Machine Learning-Based Forecasting of Study Completion in Oncology Trials

MGIntelligence

About MGIntelligence

- Accelerating drug discovery with cutting-edge generative AI
- Empowering pharma R&D through predictive analytics & quantum insights
- Designing novel molecules with AI-guided creativity
- Predicting clinical trial outcomes with data-driven accuracy
- Optimizing real-world performance across the drug lifecycle
- One unified platform for AI-powered innovation in healthcare

MGIntelligence empowers researchers, biotech firms, and pharma leaders to make faster, smarter decisions-with AI at the core.

Objective

Objective

- Develop an AI model to predict the final status of oncology clinical trials using structured metadata (e.g., sponsor, phase, enrollment, intervention, location).

Why It Matters

- Clinical trial failure is costly — up to 50% fail due to design flaws, recruitment issues, or feasibility challenges.

AI-driven early prediction empowers stakeholders to:

- Accelerate drug development
- Reduce R&D costs
- Prioritize high-potential trials
- Optimize trial design & site selection
- Support smarter investment decisions

Data Source

- Source- ClinicalTrials.gov - The largest publicly accessible database for clinical trials
- Total Studies: 282 oncology trials (CAR-T related)
- Therapeutic Area: Multiple Myeloma
- Geography: Global (multi-country trial data)

Sample Data

Conditions	Interventions	Sponsor	Phases	Enrollment	Primary Completion Time	Locations	Study Status
Refractory Multiple Myeloma|Relapsed Multiple Myeloma	BIOLOGICAL: Manufactured Anti-BCMA CAR-T cells|DRUG: Fludarabine|DRUG: Cyclophosphamide	Thomas Martin, MD	PHASE1	5	2570	University of California, San Francisco, San Francisco, California, 94143, United States	ACTIVE_NOT_RECRUITING
Multiple Myeloma	OGICAL: Anti-BCMA CAR-T cells|DRUG: Fludarabine|DRUG: Cyclophosphamide|DRUG: Immune inhit	Hrain Biotechnology Co., Ltd.	EARLY_PHASE1	10	847	Shanghai Changzheng Hospital, Shanghai, Shanghai, 200003, China	UNKNOWN
Myeloma-Multiple|Myeloma, Plasma-Cell	amide|DRUG: Fludarabine|BIOLOGICAL: Anti-B Cell Maturation Antigen (BCMA) chimeric antigen rec	National Cancer Institute (NCI)	PHASE1	35	1570	National Institutes of Health Clinical Center, Bethesda, Maryland, 20892, United States	ACTIVE_NOT_RECRUITING
Multiple Myeloma	DRUG: T cell infusion agent targeting BCMA chimeric antigen receptor	PersonGen BioTherapeutics (Suzhou) Co., Ltd.	EARLY_PHASE1	3	70	No.3, Qingchun East Road, Hangzhou, Zhejiang, 310020, China	COMPLETED

Variables for ML Classification

Table 1: Features for ML Classification

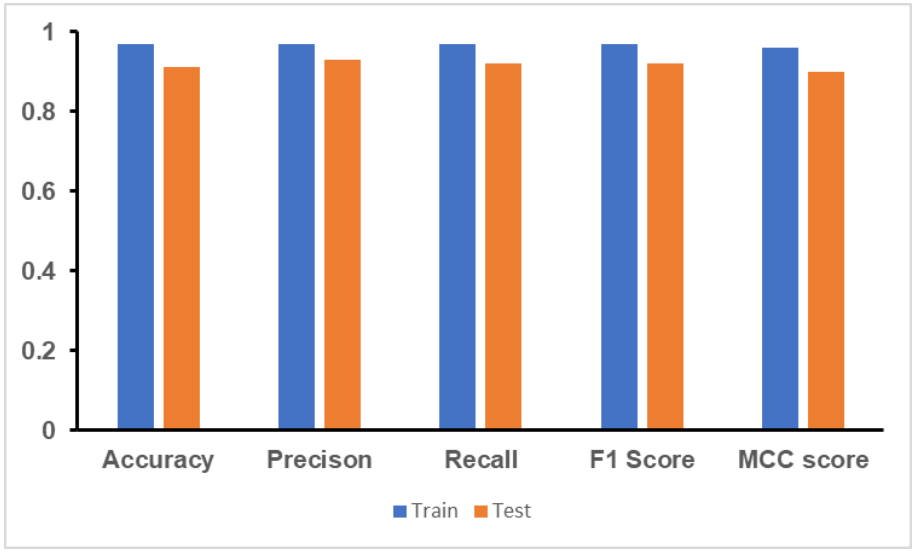
Independent Variable	**Dependent Variable**
Conditions	Study Status
Interventions	
Sponsor	
Phases	
Enrollment	
Primary completion time	
Locations	

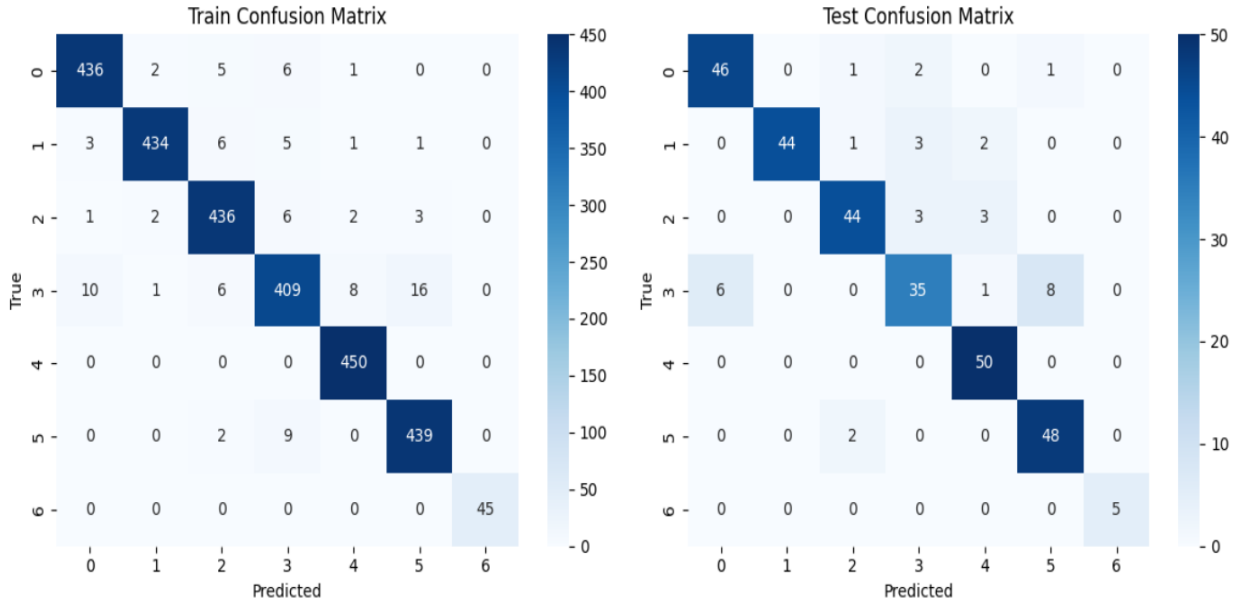
Table 2: Target Labels and Class Encoding

Study Status	Classification
Active-not recruiting	0
Completed	1
Not yet recruting	2
Recruiting	3
Terminated	4
Unknown	5
Withdrawn	6

Model Performance – Random Forest Algorithm

Objective: To forecast the final status of oncology clinical trials using Random Forest on structured metadata.





Performance Metrics (Train vs. Test):

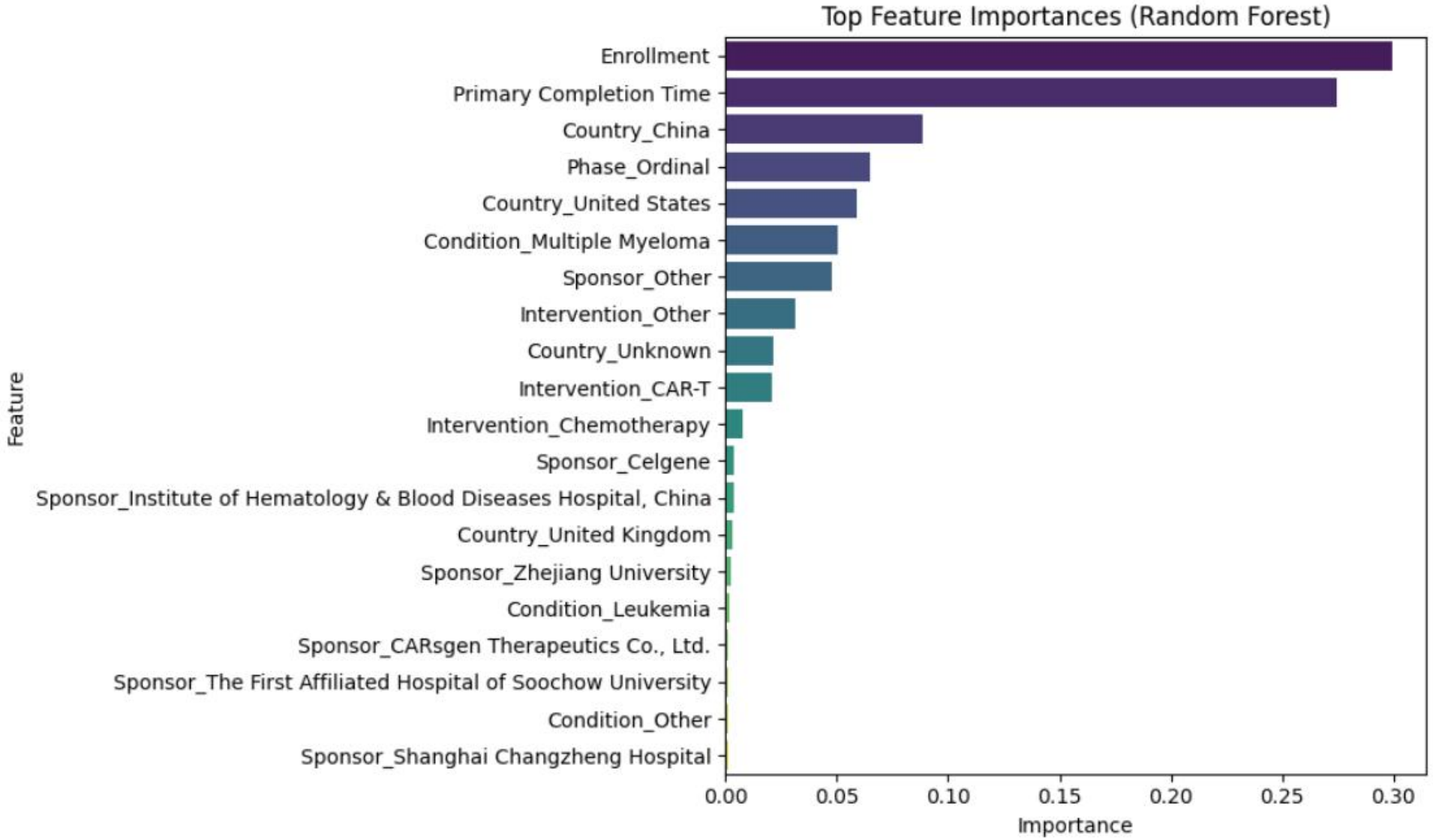
- **Accuracy:** ~91% (Test)
- **Precision, Recall, F1 Score, MCC** all consistently high, indicating robust generalization and minimal overfitting.

Confusion Matrices:

- **Train Set:** High prediction accuracy across all status categories
- **Test Set:** Strong generalization with balanced classification across multiple trial statuses

Random Forest shows strong potential for real-world deployment in early-stage trial risk assessment and portfolio prioritization.

Top Predictive Features Identified by Random Forest



Enrolment is the most important feature in predicting study status

Key Takeaways & Clinical Relevance

- Successfully developed a machine learning model to forecast trial outcomes in oncology using publicly available metadata.
- Random Forest achieved high predictive performance (~91% accuracy), with robust generalization.
- Key predictors include: enrollment, completion time, and sponsor type — all critical factors in feasibility and planning.
- This approach enables early identification of high-risk trials, helping sponsors and CROs save costs and improve development strategy.

Strategic Value for Sponsors & CROs

- **Early risk identification: Detect high-risk trials before resource commitment**
- **Trial design optimization: Tailor protocols, duration, and site strategy to boost feasibility**
- **Portfolio prioritization: Focus on trials with high predicted success likelihood**
- **Cost efficiency: Reduce sunk costs by deprioritizing likely-to-fail studies**
- **Data-driven decisions: Integrate model insights into strategic planning and investment**

Future Directions

- Expand modeling to other therapeutic areas (e.g., immunology, rare disease)
- Integrate unstructured data (e.g., trial protocols, publications) using NLP
- Develop an interactive dashboard for trial risk scoring and portfolio insights
- Offer this as a custom analytics service to pharma/CRO partners.

Contact or Call to Action

Website: www.mgintelligence.org

Email: gayathrikg@mgintelligence.org