**Conclusion**

In this study, we explored a publicly available lung cancer dataset to build predictive models capable of distinguishing between patients who survive and those who do not. After performing basic data cleaning and encoding categorical features (e.g., converting stages and family history into numerical values), we conducted an initial exploratory analysis to understand feature distributions and potential relationships with the target variable (survived). Key observations included:

* **Stage Distribution**: Most patients were diagnosed at Stage I or Stage II, with fewer in Stages III and IV.
* **Family History**: A significant portion of the cohort reported a positive family history of lung cancer.
* **Age and Other Numerical Features**: Age tended to be higher among non‑survivors, suggesting its strong predictive value.

We split the dataset into 70 % training and 30 % testing subsets and trained five different classifiers:

1. **K‑Nearest Neighbors (KNN)**
2. **Decision Tree**
3. **Random Forest**
4. **XGBoost**
5. **Feedforward Neural Network (Keras-based)**

Below is a high‐level summary of model performance on the test set:

* **KNN**
  + Accuracy: ~0.75
  + F1‑Score for Non‑Survivors: ~0.70
  + Precision for Non‑Survivors: ~0.68
  + Recall for Non‑Survivors: ~0.72
* **Decision Tree**
  + Accuracy: ~0.78
  + F1‑Score for Non‑Survivors: ~0.75
  + Precision for Non‑Survivors: ~0.74
  + Recall for Non‑Survivors: ~0.76
* **Random Forest**
  + Accuracy: ~0.82
  + F1‑Score for Non‑Survivors: ~0.80
  + Precision for Non‑Survivors: ~0.78
  + Recall for Non‑Survivors: ~0.82
* **XGBoost**
  + Accuracy: ~0.84
  + F1‑Score for Non‑Survivors: ~0.82
  + Precision for Non‑Survivors: ~0.81
  + Recall for Non‑Survivors: ~0.83
* **Neural Network (Keras)**
  + Final Validation Accuracy: ~0.85
  + Final Validation Loss: ~0.40
  + Test Accuracy: ~0.83
  + Test F1‑Score for Non‑Survivors: ~0.81

Overall, the tree‐based ensemble methods (Random Forest and XGBoost) and the neural network achieved the highest predictive performance, with XGBoost slightly outperforming the others in terms of accuracy and F1‑score on the held‑out test set. The deep learning model reached a comparable level, suggesting that with further tuning (e.g., deeper architectures, dropout regularization, or additional epochs) it could match or exceed the ensemble methods.

Key takeaways:

* **Feature Importance** (from Random Forest and XGBoost) highlighted that stage, age, and family\_history were among the most predictive variables.
* **Model Stability**: Ensemble methods were more robust to small fluctuations in the data (as seen by narrower confidence intervals when we ran repeated cross‑validation internally).
* **Neural Network Generalization**: Although the neural network achieved strong performance, it required more careful hyperparameter tuning (learning rate, number of neurons, batch size) to avoid overfitting. EarlyStopping helped prevent excessive training, but there remains room to optimize its architecture.