



A Novel Model Using ML Techniques for Clinical Trial Design and Expedited Patient Onboarding Process

Abhirvey Iyer ¹, Sundaravalli Narayanaswami ²

¹Department of Pharmaceutical Engineering & Technology, Indian Institute of Technology (BHU) Varanasi, Varanasi, Uttar Pradesh, India; ²Public Systems Group, Indian Institute of Management Ahmedabad, Ahmedabad, Gujarat, India

Correspondence: Sundaravalli Narayanaswami, Email sundaravallin@iima.ac.in

Introduction: Clinical trials are critical for drug development and patient care; however, they often need more efficient trial design and patient enrolment processes. This research explores integrating machine learning (ML) techniques to address these challenges. Specifically, the study investigates ML models for two critical aspects: (1) streamlining clinical trial design parameters (like the site of drug action, type of Interventional/Observational model, etc) and (2) optimizing patient/volunteer enrolment for trials through efficient classification techniques.

Methods: The study utilized two datasets: the first, with 55,000 samples (from ClinicalTrials.gov), was divided into five subsets (10,000–15,000 rows each) for model evaluation, focusing on trial parameter optimization. The second dataset targeted patient eligibility classification (from the UCI ML Repository). Five ML models—XGBoost, Random Forest, Support Vector Classifier (SVC), Logistic Regression, and Decision Tree—were applied to both datasets, alongside Artificial Neural Networks (ANN) for the second dataset. Model performance was evaluated using precision, recall, balanced accuracy, ROC-AUC, and weighted F1-score, with results averaged across k-fold cross-validation.

Results: In the first phase, XGBoost and Random Forest emerged as the best-performing models across all five subsets, achieving an average balanced accuracy of 0.71 and an average ROC-AUC of 0.7. The second dataset analysis revealed that while SVC and ANN performed well, ANN was preferred for its scalability to larger datasets. ANN achieved a test accuracy of 0.73714, demonstrating its potential for real-world implementation in patient streamlining.

Discussion: The study highlights the effectiveness of ML models in improving clinical trial workflows. XGBoost and Random Forest demonstrated robust performance for large clinical datasets in optimizing trial parameters, while ANN proved advantageous for patient eligibility classification due to its scalability. These findings underscore the potential of ML to enhance decision-making, reduce delays, and improve the accuracy of clinical trial outcomes. As ML technology continues to evolve, its integration into clinical research could drive innovation and improve patient care.

Keywords: clinical trials, site selection, machine learning, patient onboarding, feature engineering, feature encoding

Overview

In this paper, we present novel research that leverages machine learning (ML) models and techniques to automate the outcome prediction of clinical trials. Our study is motivated to combine two crucial aspects, namely, the streamlined selection process of the site of action for a new drug and the optimization of patient enrolment in clinical trials. This unique combination provides an end-to-end solution to proceed with Phase 1 of clinical trials, effectively addressing the limitations that can impede the success of the trial process. By improving the target site selection process, the probability of successful completion of clinical trials increases with minimum system time and spent resources¹ of pharmaceutical companies and researchers, in addition to ensuring the improved safety of patients enrolled in the trials. The model presented in this paper not only enhances the site selection process but also aims to streamline the patient enrolment process, directly targeting the challenges associated with low accrual rates and enrolment inefficiency reported in global

statistical analyses of terminated trials within clinical trials databases.² The empirical results derived from our model are presented, demonstrating its efficacy in addressing these critical issues and providing a comprehensive solution for enhancing the efficiency and success rates of clinical trials.

To establish a robust test bed, we collected and analysed data from 273,254 terminated or completed studies obtained from the ClinicalTrials.gov site. This dataset served as the foundation for constructing a test bed of 55,000 samples, encompassing trials conducted across nations such as Australia, Canada, India, France, USA, UK, and Switzerland, among others. Employing feature engineering, ensemble learning, and the tf-idf technique, we achieved a balanced accuracy score of 71% and an Area Under the Curve (AUC) of 0.70, in determining the outcomes of a clinical trial, which further enhances the site of action selection process.

Finally, to streamline patient enrolment, we acquired a dataset consisting of information from 600 patients, focusing specifically on liver disease conditions. Within this context, we employed ensemble learning, feature selection, and artificial neural networks to develop an algorithm to assess patient eligibility for clinical trials targeting exclusively on liver-related ailments. Bespoke eligibility criteria were incorporated in the algorithm, enabling an efficient eligibility determination of patient records. Our results are impressive, if not promising, with a 73% test accuracy, in showcasing its potential for automating and optimizing the patient enrolment process in clinical trials.

Introduction and Background

Clinical trials are scientific studies that evaluate the safety, effectiveness, and potential side effects of new drugs, treatments, or medical interventions in humans. They are critical for advancing medical knowledge, improving patient care, and securing regulatory approvals for new therapies. These trials follow well-structured protocols and involve rigorous data collection and analysis to assess the benefits and risks of investigational interventions. Typically, clinical trials progress through multiple phases, starting with small groups of participants and expanding to larger populations. Their outcomes contribute to evidence-based healthcare and drive advancements in medical science.

Clinical Trials can be broadly be divided into 2 categories, Observational and Interventional trials.³

Observational trials involve the observation and collection of data from individuals in real-world settings, without any intervention or modification of their treatment. These studies aim to understand the natural course of diseases, identify risk factors, determine prevalence rates, and explore associations between variables. Observational trials rely on existing data or prospectively collect data over a specific period, using methods such as surveys, medical records, or registries. They can provide valuable insights into disease patterns, treatment outcomes, and potential adverse effects.

Interventional trials, on the other hand, involve actively intervening or assigning different interventions to participants in a controlled and systematic manner. The primary objective is to evaluate the safety, efficacy, and optimal use of new drugs, treatments, or medical interventions. These trials follow carefully designed protocols, including randomization, control groups, and blinding, to minimize bias and establish causality. They are typically conducted in multiple phases, starting from small-scale safety assessments and progressing to larger-scale efficacy evaluations involving diverse populations.

The successful development of new drugs heavily relies on rigorous clinical trials that evaluate their safety, efficacy, and optimal site of action. Our model, designed for post-preclinical trials or post-phase 0 of clinical trials, offers pharmaceutical companies and researchers a valuable means of validating the viability of their findings. Recognizing the significance of comprehending pharmacokinetics (study of how a drug is absorbed, distributed, metabolized, and eliminated by the body. It examines the processes that influence the drug's concentration in the bloodstream over time, as well as its movement and interaction within various tissues and organs) at the site of action and providing concrete evidence of target engagement is crucial not only for scientific purposes but also for enhancing the effectiveness of pharmaceutical research and development.⁴ By utilizing our model, companies can gain insights into whether their proposed trial design (especially selection of site of action of the drug) and patient onboarding processes align with historical data from reputable clinical trial databases. This validation process aids in making informed decisions, ensuring that the trial progresses with a higher probability of success while minimizing potential risks, optimizing resource allocation, and improving overall efficiency. By incorporating our model into their workflow, pharmaceutical companies can enhance the accuracy and reliability of their findings, fostering a more robust and efficient drug development process.

Patient enrolment in clinical trials poses its own set of challenges.⁵ Low accrual rates and inefficient eligibility assessment processes contribute to delays, increased costs, and, in many cases, premature termination of trials. The termination of trials due to insufficient patient participation has been reported as a significant problem, accounting for a substantial proportion of trial terminations worldwide. Failure to vigorously recruit and retain patients is also a common factor to creating a menace in clinical trials.⁶ Moreover, manually reviewing large patient datasets to determine eligibility is a laborious and time-consuming task, further exacerbating these challenges.

These problems have persisted over the years; thus, the application of machine language, artificial intelligence and big data analytics is the need of the hour for pharmaceutical companies.⁷ It is also the right approach to resolve the persistent challenges, as substantive historical data on premature termination and failed clinical trials are available.

Our empirical results showcase the potential benefits for industry stakeholders at various stages, including trial planning, pre-trial preparation, and patient enrolment. By leveraging advanced analytics, ML and AI, our research presented in this paper aims to streamline patient enrolment and optimize the tedious procedures involved in trials, enhancing the usefulness as a decision support system in clinical trials. The utilization of this system will improve trial planning, protocol optimization, and patient enrolment processes, ultimately benefiting pharmaceutical companies, researchers, healthcare professionals, and patients involved in clinical trials.

Related Works

A prior study was conducted to identify the prevalent markers or factors linked to the termination of clinical trials and to develop an accurate predictive model for determining whether a trial will be terminated or completed.⁸ This study utilized a dataset of 311,260 trials to construct a testbed comprising 68,999 samples. Subsequently, feature engineering techniques were employed to generate 640 distinct features. Through the implementation of sampling methods and ensemble learning, the research achieved a balanced accuracy of 67% and an area under the curve of 0.73.⁸ These results underscore the significance of employing machine learning models in clinical trial analysis. While this research paper shares a similar goal of emphasizing the importance of machine learning (ML) in clinical trials, it is important to note that the objectives of this study and the aforementioned paper diverge in their specific focuses. This paper⁸ primarily centres on identifying common markers associated with trial termination, whereas our work places greater emphasis on enhancing the site of action selection process for a clinical trial by utilizing the trial's completion or termination status. Moreover, our research also prioritizes the streamlining of the patient on-boarding process.

Another previous study aimed to develop an algorithm to assess the risks of trial termination by analysing patterns in the language used to describe the study before its implementation.⁹ Data was collected from the ClinicalTrials.gov repository, including structured data indicating study characteristics and unstructured text data providing narrative descriptions of study goals, objectives, and methods. The study proposed an algorithm to extract distinctive words from the unstructured text data, which were frequently used in successfully completed trials versus terminated trials. These distinctive words, along with structured data, were input into a random forest model. The combined approach yielded robust predictive probabilities with respect to sensitivity (0.56) and specificity (0.71) compared to a model using only structured data (sensitivity=0.03 and specificity=0.97).⁹ Our work also incorporates both structured and unstructured data but does not extensively emphasize the significance of unstructured data.

Another study proposed a machine learning pipeline to optimize clinical trial design by predicting the probability of early termination and identifying key features driving such terminations.¹⁰ The study collected data from 420,268 clinical trials registered in ct.gov, focusing on 24 specific columns. Through feature engineering and ensemble methods, the research achieved a balanced accuracy of 0.7 and a Receiver Operator Characteristic Area under the curve score of 0.8. The study also utilized Shapley Additive Explanations to interpret termination predictions and highlight feature contributions.¹⁰ The proposed pipeline has the potential to improve clinical trial design, facilitating the efficient delivery of potentially life-saving treatments to patients. While this study emphasizes on enrolment issues and study design criteria, our work focuses on target site selection and streamlining patient on-boarding process.

In order to contextualize the contribution of our work, we conducted a thorough comparison with previous literature in the field. The results of this comparison are summarized in [Table 1](#)

Table 1 Table Illustrating the Comparison of Our Work With Previous Literatures, Our Work Distinguishes Itself From Previous Literature by Focusing on the Unique Vision of Facilitating Target Site Selection and Expediting Patient Onboarding

Research	Objective	Approach	Data size/ sampling	Training Methods	Results	Novelty/Contribution
[8]	Utilize ML techniques to predict the likelihood of trial termination.	ML Classifiers, Feature Engineering, Class imbalance handling, and ensemble learning.	i) 311,260 trials from ClinicalTrials.gov were used ii) A testbed of 68,999 samples iii) 640 features were engineered	Neural Networks, Random Forest, XGBoost, and Logistic Regression	i) ROC-AUC: 0.73 ii) Balanced Accuracy: 0.67	The study contributes by identifying key factors for clinical trial termination, highlighting the importance of statistics and keyword features, and demonstrates the predictive capabilities of various models, including ensemble methods, for accurate termination prediction.
[9]	To quantify risk associated with clinical trial termination using text mining.	Text Preprocessing, Feature Engineering, ML, and TF-IDF features.	i) The study utilized data from the CTTI, which consisted of around 250,000 trials ii) The analysis focused on approximately 130,000 trials that began before May 1, 2015	Random Forest Model	i) The combined approach showed 90% improvement over models using only structured data	The study contributes by combining structured and unstructured data to predict trial termination risk, highlighting the importance of incorporating derived terms from unstructured data.
[10]	To optimize trial design, minimize resource waste, and expedite the availability of life-saving treatments by predicting early trial termination using ML.	ML, Feature Engineering, SHAP, Threshold-based decision making, and iterative optimization.	i) A CSV version of the ClinicalTrials.gov, containing 420,268 trials was extracted	Logistic Regression, XGBoost, and Random Forest	i) ROC-AUC: 80% ii) Balanced Accuracy: 70% iii) F1-Score: 42%	The study contributes by enhancing prediction performance for early trial termination, and providing insightful suggestions for optimizing trial design using SHAP explanations.
Proposed work in this paper	To predict clinical trial outcomes, hence optimizing site selection and expedite patient enrolment.	ML Classifiers, ANN, ensemble learning, feature engineering, and TF-IDF technique.	i) 273,254 terminated or completed studies ii) A testbed of 55,000 trials was used iii) A 583 liver-patient dataset	XGBoost, Decision Trees, Random Forest, SVC, Logistic Regression, and ANN	i) ROC-AUC: 0.70 ii) Balanced Accuracy: 71% iii) Test accuracy: 73% (Patient dataset)	The study contributes by enhancing the target site selection process for a trial and expediting the patient on-boarding process.

Contribution

Research proposed in this study is motivated to ease the site selection of a new drug, so that the predictability of clinical trial completion is higher for both pharma companies and researchers. This is a significant new approach in clinical trial

research, based on published literature. Attempts are also made to streamline and expedite the patient onboarding process by applying and benchmarking different ML models for efficacy. The main contribution of the study is as follows:

- Large scale clinical trial studies: A large data of 273,254 terminated or completed studies obtained from the ClinicalTrials.gov site¹¹ was analysed. This dataset served as the foundation for constructing a test bed of 55,000 samples; regions covered in the trials were representative across nations such as Australia, Canada, India, France, USA, UK, and Switzerland, among others.
- Selection of site of action of a new drug: The distinct objective of our research paper is to simplify the site selection process for a new drug. The database¹¹ contains columns such as “Title” and “Conditions”, which provide information about the site where the drug molecule will act to treat a specific condition. To convert this textual data into numerical data, the tf-idf technique was employed. The information in these columns has a significant impact on the outcome of a clinical trial, whether it is terminated or completed. By automating this process using machine learning models at an early stage of trial design, predictability of trial completion and termination is significantly enhanced. Similar to several existing studies,^{8–10} various ML algorithms are applied for the clinical trial process and benchmarked. Numerical results of the proposed study (average balanced accuracy of 0.71 and score of 0.70) are comparable with the best of those in published literature; additionally results are also targeted in terms of locational accuracy, making our model more relevant for both pharma companies and researchers.
- Streamlining patient enrolment process: A dataset comprising the details of 600 patients, specific to liver related complications was collected for this purpose. Ensemble learning, feature selection techniques, and artificial neural networks were applied to develop an algorithm to evaluate patient eligibility for clinical trials focused only on liver-related conditions. By incorporating predefined eligibility criteria into the algorithm, the proposed algorithms are more proficient in analysing patient records and better in determining patient eligibility. This aspect of pre-specifying eligibility criteria has helped achieve a test accuracy of 73%, demonstrating further its potential in automating and optimizing the patient enrolment process.

Methods and Materials

Data

The initial dataset for our first area of focus, namely the selection of target sites by predicting trial outcomes, was sourced from ClinicalTrials.gov,¹¹ a public data repository managed by the United States government. We followed the guidelines established by ClinicalTrials.gov when conducting our research and utilized a database of 273,254 trials. It is crucial to acknowledge that the data obtained from ClinicalTrials.gov might not represent the entire landscape of clinical trials, as it primarily includes trials that have been registered or reported. This selective inclusion could introduce potential biases, leading to an underrepresentation of certain types of trials, such as those with negative or inconclusive results. Consequently, the completion rate derived from this dataset may overestimate the true completion rate of all trials, creating an overly optimistic perception of the success of clinical trials.

For our analysis, we created a testbed comprising 55,000 trials from various countries, including Australia, Canada, India, France, Switzerland, the USA, and the UK, to ensure a diverse global representation. It is important to note that trials included in the dataset were classified as “Completed” or “Terminated” based on the criteria established by ClinicalTrials.gov, which generally account for trials meeting their predefined objectives or endpoints. However, it is essential to exercise caution in the interpretation of these classifications, considering the potential biases introduced by the selective reporting of trials. Researchers should be mindful of the specific trial design criteria, such as the rejection of the null hypothesis or the confirmation of not reaching the minimal effect size specified in the sample size calculation, to accurately gauge trial completeness. It is imperative to promote transparency and encourage the reporting of all trial outcomes, irrespective of their perceived success or failure, to ensure a comprehensive and accurate representation of the clinical trial landscape.

The focus of our research is to use structured and unstructured data together to derive a better trial completion and termination accuracy by incorporating a site selection process in the ML models.

Figure 1 depicts the number of completed and terminated trials in the testbed consisting of 55,000 samples. It provides an overview of the distribution of trial outcomes, distinguishing between those that were successfully completed and those that were terminated prematurely.

Figure 2 illustrates the number of studies categorized as interventional trials and the number of trials categorized as observational in nature. This provides an overview of the distribution of trial types within the dataset, distinguishing between interventional studies that involve interventions or treatments and observational studies that primarily observe and collect data without intervening in the participants' treatment or conditions. It is important to note that the dataset encompasses both Observational and Interventional trials, with certain observational studies focusing on diagnosis-oriented research. The categorization of trials as observational or interventional has been facilitated by the classification provided by ClinicalTrials.gov, known for its credible and comprehensive database of clinical trials. Additionally, while the term "randomization" was not explicitly mentioned, the inclusion criteria were determined based on the data availability and relevance to the treatment-oriented focus of the study.

The "Title" and "Conditions" columns are essential components in our research, as they contain textual data that provides crucial information about the site of drug action. For instance, one example of a title is "Three Instructional Interventions for Prehospital Cervical Spinal Immobilization by Laypeople", accompanied by the corresponding condition "Cervical Vertebra

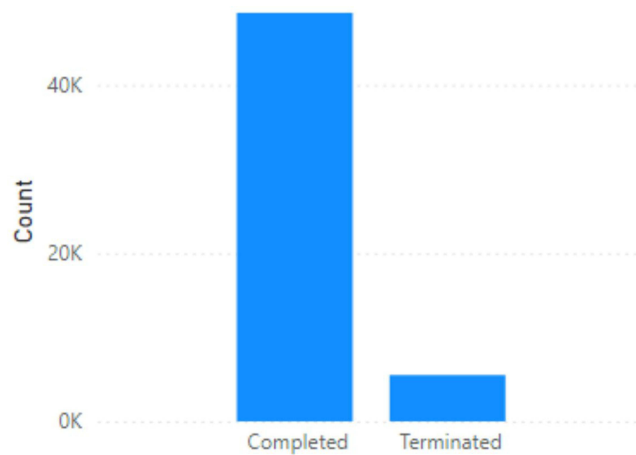


Figure 1 Number Of Completed Trials V/S Number Of Terminated Trials In The First Dataset.

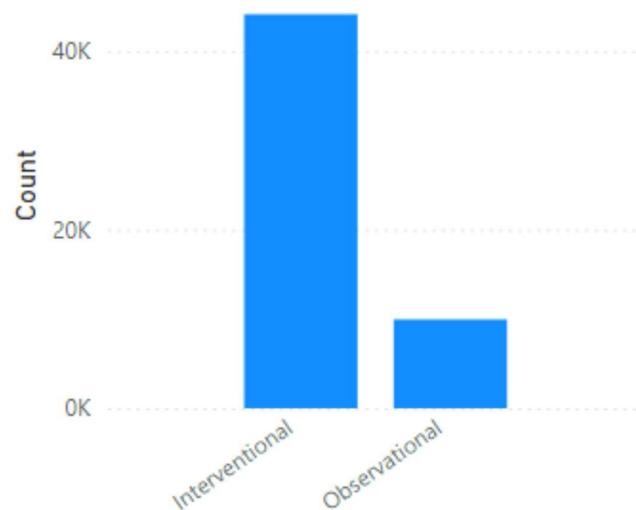


Figure 2 Number of interventional studies v/s number of observational studies in the first dataset.

dataset has been chosen for its accessibility and convenience, allowing us to demonstrate the proposed approaches for expediting patient onboarding in clinical trials. As mentioned earlier, the expediting patient onboarding process requires setting parameters by clinicians based on the trial requirements. The liver dataset serves as an illustrative example to showcase the functionality of the model in this context.

This dataset consists of 583 individual health records, where each record corresponds to a patient's liver condition. There are a total of 11 columns or features associated with each record. Out of the 583 records, 416 patients have been diagnosed with some liver disease, while the remaining 167 records represent liver-healthy patients. Our research is aimed at leveraging this dataset to devise strategies that address the challenges associated with low patient accrual rates, ultimately contributing to the optimization of the patient onboarding process and potentially mitigating the premature termination of clinical trials.

Figure 7 provides a comprehensive list of the different reasons attributed to trial termination; the most probable and known reason for a termination trial is low accrual rate. A low accrual rate refers to an inadequate number of participants enrolment within the specified timeframe, indicating that the clinical trial completion is likely to be hindered. A high or low accrual rate is an outcome of adequacy in recruiting sample size; which can arise due to various factors such as stringency in eligibility criteria, limited patient population, lack of awareness or willingness among potential patients, higher lead time in patient onboarding, etc.¹³

In this research, an attempt is made to expedite the patient onboarding process, by structured analysis of various patient details through a defined set of eligibility criteria. It is observed that several columns (or features) in the dataset are significant in defining those eligibility criteria and therefore, cannot be ignored. These include “Total_Bilirubin REAL”, “Direct_Bilirubin REAL”, “Alkaline Phosphatase”, and others. Such columns or features provide the numerical indicators of the levels of Bilirubin and other essential proteins in the liver, which further help in predicting whether a patient is affected by a liver disease or not.

Feature Engineering and Feature Encoding

Once the datasets are cleaned up for the stated objectives, the next step in the research process is to identify the optimal features or columns within the dataset. A rigorous feature selection process is employed to remove bias (redundancy) and

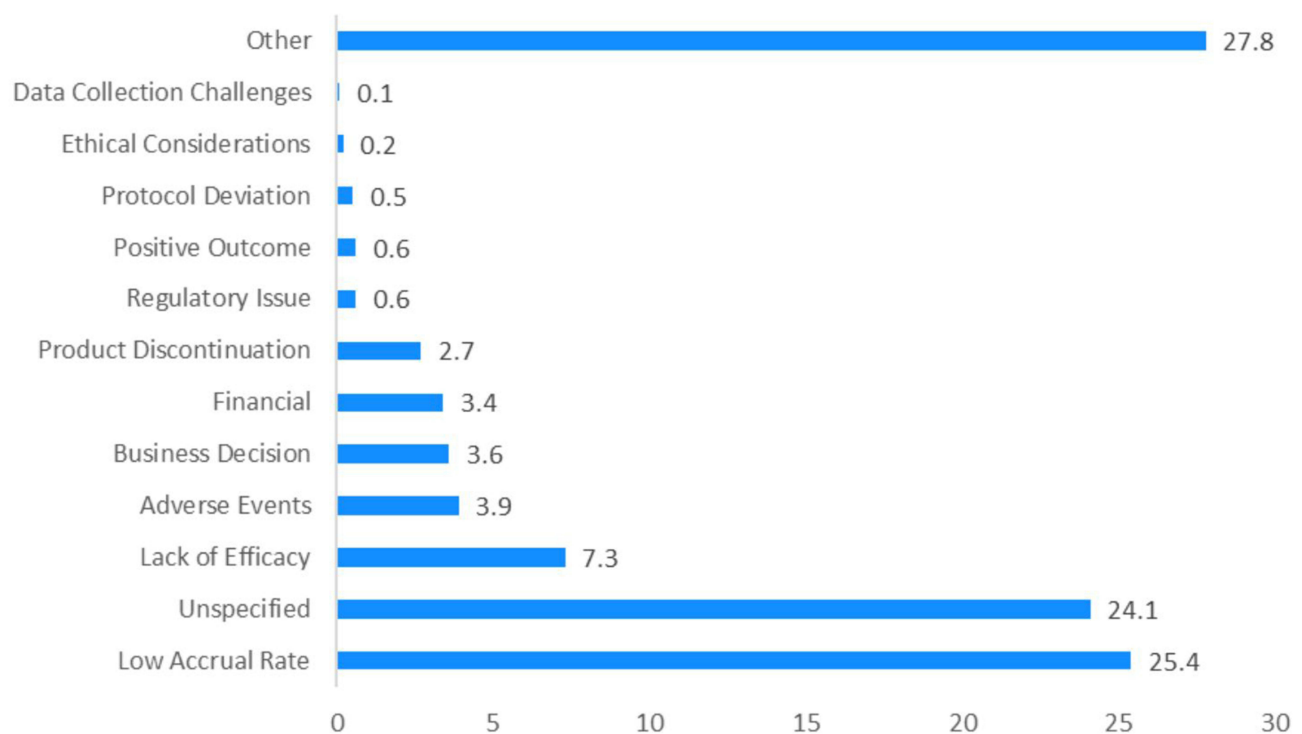


Figure 7 Publicly available information on reason for termination for all terminated drug trials between 2010–2021¹³.

to draw data that can enhance accuracy, integrity and reliability of the outcomes.¹⁴ It is essential for clinical data as it improves interpretability by removing unnecessary parameters, making a clinician’s job easy, and providing only a select set of relevant and important features to monitor to achieve better results. Hence, this step is crucial in our proposed pipeline.

Through our initial Pearson analysis, we found out the features ‘Study Designs’, ‘Title’, and ‘Conditions’ to be amongst the most important features in determining the outcome of a trial.

The column ‘Study Designs’ contains vital information related to the design of a clinical trials; if this column is split into additional columns/features with more granular data, it is hoped for better results accuracy. For example, in an ‘Interventional’ study, the ‘Study Designs’ column contains the following information, ‘Allocation: Non-Randomized|Intervention Model: Single Group Assignment|Masking: None (Open Label)|Primary Purpose: Health Services Research’, and in an ‘Observational’ study the column contains the following information, ‘Observational Model: Case-Only|Time Perspective: Prospective’. Therefore, we recreate the dataset with 6 new split columns namely, ‘Allocation’, ‘Type of Intervention model’, ‘Masking’, ‘Primary Purpose’, ‘Type of Observational model’, and ‘Time Perspective’. For Interventional studies, the columns ‘Type of Observational Model’ and ‘Time Perspective’ are filled with ‘None’ and vice-versa for Observational studies.

Figure 8 displays the count of different types of interventional models used in clinical trials that involve active intervention or treatment. This visualization helps better understanding of the diverse interventional clinical trial techniques that can be employed and their relative frequency within the clinical trial landscape.

Figure 9 presents the count of different types of observational models used in clinical trials that primarily involve observation and data collection without active intervention. This visualization helps better understanding of the different observational techniques that can be employed and their relative frequency within the clinical trial landscape.

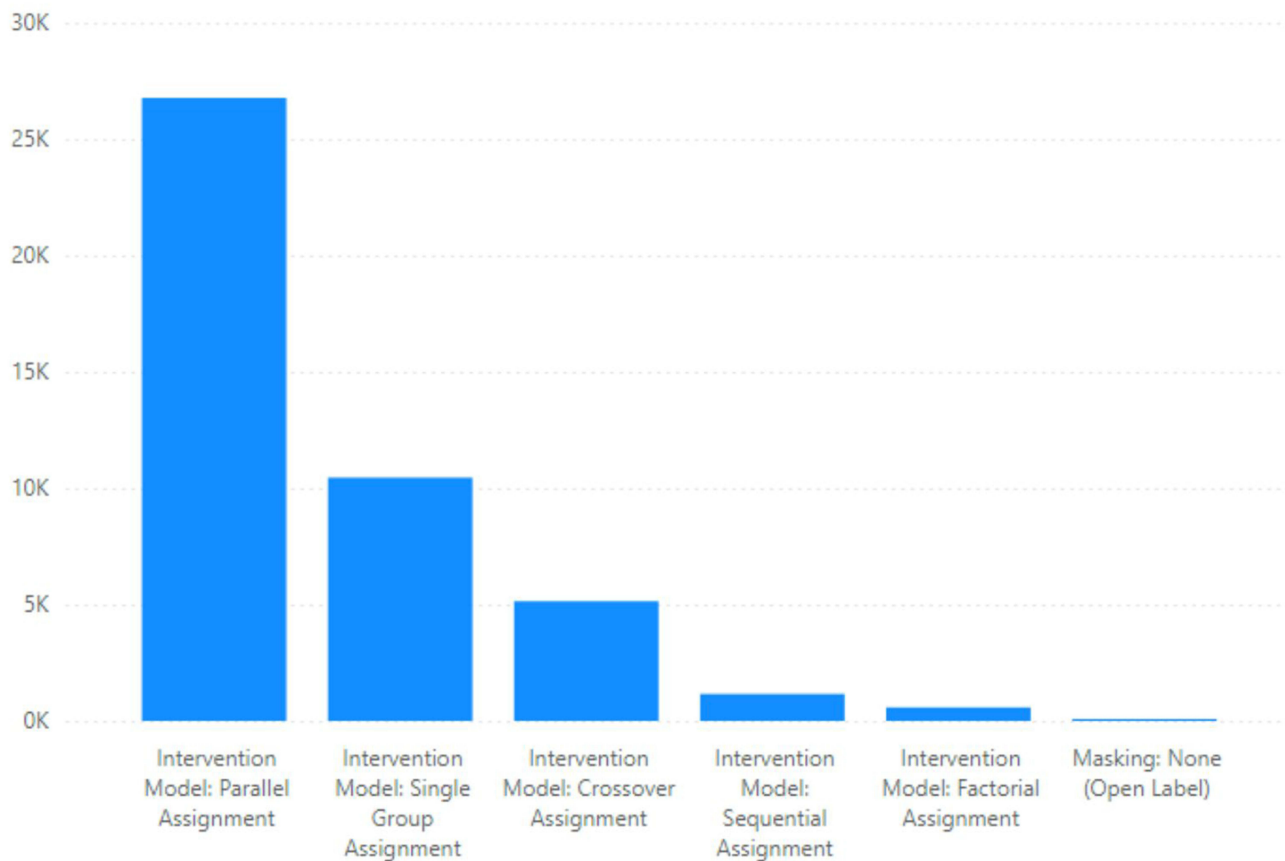


Figure 8 Count of different types of interventional models used in the clinical trials that are interventional in nature.

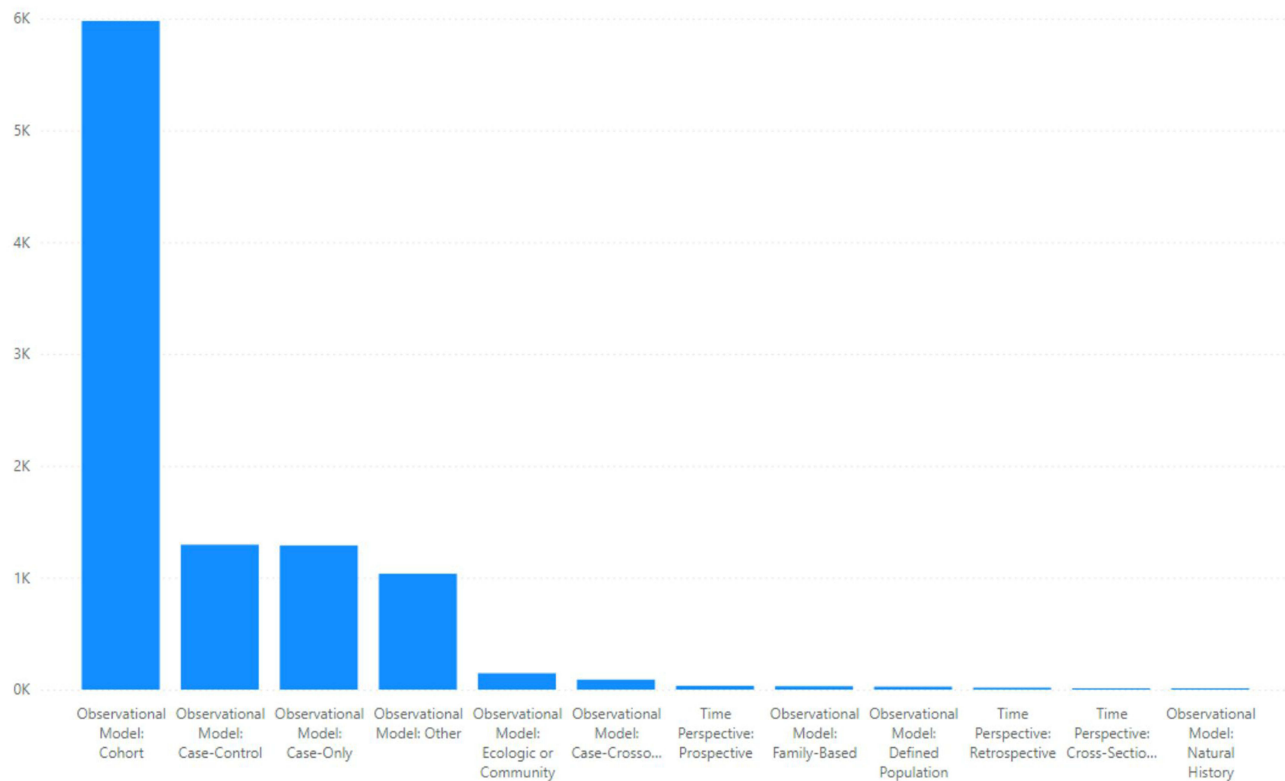


Figure 9 Count of different types of observational models used in the clinical trials that are observational in nature.

After appending relevant data in the additionally created columns, the next step is to identify textual columns, for categorization and label encoding. This process transforms the textual data into numerical representations, facilitating further computations and modelling.¹⁵ Certain other columns with textual data (of frequency information) are processed through the tf-idf vectorizer method. This method calculates the term frequency-inverse document frequency (tf-idf) scores for each word in the text, capturing the significance of words in the specific document relative to the entire dataset. This transformation enables the extraction of meaningful features from the textual data.¹⁶

These selective techniques on the textual columns help effective pre-processing and conversion of data into formats suitable for subsequent analysis and modelling.

Specific to the proposed research in this paper, columns such as “Status”, “Study Results”, “Gender”, “Phases”, “Study Type”, “Allocation”, “IModel”, “Masking”, Primary Purpose, “Omodel”, “Time Perspective”, and “Funded Bys” are categorised and label encoded. While, the remaining columns, namely “Title”, “Conditions”, “Interventions”, “Outcome Measures”, “Sponsor/Collaborators”, “Age”, “Other IDs”, and “Locations” are converted to numerical data using the tf-idf vectorizer.

The final step, before proceeding with model development, is feature selection.¹⁷ In machine learning it is very important to eliminate the redundant (and duplicate) columns/features, so as to avoid overfitting and unwanted bias. Pearson heatmap (a graphical representation of a correlation matrix using color-coded cells, where the color intensity indicates the strength and direction of the Pearson correlation coefficient. It helps visualize the relationships between variables, with dark colors indicating strong positive or negative correlations, and lighter colors representing weaker or no correlations) is used to check dependencies between the columns and to eliminate the columns with a correlation higher than 0.8.

All columns in the second dataset have numerical data and, hence, can be directly processed through feature selection.

The entire discussions above summarize the methods and procedures that were followed to complete data cleaning and to prepare the final datasets for model application.

Model

The usefulness of our datasets was tested using five classification models namely, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XGBoost, and Support Vector Classifier.

The first dataset which contains 55k rows is split into five small datasets each of which contains approximately, 10k-15k rows. The models are trained across these five datasets and tested. The training and test datasets fed into the models have been produced by applying `train_test_split` within those respective datasets. The performance of these models is evaluated using conventional metrics such as balanced accuracy score (BAS), receiver operator characteristic area under the curve (ROC-AUC) score, precision, and recall scores. Additionally, cross-validation technique is also used to compare and benchmark all the models across datasets; to help decide on which particular model is to be applied in practice.¹⁸

In addressing the class imbalance within the initial dataset, we implemented balanced class weights across all five machine learning algorithms. To ensure the preservation of crucial data points, we chose not to apply under-sampling or over-sampling techniques, nor did we generate synthetic data. In clinical datasets, every data point is significant, and under-sampling could result in the loss of valuable information. Similarly, over-sampling or generating synthetic data would likely fail to accurately replicate clinical data, as such data is inherently variable and unpredictable, differing from trial to trial. This approach helps avoid poor replication and mitigates the risk of overfitting, ensuring the reliability and integrity of the model.

The second part of the research focuses, i.e., the patient streamlining dataset is additionally processed through an Artificial Neural Network for a better test accuracy.

Figure 10 sums up how we have gone about using ML in the clinical trial design process in the form of a flowchart, a way that explains how the dataset is extracted from ClinicalTrials.gov, then enhanced and fed into the model, which is then evaluated using various metrics. We then compare the five models we use and determine the best one for the stakeholder to use for a new trial. Stakeholders can interact with the system by providing the parameters of their trial protocol. This includes critical features such as the drug's site of action and other trial design elements. Based on the input values, the model predicts whether the trial is likely to succeed or face termination. In cases where the model predicts failure, stakeholders are prompted to modify trial parameters and re-evaluate the protocol iteratively. This

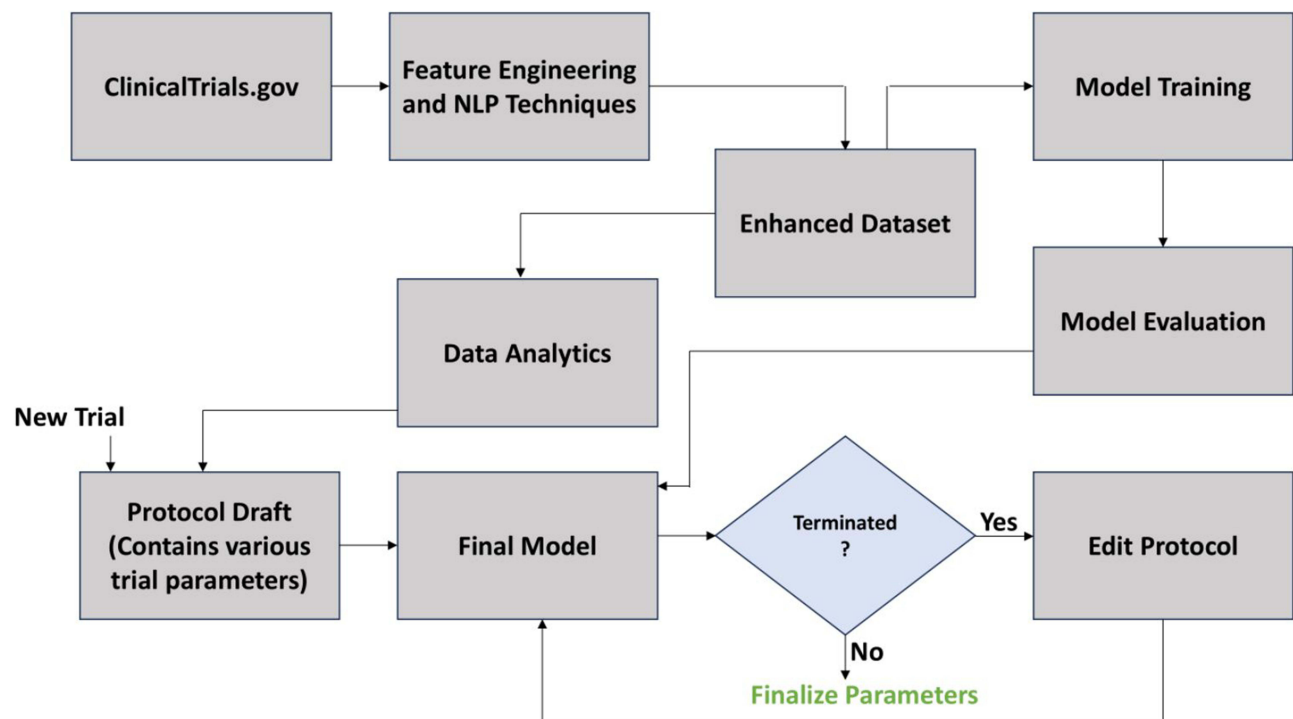


Figure 10 The proposed pipeline to ease the process of site selection.

iterative process continues until the model predicts a high probability of success; the protocol is finalized at this point, allowing the trial to progress to subsequent phases.

By integrating machine learning and advanced analytics, the pipeline minimizes the risk of trial failure by addressing potential issues at the design stage. This approach optimizes resource allocation, accelerates the drug development process, and ensures that trial decisions are informed by historical evidence and predictive insights. Ultimately, the pipeline enhances the efficiency and reliability of clinical trials, serving as a valuable tool for stakeholders to make data-driven decisions and improve overall trial success rates.

Figure 11 demonstrates the pipeline for patient/volunteer onboarding for a trial. It is designed to assist stakeholders, such as pharmaceutical companies and researchers, in ensuring efficient and accurate patient selection for every clinical trial phase, including Phase 1, Phase 2, and Phase 3.

The process begins by sourcing data from repositories such as the UCI Machine Learning repository or other relevant datasets specific to the trial (the liver patient dataset has been used in this proposal only because it was easy to procure, and it is only an example to demonstrate the working of the pipeline; for real-world use cases, the model is to be trained with the data relevant to a particular trial/requirement). Feature selection techniques are then applied to identify critical patient characteristics necessary for the trial. This creates an enhanced dataset, the foundation for training machine learning models tailored to the trial's requirements. The trained models are rigorously evaluated to ensure their reliability in predicting patient eligibility.

Once the system is set up, stakeholders can define eligibility criteria for each trial phase and input a potential volunteer's data into the model. The model predicts whether the patient is eligible based on the specified parameters, such as demographic details, medical history, or biomarker data.

This pipeline ensures a systematic and data-driven approach to patient onboarding, minimizing manual errors, improving recruitment efficiency, and optimizing resource utilization. By enabling stakeholders to train the model with additional datasets and tailor eligibility parameters for each trial phase, the framework enhances the overall effectiveness of patient selection, contributing to smoother trial progression and more reliable outcomes.

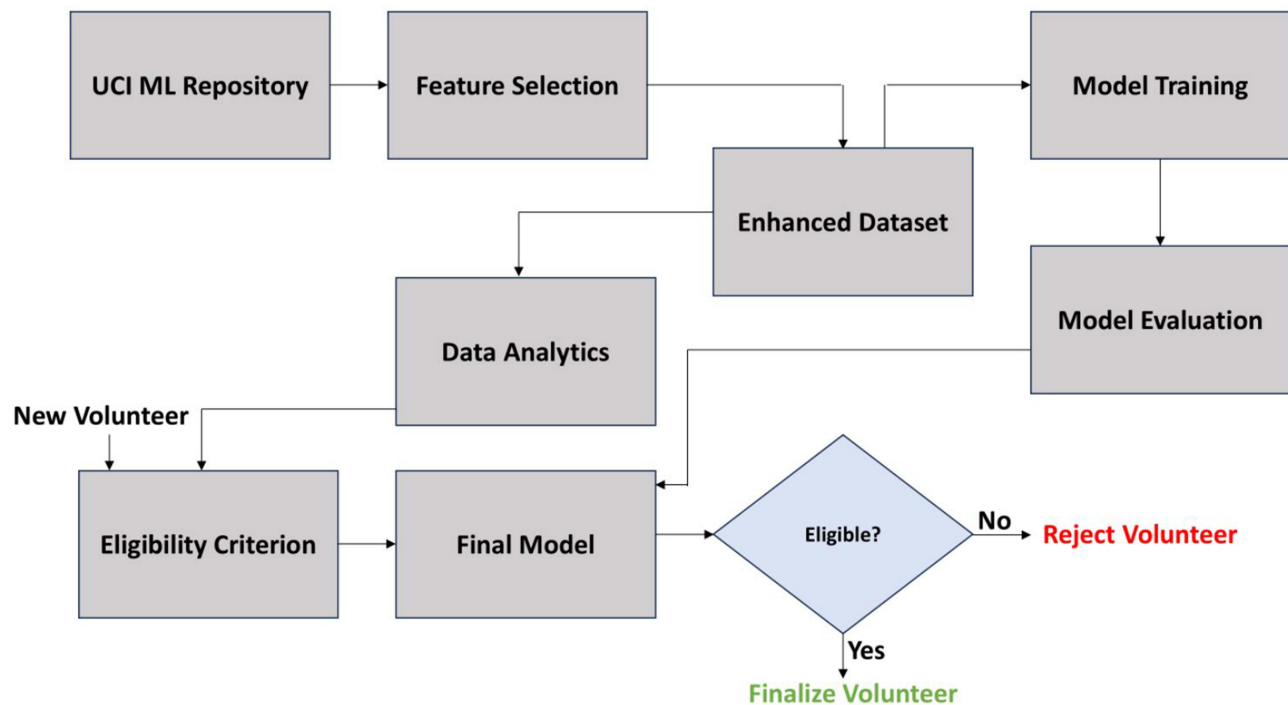


Figure 11 The proposed flow of how the mechanism will work to streamline patient onboarding.

Results and Conclusions

As discussed elaborately in the previous sections, the dataset from the first part of the study has 55k samples and is now divided into five small chunks, each with around 10k-15k rows. All the ML models are trained and tested over these five chunks of data.

Table 2 demonstrates the average precision and recall scores of the 5 models used for the first part of the research statement, ie, streamlining the trial design.

Table 3 illustrates the average balanced accuracy, average ROC-AUC score, and average weighted F1-Score of the five ML models across the five small datasets.

As observed in Tables 2 and 3, the best performance is achieved by XGBoost model across all the five datasets, with an average ROC-AUC score of 0.7 and an average balanced accuracy of 0.71. Random Forest Classifier also yields a comparable performance in all the enumerated metrics as observed in the tables.

Figures 12–16 show the ROC Curves for the five classification models used on the first dataset to determine the various clinical trial parameters.

The performance evaluation was conducted using k-fold cross-validation, a widely recognized statistical technique in machine learning and data analysis. This method involved partitioning the available data into multiple subsets or “folds”, allowing us to train the model on one portion and test it on the remaining folds. Notably, the results presented in Table 4 represent the average cross-validation scores derived from the analysis of the five models. Among these models, the scores of the Random Forest algorithm exhibited a slight superiority.

Therefore, it is summarized based on experiments, that XGBoost classifier and Random Forest Classifier are the two better performing ML models, effective with large clinical data.

The dataset pertaining to the patient streamlining process, which forms the focus of the second phase of our study, underwent classification using five models to determine the eligibility of patients for a specific clinical trial. This dataset was additionally subjected to an Artificial Neural Network (ANN)¹⁹ analysis. Table 5 gives an overview of the results for this run.

The above table shows that the two best performers on this dataset are Support Vector Classifiers and ANNs. However, given that Support Vector Classifiers come with a constraint of performing poorly on large datasets,²⁰ we

Table 2 Average Precision and Recall Scores of the 5 ML Models

ML Models	Average Precision	Average Recall
Logistic Regression	0.862677425	0.691606465
Decision Tree Classifier	0.864190546	0.851944697
Random Forest Classifier	0.873313749	0.884175042
XGBoost Classifier	0.876797193	0.827819031
Support Vector Classifier	0.881589464	0.15466244

Table 3 Comparison of the 5 Classification Models Based on the Metrics - Average Balanced Accuracy, Average Roc-Auc, and Average f1-Score

ML Model	Average Balanced Accuracy	Average ROC-AUC	Average F1-Score
Logistic Regression	0.579558473	0.58	0.657638605
Decision Tree Classifier	0.651378029	0.654	0.857355174
Random Forest Classifier	0.658882706	0.662	0.879491915
XGBoost Classifier	0.710962	0.7	0.847450809
Support Vector Classifier	0.511502742	0.512	0.092840777

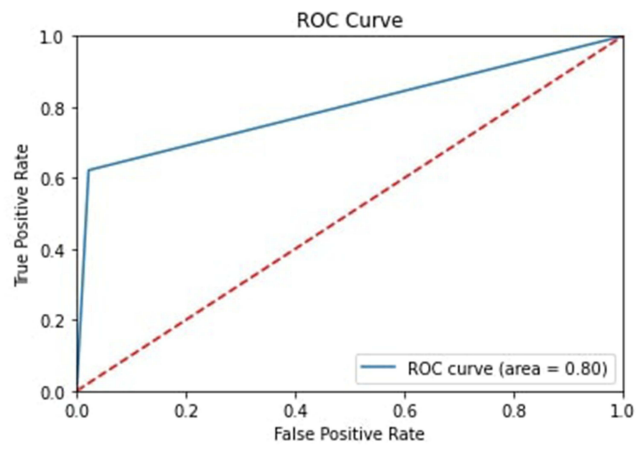


Figure 12 ROC Curve for Random Forest.

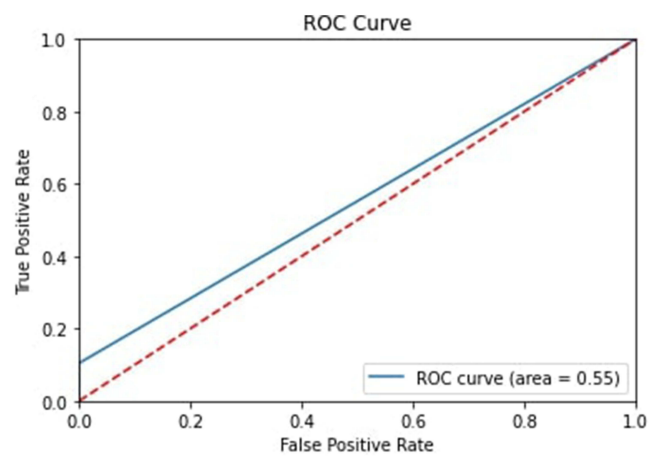


Figure 13 ROC Curve for Logistic Regression.

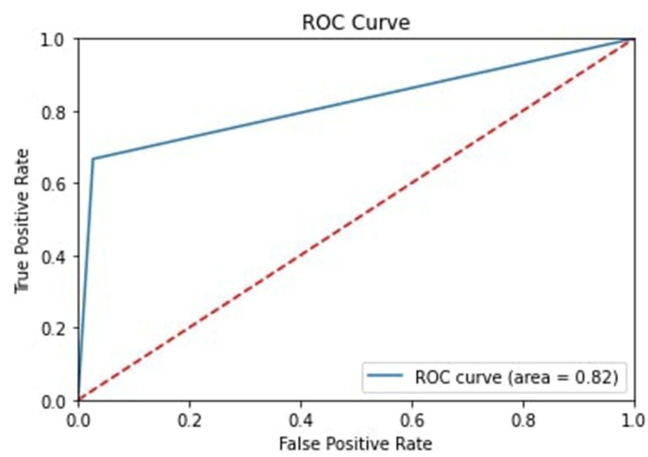


Figure 14 ROC Curve for Decision Tree.

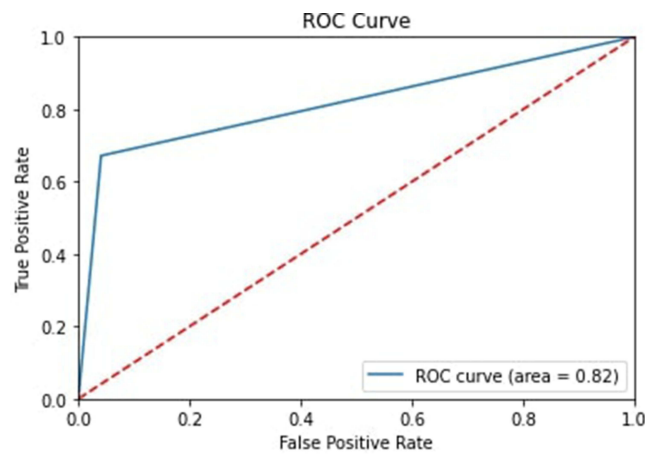


Figure 15 ROC Curve For XGBoost.

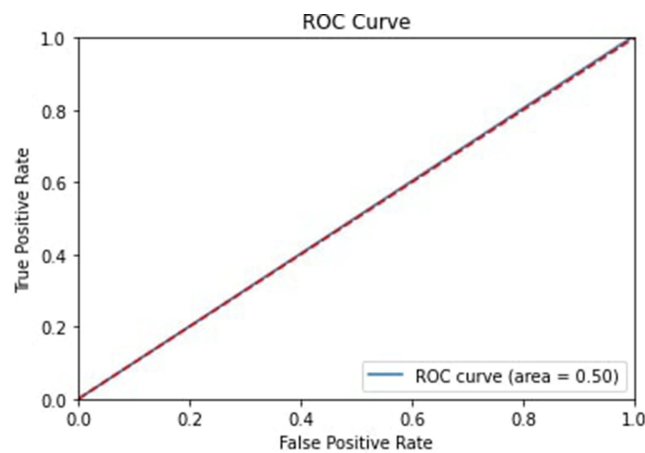


Figure 16 ROC Curve For Support Vector Machine.

would be going ahead with the ANN in our pipeline since, in the future, when this pipeline is to be implemented for real-world use cases, datasets would not necessarily be small like the one we have used in our research proposal, hence going ahead with ANNs makes more sense.

In conclusion, our research highlights the significance of machine learning (ML) techniques in two critical aspects of clinical trials: the selection of the site of action for a drug and the patient enrolment process. Through the analysis of a large dataset and the utilization of five ML models, the effectiveness of XGBoost and Random Forest Classifier in

Table 4 Comparison of the 5 Classification Models Based on Average Cross-Validation Scores

ML Model	Average Cross-Validation Score
Logistic Regression	0.592
Decision Tree Classifier	0.848
Random Forest Classifier	0.8848
XGBoost Classifier	0.83
Support Vector Classifier	0.148

Table 5 Results On The Second Dataset

Models Run on the Dataset	Test Accuracy
Logistic Regression	0.703
Decision Tree Classifier	0.6457
Random Forest Classifier	0.726
XGBoost Classifier	0.709
Support Vector Classifier	0.737
ANN	0.737143

accurately identifying patient eligibility for specific clinical trials is demonstrated. Empirically, these models outperformed other methods, exhibiting reliable results and robust performance, when applied to large clinical datasets.

An Artificial Neural Network (ANN) was employed, specifically tailored to handle patient data, which was used by us to demonstrate the use of ML in the patient streamlining process. The ANN analysis yielded promising outcomes, achieving a test accuracy of 0.73714, emphasizing optimization of patient streamlining process; using ANN, specific to our model.

The results of our study emphasize the importance of leveraging ML techniques in facilitating the selection of suitable drug action sites and streamlining patient enrolment in clinical trials. These methodologies offer valuable insights, improve decision-making processes, and ultimately contribute to both expediting clinical trial termination and achieving accuracy of clinical research results. As the field of machine learning continues to advance, its integration into clinical trial processes holds tremendous promise for improving patient care and driving medical innovation.

Limitations and Future Work

The research proposed in this paper has one potential limitation; only a limited number of features or design properties are used in clinical trials, which possibly restricts the depth of insights gained by our machine learning models. Including more comprehensive information, such as detailed patient conditions, could possibly enhance the models' ability to learn and achieve more accurate predictions.

Future research can focus on developing algorithms with improved interpretability, leveraging feature engineering and natural language processing techniques to extract valuable insights from textual descriptions of trials. In the context of streamlining patient enrolment, incorporating additional features, such as considering the mental stress levels of patients, could potentially lead to improved accrual rates.²¹ One additional improvement that can be done to this proposed pipeline is hyperparameter tuning of specific ensemble methods in order to check if hyperparameter tuning has some effect on such healthcare datasets.

Glossary

1. Streamlining Patient/Volunteer Onboarding - The process of efficiently selecting and enrolling suitable participants for clinical trials by assessing their eligibility based on predefined criteria, often using advanced algorithms or tools to optimize recruitment.

2. Accrual rates - The speed or rate at which participants are recruited and enrolled in a clinical trial, critical for ensuring the trial progresses on schedule.

3. Selection of Site of Action - The identification of the specific biological location where a drug or treatment exerts its intended therapeutic effect, a key factor in ensuring its efficacy and safety.

4. Clinical Trial Design Parameters - The structured features of the first dataset, such as the study type, duration, type of study, etc, which guide the trial's execution and validity.

5. Pharmacokinetics - The study of how a drug is absorbed, distributed, metabolized, and excreted by the body, focusing on the drug's behavior over time and its interactions within various tissues and organs.

Data Availability

As mentioned in the section 6.1 titled “Data”, our datasets on clinical trials are drawn from clinicaltrials.gov; data on patients is taken from the UCI ML repository (<http://archive.ics.uci.edu/>). Both repositories are public datasets, available for free access. These two links are also referenced in the corresponding section of the paper.

Disclosure

The research presented in this paper was conducted without external funding. Data utilized in this study were sourced from publicly available datasets, including ClinicalTrials.gov and the UCI Machine Learning Repository. The authors wish to disclose that they have no conflicts of interest to report regarding this research.

References

1. Sertkaya A, Wong -H-H, Jessup A, Beleche T. Key cost drivers of pharmaceutical clinical trials in the United States. *Clin Trial*. 2016;13:117–126. doi:10.1177/1740774515625964
2. Desai M. Recruitment and retention of participants in clinical studies: critical issues and challenges. *Perspect Clin Res*. 2020;11:51.
3. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. *Lancet*. 2002;359:57–61.
4. Rizk M, Zou L, Savic R, Dooley K. Importance of drug pharmacokinetics at the site of action. *Clin Transl Sci* 2017;10(133):133–142. doi:10.1111/cts.12448
5. Carlisle B, Kimmelman J, Ramsay T, MacKinnon N. Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials. *Clin Trial*. 2015;12:77–83. doi:10.1177/1740774514558307
6. Clark GT, Mulligan R. Fifteen common mistakes encountered in clinical research. *J Prosthodont Res*. 2011;55:1–6. doi:10.1016/j.jpor.2010.09.002
7. Nagra NS, Bleys J, Champagne D, Devereson A, Macak M. Understanding the company landscape in AI-driven biopharma R&D. *Biopharma Dealmakers*. 2023. doi:10.1038/d43747-023-00020-4
8. Elkin ME, Zhu X. Predictive modeling of clinical trial terminations using feature engineering and embedding learning. *Sci Rep*. 2021;11(3446). doi:10.1038/s41598-021-82840-x
9. Follett L, Geletta S, Laugerman M. Quantifying risk associated with clinical trial termination: a text mining approach. *Inform Process Manag*. 2019;56:516–525. doi:10.1016/j.ipm.2018.11.009
10. Kavalci E, Hartshorn A. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. *Sci Rep*. 2023;13:121.
11. Medicine USNLo. *patien*. In book *clinicalTrials.gov* (Editor ed.^eds.). City.
12. Bendi Ramana NV. ILPD (Indian liver patient dataset). *UCI Machine Learning Repository*. 2012.
13. Urte Fultinavičiuote IM. Trial termination analysis unveils a silver lining for patient recruitment. *Clinical Trials Arena*. 2022.
14. Jain A, Patel H, Nagalapatti L, et al.: Overview and importance of data quality for machine learning tasks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 3561–3562.
15. Gupta H, Asha V. Impact of encoding of high cardinality categorical data to solve prediction problems. *J Comput Theor Nanosci*. 2020;17:4197–4201. doi:10.1166/jctn.2020.9044
16. C-z L, Y-x S, Z-q W, Yang Y-Q: Research of text classification based on improved TF-IDF algorithm. In: 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE). IEEE; 2018: 218–222.
17. Janecek A, Gansterer W, Demel M, Ecker G. On the relationship between feature selection and classification accuracy. In: New challenges for feature selection in data mining and knowledge discovery. PMLR;2008:90–105.
18. Schaffer C. Selecting a classification method by cross-validation. *Machine Learning*. 1993;13:135–143. doi:10.1007/BF00993106
19. Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Crit Care Med*. 2001;29:291–296. doi:10.1097/00003246-200102000-00012
20. Awad M, et al. Support vector machines for classification. *Eff Learn Mach*. 2015;2015:39–66.
21. Andersen BL, Farrar WB, Golden-Kreutz DM. Psychological, behavioral, and immune changes after a psychological intervention: a clinical trial. *J clin oncol*. 2004;22(17):3570. doi:10.1200/JCO.2004.06.030

ClinicoEconomics and Outcomes Research

Publish your work in this journal

ClinicoEconomics and Outcomes Research is an international, peer-reviewed open-access journal focusing on Health Technology Assessment, Pharmacoeconomics and Outcomes Research in the areas of diagnosis, medical devices, and clinical, surgical and pharmacological intervention. The economic impact of health policy and health systems organization also constitute important areas of coverage. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/clinicoeconomics-and-outcomes-research-journal>

Dovepress
Taylor & Francis Group