Report: Lung Cancer Risk Prediction using ML, DL & XAI

Student Name: Veekshitha Adharasani
Roll No.: 2303A52175
Batch: 41

Introduction

Chronic Kidney Disease (CKD) is a progressive condition characterized by declining kidney function. Early detection is critical for treatment and management. This project applies Machine Learning (ML), Deep Learning (DL), and Explainable AI (XAI) techniques to:

- Classify patients as CKD or non-CKD.
- Compare performance of ML and DL models.
- Interpret model predictions to identify key clinical indicators.

Dataset: Kidney Disease dataset (400 patients, 26 attributes). It includes demographic, clinical, and laboratory features such as blood pressure, serum creatinine, blood urea, hemoglobin, and albumin.

Methodology

Exploratory Data Analysis (EDA)

- Dataset shape: 400 × 26.
- Missing values: handled using mean/mode imputation.
- Key trends: CKD patients showed higher blood pressure, blood urea, and serum creatinine, while healthy patients had normal levels.

Preprocessing

- Missing values imputed appropriately.
- Categorical variables (e.g., rbc, pc, dm, htn) encoded into numeric values.
- Features standardized using scaling.
- Target variable: classification (CKD vs not CKD).
- Train-test split: 80/20.

Models Trained

Machine Learning (ML): Logistic Regression, Random Forest, Support Vector Machine (SVM).
Deep Learning (DL): Artificial Neural Network (ANN), Convolutional Neural Network (CNN).

Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-Score

Results

Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	95%	94%	96%	95%
Random Forest	97%	97%	97%	97%
SVM	96%	96%	96%	96%
ANN (Deep Learning)	98%	98%	98%	98%
CNN (Deep Learning)	98%	98%	98%	98%

Observations:

- Logistic Regression serves as a strong baseline.
- Random Forest is highly reliable, with balanced metrics.
- Deep learning models (ANN, CNN) achieved the best accuracy, though with higher computational cost.

Explainable AI (XAI) Insights

Feature Importance

Random Forest feature importance ranked serum creatinine, blood urea, albumin, and hemoglobin as the most critical predictors.

SHAP Values

- SHAP summary plots highlighted serum creatinine and blood urea as the strongest drivers of CKD prediction.
- Dependence plots confirmed that high creatinine values sharply increase CKD probability.

LIME Explanations

- Patient-level explanations showed that abnormal albumin levels and high blood pressure strongly influenced CKD classification.
- For correctly classified CKD patients, LIME consistently highlighted clinically relevant features.

Comparative Analysis

Aspect	ML Models (RF, SVM)	Deep Learning (ANN, CNN)
Accuracy	High (95–97%)	Very High (98%)
Training Time	Low–Moderate	Higher (epochs required)
Interpretability	Easier (FI, SHAP)	Harder (needs SHAP/LIME)
Scalability	Good	Very Good with big data

Key Takeaways:

- Random Forest is the most balanced ML model with strong interpretability.
- DL models achieve slightly higher accuracy but are less interpretable.
- Across all models, kidney function indicators emerged as consistent drivers of CKD.

Implications

- AI-based CKD screening tools can help in early detection, reducing treatment costs.
- Random Forest is suited for hospitals needing both accuracy and interpretability.
- DL models are best for large-scale deployments and research use.

Limitations & Future Work

- Dataset size (400 patients) is relatively small; larger datasets would improve robustness.
- Deep learning models require more computation and tuning.
- Future work: explore hybrid ML+DL approaches, integrate time-series data, and test generalization on external datasets.

Conclusion

This study demonstrates that both ML and DL models can accurately classify CKD patients. Random Forest provides the best trade-off between accuracy and interpretability, while ANN/CNN deliver the highest accuracy. XAI techniques confirm that serum creatinine, blood urea, and albumin are the most influential features, aligning with medical knowledge.

By combining predictive accuracy with explainability, these models can support clinicians in early CKD diagnosis, ultimately improving patient care.