AI-Driven Drug Sensitivity Prediction in Cancer Cell Lines for Precision Medicine

Valeria V. Mudzindiko & Miltone Awiti
Course: Artificial Intelligence in Healthcare – SAT 5114
Instructor: Dr. Guy Hembroff | Group 29

Background

- Cancer treatment is shifting towards precision medicine, which tailors therapy to individual genetic profiles.
- Predicting drug sensitivity helps oncologists select the most effective treatment for each patient.
- The Genomics of Drug Sensitivity in Cancer (GDSC) dataset provides a robust foundation with genomic and pharmacological data from 1,002 cancer cell lines and 621 drugs.
- Machine learning (ML) can identify key markers and patterns in the data to predict treatment outcomes.

Study Objectives

- Develop and compare ML models that predict cancer cell line responses to therapeutic compounds.
- Identify genetic and molecular biomarkers that influence drug sensitivity.
- Evaluate model performance using clinical-relevant metrics.
- Provide interpretability and transparency using SHAP values to guide clinical use.

Literature Review

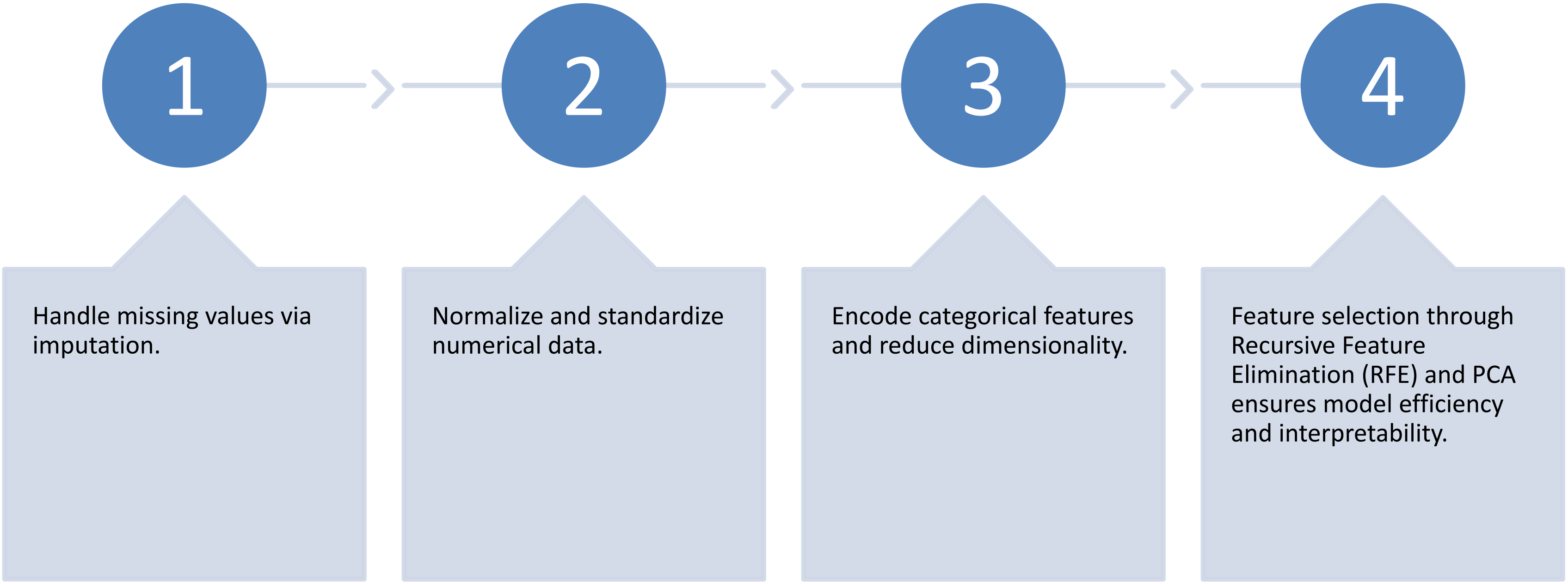
- AI models outperform traditional statistical methods in predicting treatment response (Quazi et al., 2022).
- Deep learning enhances predictions of how lung cancer patients respond to specific drugs (Cortes-Ciriano et al., 2022).
- Challenges include data imbalance, lack of model interpretability, and computational demands.
- Tools like SHAP and LIME offer potential solutions to explain 'black-box' AI decisions.

Dataset Description

- GDSC1 and GDSC2 datasets include over 484,000 samples, 1,002 cancer cell lines, 621 compounds
- Data includes IC50 values, gene expression, mutation status, and drug response profiles.
- Provides high-dimensional input for building robust ML models.
- Source: Kaggle and CancerRxGene official site.

https://www.kaggle.com/code/samiraalipour/genomics-of-drug-sensitivity-in-cancer

Data Preprocessing



1. Handle missing values via imputation.
2. Normalize and standardize numerical data.
3. Encode categorical features and reduce dimensionality.
4. Feature selection through Recursive Feature Elimination (RFE) and PCA ensures model efficiency and interpretability.

AI Models Used

- Logistic Regression: Baseline binary classification model.
- Random Forest: Ensemble of decision trees; high accuracy and feature importance interpretation.
- XGBoost: Boosted trees optimized via gradient descent; state-of-the-art in structured data.
- SVM: Effective in high-dimensional spaces.
- KNN: Simple, distance-based classifier for comparison purposes.

Model Evaluation Metrics

Accuracy: Overall correctness.

Precision: Proportion of predicted positives that are correct.

Recall: Ability to find all relevant positive cases.

F1 Score: Harmonic mean of precision and recall.

Confusion Matrix: Shows TP, TN, FP, FN to evaluate performance.

SHAP & Interpretability

SHAP (SHapley Additive exPlanations) values show how much each feature contributes to a prediction.

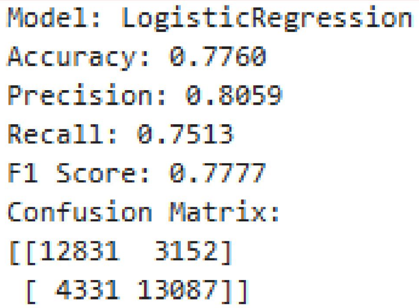
Helps explain model outputs for clinical interpretability.

Supports trust in AI predictions by clinicians.

Key features like EGFR mutations and TP53 status show strong influence on drug sensitivity.

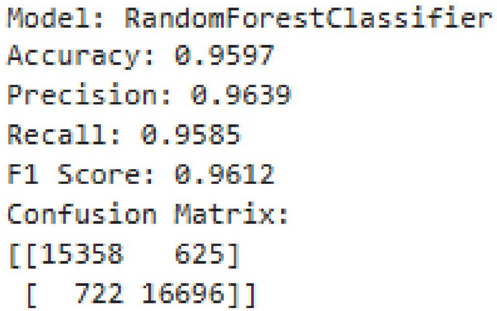
Performance Metrics for selected models

Logistic Regression



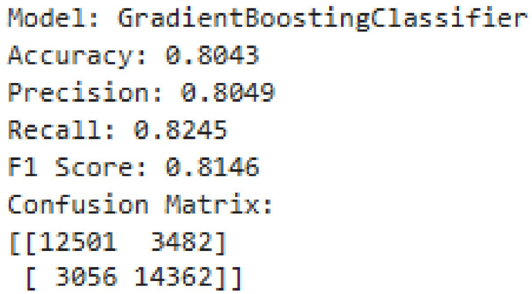
```
Model: LogisticRegression
Accuracy: 0.7760
Precision: 0.8059
Recall: 0.7513
F1 Score: 0.7777
Confusion Matrix:
[[12831  3152]
 [ 4331 13087]]
```

Random Forest



```
Model: RandomForestClassifier
Accuracy: 0.9597
Precision: 0.9639
Recall: 0.9585
F1 Score: 0.9612
Confusion Matrix:
[[15358   625]
 [  722 16696]]
```

Gradient Boosting Classifier



```
Model: GradientBoostingClassifier
Accuracy: 0.8043
Precision: 0.8049
Recall: 0.8245
F1 Score: 0.8146
Confusion Matrix:
[[12501  3482]
 [ 3056 14362]]
```

Results Summary

- Random Forest achieved highest predictive performance.
- XGBoost also showed competitive results.
- SHAP identified top features contributing to model predictions.
- Tissue-specific accuracy revealed variability in prediction strength across cancer types.

Future Work

- Incorporate deep learning models for more complex feature learning.
- Address class imbalances using SMOTE or weighted losses.
- Integrate multi-omics datasets to improve prediction.
- Develop an AI-powered clinical decision support system (CDSS).

Conclusion

- AI models significantly enhance drug sensitivity prediction in cancer.
- Random Forest and XGBoost models performed best.
- Project supports precision oncology by matching treatment to genomic profiles.
- Future efforts will aim at deployment in clinical settings.