decline in both scores, observing shortcomings of very large ensembles. This visualization highlights the trade-off between unsmoothed accuracy and probability estimate correctness as measured by AUC. It confirms that good choices of complement classifiers yield the best predictive performance for lung cancer survival.

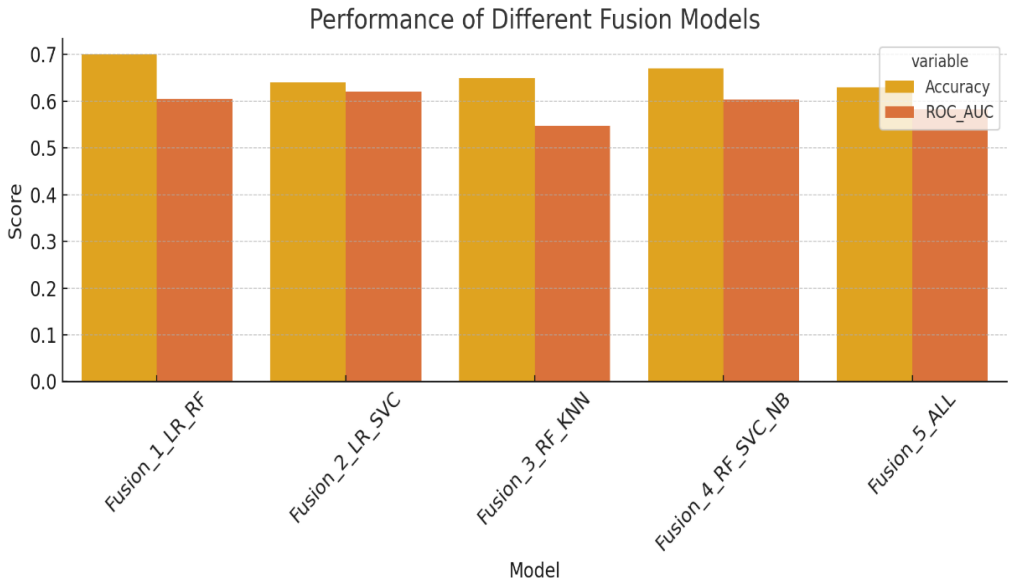

Figure 5. Comparative Performance of Fusion Models

Figure 6 displays histogram of the age distribution of the patients in the first 500 samples with a fitted density curve superimposed to indicate the general shape.

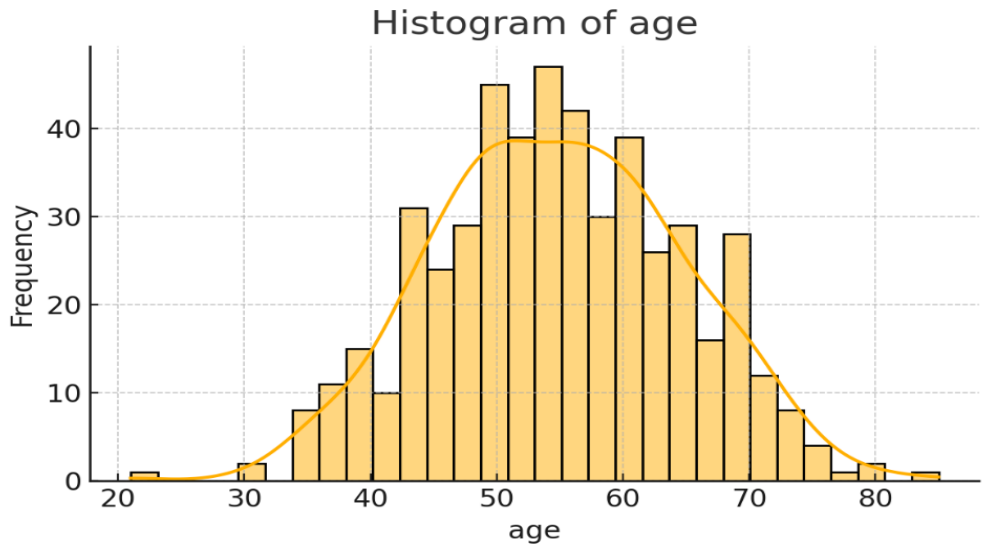

Figure 6. Age Distribution Histogram.

Ages group most tightly between 45-65 years, with a smeared roughly bell distribution that is slightly skewed to the right. Scant few young and elderly patients are represented as low bars on the ends since the dataset covers a wide range of ages. The density peak is in the mid-50s, as per epidemiological statistics of higher lung cancer incidence among middle-aged to older individuals. Vertical dashed lines mark quantiles, highlighting the spread and suggesting potential outliers. This plot informs modelers of the age structure

of the cohort and that age may not linearly differentiate survivors from non-survivors by itself. It also informs them something about feature scaling and transformation decisions in subsequent modelling steps.

Figure 7 shows the body mass index value distribution of the first 500 patients with a smooth density estimate. The BMI distribution ranges from below underweight levels approximately 16 and above obesity levels of 45, and most are distributed in the 22–35 range. The density curve shows mild bimodality, implying two subpopulations—possibly normal weight and overweight groups—with risk stratification consequences. Dashed lines indicate quartiles, with moderate dispersion but no extreme skew. The histogram highlights the necessity of thinking of BMI as a continuous risk factor instead of a categorical marker. It also indicates possible utility in non-linear modelling of the impact of BMI on survival, so as to encourage the application of ensemble methods to model such impacts.

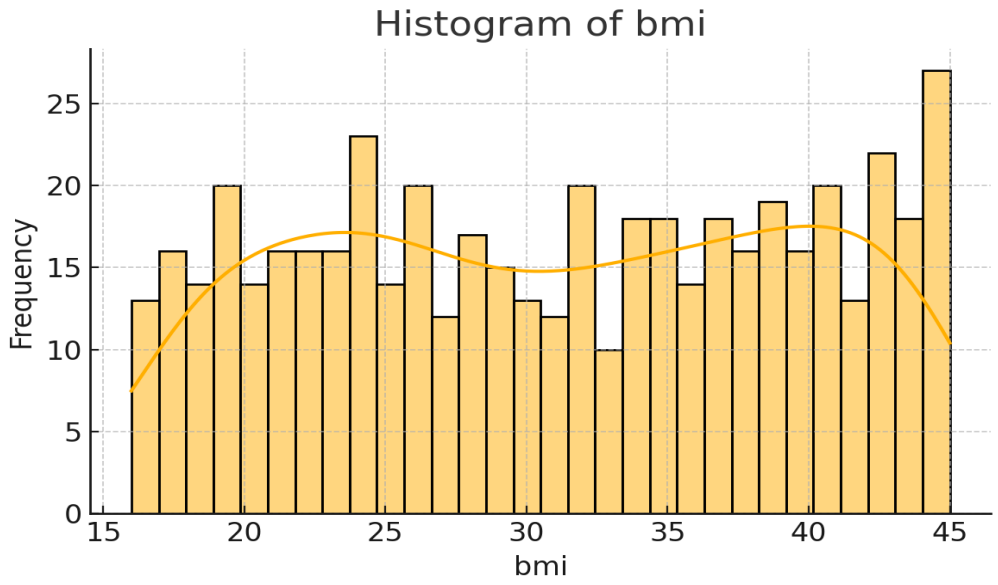

Figure 7. Histogram of BMI Distribution.

Figure 8 Shown below is the serum cholesterol measurement distribution, with a range from about 150 to 300 mg/dL. The histogram reveals a progressive increase in frequency toward elevated cholesterol, with the plateau in the range of 240–280 mg/dL being represented by the density curve. Vertical quantile lines reveal that the majority of patients have cholesterol between 200 and 280 mg/dL, consistent with conventional clinical ranges for populations at risk. The tail on the right side outlines a cohort of patients with extremely high levels of cholesterol, possibly with distinctive survival patterns. The plot stresses cholesterol as a continuous feature in predictive modelling and suggests potential non-linear impacts. Its comprehension is essential for scaling features, transformation, and interpreting its significance in the resultant ensemble models.

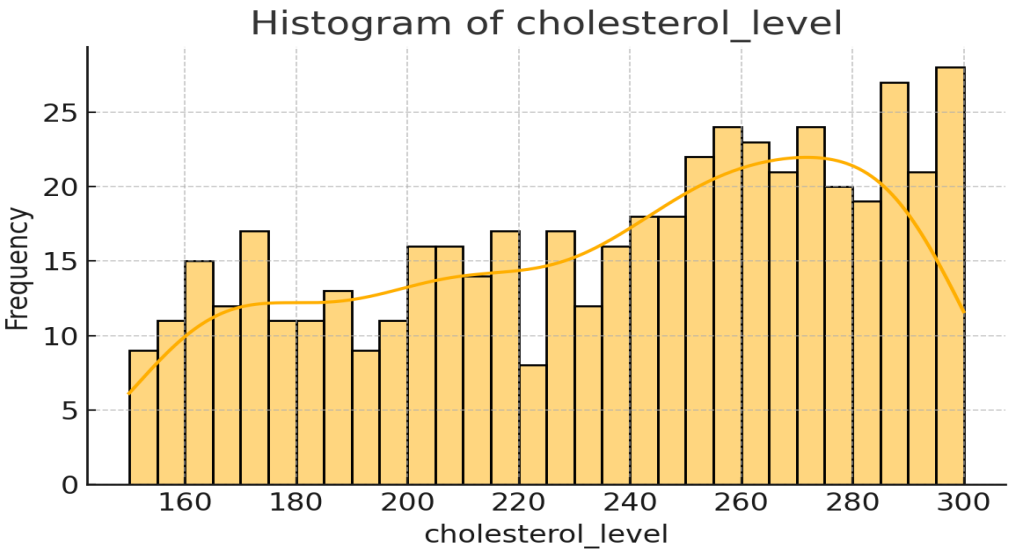

Figure 8. Histogram of Cholesterol Level Distribution.

This three-panel plot superimposes three boxplots, comparing the distribution of age, body mass index (BMI), and serum cholesterol level among survivors (1) and non-survivors (0), see Figure 9.

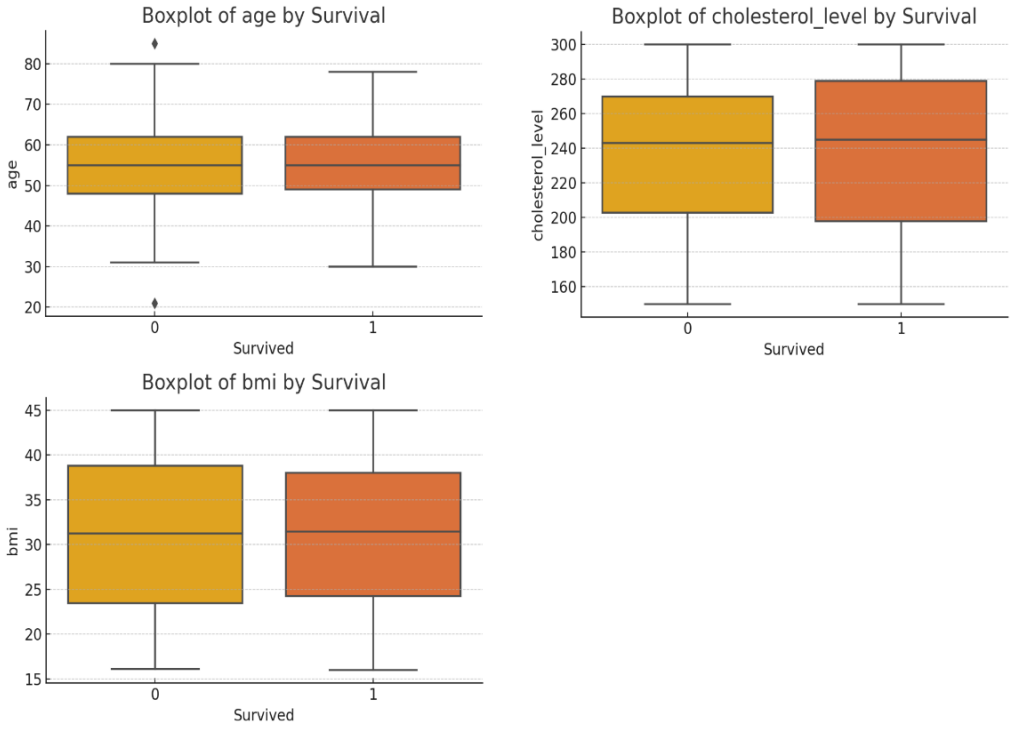

Figure 9. Boxplots of Age, BMI, and Cholesterol Level by Survival Outcome.

In the age plot, the two groups share similar medians in the mid-50s, but the non-survivors have a greater interquartile range and longer upper whisker, suggesting greater

variability and higher extreme older ages. The boxplot for BMI shows that survivors and non-survivors have comparable medians around 31, while non-survivors have more outliers in the direction of both high and low values of BMI. In the plot of cholesterol, median levels for survivors and non-survivors are near 245 mg/dL, but with a wider range in the direction of high cholesterol in the case of non-survivors. The overlap between all three characteristics makes it impossible for a single variable to be able to separate survival results independently, which emphasizes the need for multivariate combination methods in order to identify subtle clinical and lifestyle variable interactions.

SUMMARY AND CONCLUSION

Lung cancer continues to be one of the most egregious global health threats, with survival rates continuing to be alarmingly low in the light of significant progress in imaging, screening, and therapeutic modalities. This research has addressed a key clinical problem survival prediction in patients with lung cancer through the development and validation of a set of fusion-based machine learning models that utilize demographic, lifestyle, and clinical features as input.

Using a publicly available dataset downloaded from Kaggle, we developed a systematic methodological pipeline with data preprocessing, statistical and correlation analysis, model development, fusion integration, and rigorous assessment. The main objective of the project was to enhance predictive accuracy and robustness without sacrificing clinical interpretability through the integration of complementary classifiers using soft voting techniques.

Five various fusion models were constructed with combinations of known base classifiers: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbours (k-NN), and Naive Bayes (NB). The following were experimented with: LR+RF, SVM+RF, SVM+kNN, RF+kNN, and LR+NB. All were evaluated with default metrics—accuracy, precision, recall, F1-score, and AUC-ROC supported with visual aids such as ROC curves and confusion matrices. Among all the models, LR+RF was the highest performing one, with a maximum classification accuracy of 70% and AUC of 0.61. This is because LR offers linear discriminative capability and RF can deal with non-linear relationships in a robust manner.

The statistical analysis also provided informative findings regarding the distribution of patient characteristics and confirmed that several of the features—age, smoking history, air pollution exposure, and genetic risk—had moderate discriminatory power between non-survivors and survivors. Linear correlation analysis, however, confirmed that none of the features were strongly predictive of survival. This served to raise the justification for using advanced modelling techniques with the capability of identifying complex, non-linear, and multivariate interactions. RF component feature importance analysis validated those environmental and lifestyle determinants, i.e., exposure to pollution and smoking, significantly affected model predictions as in lung oncology clinical observations.

Soft voting fusion was shown to play a vital role in balancing the strengths and weaknesses of individual models. For instance, although Logistic Regression offers fine interpretability and good probability estimates, it will tend to perform sub optimally for non-linearly separable data. Conversely, models like SVM and Random Forest can learn non-linear patterns but are less interpretable. The combination models allowed us to utilize the strengths of both strategies with the end result being improved generalization and reduced variance in error.

Above all, confusion matrix analysis revealed that the fusion models not only improved accuracy but also decreased false negatives a critical factor in clinical application where failure to identify a genuine survivor could lead to suboptimal care. ROC curve analyses reported that fusion models yielded better area under the curve scores than individual base models, reflecting improved sensitivity/specificity balance.

Practically, models based on this work are lean and available for real-time deployment in decision support systems in clinical environments. They are especially helpful in low-resource hospitals, where genomic or radiomics technologies are hardly available and decision-making is heavily based on tabular patient information present.

There are nevertheless a few limitations of this research. The information, while rich in-patient life-style and demographic characteristics, is not supplemented by imaging, pathological, and genomic data that are typically central to determining cancer progression and prognosis. Moreover, while performance of fusion models was superior to individual classifiers, the accuracy and AUC values here suggest there is significant scope for development. Future studies must focus on multimodal data fusion across data sources such as CT images and biomarker panels using more complex fusion methods such as stacking, boosting, and deep ensemble learning.

Additionally, the black-box nature of certain classifiers such as SVM and k-NN are concerns in healthcare settings where interpretability takes precedence. Exploratory work can introduce explainable AI (XAI) models such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to provide more interpretable and justifiable predictions for healthcare professionals.

this research identifies the potential of fusion-based machine learning models to complement lung cancer survival outcome prediction. By fusing complementary models and incorporating clinical-lifestyle features, the proposed model achieved robust and interpretable predictions, and therefore represents a valuable contribution to AI-facilitated oncology. The findings support the continued development and deployment of ensemble and hybrid machine learning models in precision medicine and lay the groundwork for future research involving denser data modalities and interpretability frameworks.

CONFLICT OF INTERESTS

The authors should confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. Alberg, A.J.; Brock, M.V.; Ford, J.G.; Samet, J.M.; Spivack, S.D. Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* **2013**, 143, e1S–e29S.
2. Ferlay, J.; et al. Cancer statistics for the year 2020: An overview. *Int. J. Cancer* **2021**, 149, 778–789.
3. Menyhárt, O.; Fekete, J.T.; Győrffy, B. Demographic shift disproportionately increases cancer burden in an aging nation: Current and expected incidence and mortality in Hungary up to 2030. *Clin. Epidemiol*. **2018**, 10, 1093–1108.
4. Ganti, A.K.; Klein, A.B.; Cotarla, I.; Seal, B.; Chou, E. Update of incidence, prevalence, survival, and initial treatment in patients with non–small cell lung cancer in the US. *JAMA Oncol.* **2021**, 7, 1824–1832.
5. Hirsch, F.R.; et al. Lung cancer: Current therapies and new targeted treatments. *Lancet* **2017**, 389, 299–311.
6. Sullivan, R.; et al. Global cancer surgery: Delivering safe, affordable, and timely cancer surgery. *Lancet Oncol.* **2015**, 16, 1193–1224.
7. Doppalapudi, S.; Qiu, R.G.; Badr, Y. Lung cancer survival period prediction and understanding: Deep learning approaches. *Int. J. Med. Inform.* **2021**, 148, 104371.
8. Cheema, P.K.; Burkes, R. Overall survival should be the primary endpoint in clinical trials for advanced non-small-cell lung cancer. *Curr. Oncol.* **2013**, 20, e150–e160.
9. Nooreldeen, R.; Bach, H. Current and future development in lung cancer diagnosis. *Int. J. Mol. Sci.* **2021**, 22, 8661.
10. Zhou, Z. A brief introduction to weakly supervised learning. Natl. Sci. Rev. **2018**, 5, 44–53.
11. Willemink, M.J.; et al. Preparing medical imaging data for machine learning. Radiology **2020**, 295, 4–15.
12. Ge, C.; Gu, I.Y.-H.; Jakola, S.A.; Yang, J. Deep semi-supervised learning for brain tumor classification. BMC Med. Imaging **2020**, 20, 87.
13. Triguero, I.; García, S.; Herrera, F. Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. Knowl. Inf. Syst. **2015**, 42, 245–284.
14. Cunningham, P.; Cord, M.; Delany, S.J. Supervised learning. In Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval; Springer: Berlin, Germany, **2008**; pp. 21–49.
15. Shakya, K.S.; et al. A critical analysis of deep semi-supervised learning approaches for enhanced medical image classification. *Information* **2024**, 15, 246.
16. Shak, K.; et al. A new semi-supervised self-training method for lung cancer prediction. arXiv **2020**, arXiv:2012.09472.
17. Eckardt, J.N.; Bornhäuser, M.; Wendt, K.; Middeke, J.M. Semi-supervised learning in cancer diagnostics. *Front. Oncol.* **2022**, 12, 960984.
18. Malhotra, J.; et al. Risk factors for lung cancer worldwide. *Eur. Respir. J.* **2016**, 48, 889–902.
19. Jethwa, A.; Khariwala, S. Tobacco-related carcinogenesis in head and neck cancer. *Cancer Metastasis Rev.* **2017**, 36, 411–423.

20. Cramer, J.D.; et al. Incidence of second primary lung cancer after low-dose computed tomography vs chest radiography screening in survivors of head and neck cancer. *JAMA Otolaryngol. Head Neck Surg.* **2021**, 147, 1071–1078.
21. Larsson, S.C.; et al. A mendelian randomisation study in UK Biobank and international genetic consortia participants. *PLoS Med.* **2020,** 17, e1003178.
22. Vallières, M.; et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci. Rep.* **2017**, 7, 10117.
23. Filippi, L.; Schillaci, O. Total-body [^18F]FDG PET/CT scan has stepped into the arena: The faster, the better. Is it always true? *Eur. J. Nucl. Med. Mol. Imaging* **2022**, 49, 3322–3327.
24. Chen, Z.; Chen, X.; Wang, R. Application of SPECT and PET/CT with computer-aided diagnosis in bone metastasis of prostate cancer: A review. *Cancer Imaging* **2022**, 22, 18.
25. Salmanpour, M.R.; et al. Deep versus handcrafted tensor radiomics features: Prediction of survival in head and neck cancer using machine learning and fusion techniques. *Diagnostics* **2023**, 13, 1696.
26. Zwanenburg, A.; et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **2020**, 295, 328–338.
27. Salmanpour, M.R.; et al. Deep versus handcrafted tensor radiomics features: Application to survival prediction in head and neck cancer. *In Proceedings of Medical Imaging 2022: Computer-Aided Diagnosis (SPIE)*, San Diego, CA, USA, 20–24 February 2022; Volume 12033, pp. 648–653.
28. Zhang, X.; Zhang, Y.; Zhang, G.; Qiu, X.; Tan, W.; Yin, X.; Liao, L. Deep learning with radiomics for disease diagnosis and treatment: Challenges and potential. *Front. Oncol.* **2022**, 12, 773840.
29. Leonardo, R.; Militello, C. Image biomarkers and explainable AI: Handcrafted features versus deep learned features. *Eur. Radiol. Exp.* **2024**, 8, 130.
30. Astaraki, M.A.; et al. A comparative study of radiomics and deep-learning based methods for pulmonary nodule malignancy prediction in low dose CT images. *Front. Oncol.* **2021**, 11, 737368.
31. Hosny, A.; Aerts, H.J.; Mak, R.H. Handcrafted versus deep learning radiomics for prediction of cancer therapy response. *Lancet Digit. Health* **2019**, 1, e106–e107.
32. Wagner, M.W.; et al. Radiomics, machine learning, and artificial intelligence—What the neuroradiologist needs to know. *Neuroradiology* **2021**, 63, 1957–1967.
33. Afshar, P.; et al. From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities. *IEEE Signal Process. Mag.* **2019**, 36, 132–160.
34. Salmanpour, M.R.; et al. Robustness and reproducibility of radiomics features from fusions of PET-CT images. *J. Nucl. Med.* **2022**, 63, S2, 3179.
35. Khalid, H.; Sajid, S.M.; Nistazakis, H.E.; Ijaz, M. Survey on limitations, applications and challenges for machine learning aided hybrid FSO/RF systems under fog and smog influence. *J. Mod. Opt.* **2024**, 71, 101–125.
36. Asif, S.; et al. Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision. *Arch. Comput. Methods Eng.* **2024**, 1–31.
37. Kononenko, I. Machine learning for medical diagnosis: History, state of the art and perspective. *Artif. Intell. Med.* **2001**, 23, 89–109.

38. Salmanpour, M.R.; et al. Prediction of TNM stage in head and neck cancer using hybrid machine learning systems and radiomics features. *In Proc. Medical Imaging 2022: Computer-Aided Diagnosis (SPIE)*, San Diego, CA, USA, 20–24 Feb 2022; Volume 12033, pp. 648–653.
39. Othman, N.A.; Abdel-Fattah, M.A.; Ali, A.T. A Hybrid Deep Learning Framework with Decision-Level Fusion for Breast Cancer Survival Prediction. *Big Data Cogn. Comput*. **2023**, 7, 50.
40. Koller, D.; Friedman, N. Probabilistic Graphical Models: Principles and Techniques—Adaptive Computation and Machine Learning; MIT Press: Cambridge, MA, USA, 2009.
41. Bellazzi, R.; Zupan, B. Predictive data mining in clinical medicine: Current issues and guidelines. *Int. J. Med. Inform*. **2008**, 77, 81–97.
42. Jolliffe, I.T.; Cadima, J. Principal component analysis: A review and recent development. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, 374, 20150202.
43. Hasan, B.M.S.; Abdulazeez, A.M. A review of principal component analysis algorithm for dimensionality reduction. *J. Soft Comput. Data Min***. 2021**, 2, 20–30.
44. Rai, S.; Mishra, P.; Ghoshal, U.C. Survival analysis: A primer for the clinician scientists. *Indian J. Gastroenterol.* **2021**, 40, 541–549.
45. Deo, S.V.; Deo, V.; Sundaram, V. Survival analysis—Part 2: Cox proportional hazards model. *Indian J. Thorac. Cardiovasc. Surg*. **2021**, 37, 229–233.
46. Cui, L.; et al. A deep learning-based framework for lung cancer survival analysis with biomarker interpretation. *BMC Bioinform.* **2020**, 21, 112.
47. Saxena, K. Lung Cancer Dataset. Kaggle 2023. Available online: https://www.kaggle.com/datasets/khwaishsaxena/lung-cancer-dataset (accessed on 4 April 2025).