Journal of Posthumanism
2025
Volume: 5, No: 8, pp. 257–271
ISSN: 2634-3576 (Print) | ISSN 2634-3584 (Online)
posthumanism.co.uk

DOI: https://doi.org/10.63332/joph.v5i8.3157

Integrating Artificial Intelligence and Data Science for Breakthroughs in Drug Development and Genetic Biomarker Discovery

Md Habibur Rahman[1], Md Abubokor Siam[2], Ahmed Shan-A-Alahi[3], Kazi Bushra Siddiqa[4], Shuchona Malek Orthi[5], Md Kazi Tuhin[6], Emran Hossain[7], Mukther Uddin[8]

Abstract

The evolving complexity of drug discovery and the demand for focused therapeutics have given further impetus to the implementation of AI and data science-based approaches. Such methods can analyze genomic and pharmacological data more rapidly, facilitating the precision design of drugs and genetic biomarkers of such drugs. This work is a machine learning framework that seeks to predict drug responses and find genetic biomarkers in the Genomics of Drug Sensitivity in Cancer (GDSC) dataset. Numerous data cleansing, normalization and one-hot encoding were also done to maintain credibility in the analysis. The robustness of a Random Forest classifier on the processing of high-dimensional biological data and excellent predictive performance was also demonstrated, with 97.7% accuracy, 98.4% precision, recall, and F1-score. The comparative studies provided better results than other models like SVM 95%, BiLSTM 80%, and GATv2 77.9%. The discriminative power of the model was proved using ROC and precision-recall curves. The framework of AI + data science in pharmacogenomics can be used to identify patterns of drug sensitivity efficiently, and, thus, promote personalized medicine and biomarker-based treatments. The method is a scalable, interpretable, and time-saving alternative to the traditional pipelines of drug discovery.

Keywords: *Artificial Intelligence, Drug Development, Genetic Biomarkers, Biomarker, Data Science, Genomics of Drug Sensitivity in Cancer (GDSC).*

Introduction

The increasing demand in personalized and targeted treatment redefined the face of modern medicine that forcing reconsidering the overall drug discovery, development, and delivery process. At the center of this revolution lies the study of pharmacology that seeks to understand the effects of drugs on biological systems by studying the interaction between chemical agents

[1] School of Business, International American University, Los Angeles, CA 90010, USA, Email: rahman@aub.ac.bd, ORCID ID: https://orcid.org/0009-0009-3830-9285.
[2] College of Business, Westcliff University, Irvine, CA 92614, USA, Email: m.siam.263@westcliff.edu, ORCID ID: https://orcid.org/0009-0001-5250-4652.
[3] Department of Technology and Computer Science, University of The Potomac, Washington DC, USA
Email: ahmed.shanaalahi@student.potomac.edu, ORCID id: https://orcid.org/0009-0007-6079-4999
[4] School of Business, International American University, Los Angeles, CA 90010, USA, Email: bushrasiddiqa82@gmail.com, ORCID ID: https://orcid.org/0009-0008-0283-9850
[5] College of Business, Westcliff University, Irvine, CA 92614, USA, Email: s.orthi.339@westcliff.edu, ORCID ID: https://orcid.org/0009-0007-5397-4561
[6] Katz School of Science and Health, Yeshiva University, 245 Lexington Avenue, New York, USA, Email: tuhinmd358@gmail.com, Orcid ID: https://orcid.org/0009-0002-6914-6182
[7] Department of Business Administration, Humphreys University, Stockton, California, USA, Email-hu0112358@student.humphreys.edu, ORCHID ID: https://orcid.org/0009-0005-2080-780X
[8] School of Business, International American University, Los Angeles, CA 90010, USA, Email: muktheruddin@ieee.org, (Corresponding Author), ORCID ID: https://orcid.org/0009-0009-0995-8761

and biological systems [1]. In this field, drug discovery is a significant endeavor that has historically involved a multi-step process that includes target selection, high-throughput screening, lead optimization, and experimental confirmation. Although they are fundamental, such techniques can be time-consuming, resource-consuming, and small in their capacity to support the molecular heterogeneity of patients and diseases.

The recent development in genomics and molecular biology has presented the genetic biomarker as a critical factor in augmenting the drug development strategy [2]. These biomarkers like gene mutations, their expression, and structural variants are measurable factors that aid in forecasting of disease progression, response to therapy as well as aspects of vulnerability of an individual patient. And, their entry into the drug development pipeline is aligned to the paradigm of precision medicine that seeks to get the right drug, to the right patient at the right time. Stratified clinical decision-making, reduction of adverse effects, and the overall enhancement of effective treatment are facilitated through the utilization of biomarkers [3]. Nonetheless, the operationalization of precision medicine is associated with severe complications that emerge because of the exponential expansion of biological data collected using high-throughput sequencing technologies and multi-omics profiling. Genomic, transcriptomic, proteomic, and metabolomic data are generally big, noisy, and very high-dimensional but are also very information-rich. The task of combining all these different types of data with pharmacological responses is a complicated issue far beyond the reach of classical statistical procedures [4]. Scalability, interpretability, and difficulty with non-linear relationships in biomedical data are common problems of traditional models. To overcome these limitations, artificial intelligence (AI) [5],

Data science, machine learning (ML), and data intelligence have been introduced as potent means of computing. Such technologies provide powerful data mining abilities to learn and model complex, high-volume data, to learn hidden patterns and come up with predictive models without explicit rule-based programming [6][7][8]. AI has already proven useful in drug discovery in diverse areas such as drug-target interaction prediction, virtual screening, toxicity prediction, and the identification of a biomarker. One of the best machine learning algorithms that is commonly used is the ensemble algorithms such as Random Forests that have good robustness, scalability, and interpretability and hence fit biomedical tools that deal with heterogeneous data.

Motivation and Contributions of the Study

The emergence of more complex diseases like cancer, and the need to produce more personalized medicine have led to a paradigm shift in the drug development and the identification of biomarkers. The conventional experimentation practices, although useful, tend to be time-consuming, costly, and limited in their scope. Recent developments in the field of AI and data science can provide a transformative opportunity to overcome these issues and speed up the discovery of both therapeutic targets and genetic biomarkers. The integration of AI into this domain not only enhances the efficiency of drug screening and response prediction but also propels the advancement of precision medicine. The combination of AI and the data-driven approach is promising to transform the drug discovery process by making it possible to apply therapeutics to the patient according to his or her genetic makeup and eventually, increasing treatment efficacy. The key contribution of the study as follows:

Utilization of the Genomics of Drug Sensitivity in Cancer (GDSC) dataset, has been used, which possesses enriched genomic and pharmacological data, and is used to assist in making

predictions on drug sensitivity.

Proper use of data preprocessing methods such as data cleaning, one-hot encoding of categorical variables, and normalization in order to equate the values of the input variables.

Division of the preprocessed dataset into subsets for testing and training in order to enable objective performance assessment and reliable model training.

Implementation of a Random Forest classification model to predict drug response based on genomic features and gene expression profiles.

The model's predictive capacity and possible clinical significance are validated by performance evaluation utilizing measures such as accuracy, precision, recall, and F1-score.

Justification and Novelty

This study is justified by the growing need to accelerate drug development and genetic biomarker discovery through advanced, data-driven approaches. Traditional methods are often slow, costly, and limited in handling complex genomic data. The proposed framework introduces a machine learning-based pipeline utilizing the GDSC dataset and a Random Forest model to predict drug sensitivity. The novelty lies in integrating AI techniques with genomic data preprocessing to uncover hidden patterns in drug response. This approach enhances predictive accuracy and supports the development of targeted therapies, contributing to more efficient and Personalized therapeutic approaches in the rapidly evolving field of precision medicine.

Organization of the Paper

The paper is organized as follows: Section II reviews related work. Section III describes the proposed methodology and evaluation metrics. Section IV presents experimental results and comparative analysis. Section V concludes the study and outlines future research directions for advancing AI-driven drug development and genetic biomarker discovery.

Literature Review

This section reviews recent advances in artificial intelligence and data science applied to drug development and genetic biomarker discovery, highlighting machine learning and deep learning techniques that accelerate drug discovery, identify key biomarkers, and enhance precision medicine through improved prediction, molecular targeting, and data-driven insights. Recently selected studies were reviewed:

Ahmad et al. (2025) highlighted the importance of genomics in early diagnosis and drug development strategies. ML algorithms like RF, GB, Deep Belief Networks, Autoencoders, SVM, CNN, and RNN are used extensively in modern genomics. Reinforcement Learning, DNN, GANs, and GNNs are used for optimized drug discovery. However, ML algorithms face data scarcity and interpretability issues, challenging accuracy and integration with experimental validation. Lung cancer therapeutics are experiencing rapid advancements with remarkable accuracy, often exceeding 95% in specific applications. More optimization is needed to efficiently integrate AI for clinical validation [9].

Mehta et al. (2025) explored the use of machine learning (ML) in understanding autoimmune diseases, which involve multiple genes and gene-by-context interactions. Traditional biochemical-genetic approaches struggle to provide answers due to the high dimensionality and epistatic interactions between genes. ML algorithms, such as the XGBoost classifier, have been

shown to help in understanding these diseases by processing big genomic data. The optimizer achieved 94.75% accuracy, revealing that other genes, such as WBC count, are compatible with inherent genes influencing autoimmune disease risk. The study emphasizes ML's potential for finding biomarkers and figuring out the genetics of autoimmune disorders, providing future directions for developing psychological therapies targeted at genetic vulnerabilities in autoimmune disease management [10].

Kalyani et al. (2024) investigated how AI-driven predictive modelling can revolutionize personalized medicine and speed up drug discovery. Through the utilization of sophisticated ML algorithms and vast biomedical datasets, the research endeavors to expedite the identification of auspicious medication candidates and customize therapies to specific patient profiles. This research holds the potential to completely transform the pharmaceutical sector by cutting down on the duration and expenses associated with medication development and enhancing patient outcomes via more focused treatments. AI technologies enable a significant improvement in the accuracy and effectiveness of medical treatments; the model shows a 20% increase in accuracy over conventional approaches [11].

Zhou et al. (2024) investigated the effectiveness of cancer drugs in mouse models, with a particular emphasis on growth kinetic models. The tumor volume data from over 30,000 mice in 930 trials were used to evaluate the semiparametric generalized additive model (GAM) and six parametric models. The researchers discovered that, when compared to other models such as von Bertalanffy and Gompertz, the exponential quadratic model was the most successful parametric model, including 87% of the studies. Because 7.5% of the growth data at the mouse group level could not be fitted by any model, GAM was employed. Both the exponential and exponential quadratic models were equally accurate in identifying the study's biomarkers and pharmacological mechanisms [12].

Spooner et al. (2023) investigated feature selection ensembles to improve high-dimensional dataset stability. They evaluated their anticipatory accuracy and strength by employing data-driven upper bounds to choose relevant features with an ensemble feature chooser. Results stabilities were shown to be up to 34% higher with an ensemble feature selector with data-driven thresholds compared with a single feature selector. Both the threshold algorithm threshold and the robust rank aggregation threshold, which were related to the information retrieval domain, were the best-performing data-driven thresholds. The approach does not compromise speed and provides more reliable and repeatable feature selections. [13].

Dwivedi et al. (2023) proposed a novel AI-based DL approach for identifying the biomarkers of NSCLC subtypes. It has a feed-forward neural network, an autoencoder, and a biomarker finding component. It was determined that the biomarkers were relevant in classifying the NSCLC subtypes. On the basis of these biomarkers, many ML models were created, and the accuracy of the Multilayer Perceptron was 95.74%. Of the 52 biomarkers, 45 have been previously documented in the literature, while the remaining 7 have not. These biomarkers are utilized to subtype non-small cell lung cancer (NSCLC) and may thus be examined to determine their potential contribution to targeted lung cancer treatment [14].

Table I collects and summarizes previous works and published research in the field of drug development and biomarker discovery, summarizing their methodology, data used, main findings, limitations, and future research, with special emphasis on predictive modelling, biomarker discovery, and the combination of artificial intelligence and data science methodologies

Author's	Methodology	Data	Key Findings	Limitation / Future Work
Ahmad et al. (2025)	ML algorithms (RF, GB, DBN, AE, SVM, CNN, RNN, RL, DNN, GANs, GNNs)	Genomics datasets from public studies	ML aids in lung cancer diagnosis and drug design; accuracy >95% in some tasks	Challenges: data scarcity, heterogeneity, interpretability; needs better clinical validation
Mehta et al. (2025)	XGBoost classifier, feature selection, data imputation, hyperparameter optimization	Genetic datasets of autoimmune patients	Achieved 94.75% accuracy; identified key features (e.g., WBC count); pathway insights	Lacks integration with experimental/clinical validation; limited dataset diversity
Kalyani et al. (2024)	Predictive modelling using AI	Biomedical datasets for drug-target interaction	AI boosts DTI prediction accuracy and reduces discovery time; 20% improvement over traditional models	Requires further validation in clinical and regulatory environments
Zhou et al. (2024)	Parametric (exponential, Gompertz, von Bertalanffy, etc.) and semiparametric GAM modeling	Tumor volume data from 30,000+ mice across 930 oncology experiments	Exponential quadratic model best fits 87% of studies; GAM handles nonparametric cases; validated for efficacy analysis and biomarker discovery	Lacks integration of AI-based nonlinear models; future work may explore deep learning for more complex growth dynamics
Spooner et al. (2023)	Ensemble feature selection using thresholding based on data	Alzheimer's disease datasets	Data-driven thresholds improve feature selection stability by up to 34%; selected features align with current AD literature; more reproducible without loss of predictive performance	Needs testing on other diseases; integration with clinical validation for early biomarker discovery

Dwivedi et al. (2023)	Explainable AI-based deep learning (autoencoder + feed-forward neural network + XAI)	NSCLC molecular data (subtypes LUAD & LUSC)	Discovered 52 biomarkers for NSCLC subtype classification with 95.74% accuracy; 14 druggable biomarkers; 7 novel biomarkers identified for further exploration	Requires clinical validation of novel biomarkers; expand to larger, diverse cohorts

Table:1 Summary of Related Works on AI And Data Science Applications in Drug Development and Biomarker Discovery

Methodology

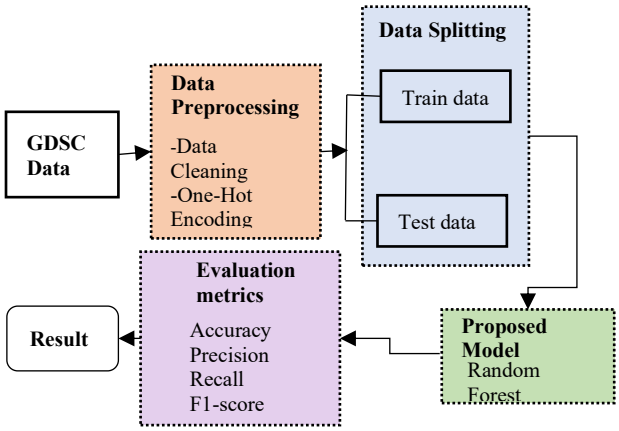

Fig:1 Proposed AI-Driven Framework for Drug Development and Genetic Biomarker Discovery

The use of an AI-based framework has been established to advance and fast-track drug development and genetic biomarker identification, utilizing the Genomics of Drug Sensitivity in Cancer (GDSC) dataset as demonstrated in Figure 1, which offers high-resolution molecular and drug response profiles across numerous cancer cell lines. The procedure begins with the GDSC data that undergoes a data preprocessing treatment consisting of the cleaning, one-hot encoding, and normalization stages to guarantee consistency and analysis preparedness. The purged data will be divided into training data and test data so as to achieve a sensible development of the model. RF classifier, due to its robustness and interpretability, is used to determine complex patterns and relationships between genomic features and drug sensitivity. Accuracy, precision, recall, and F1-score are used to measure the performance of the model, making it possible to confirm its prediction skills. The knowledge acquired in this way allows identifying the crucial genetic biomarkers and leads to an improvement in the prediction of drug sensitivity, contributing to the future development of personalized medicine and the use of targeted cancer treatment.

Data Collection

The Genomics of Drug Sensitivity in Cancer (GDSC) dataset includes information on more than 1,000 cancer cell lines that correspond to more than 30 different forms of cancer. It includes profiles of genome mutation, copy number variations as well as a gene expression level. Pharmacological screening encompasses sensitivity to approximately 300 anticancer agents assayed by parameters of known drug concentrations such as dose-response activity such as IC50 and AUC values. The visualizations of the data are given below:

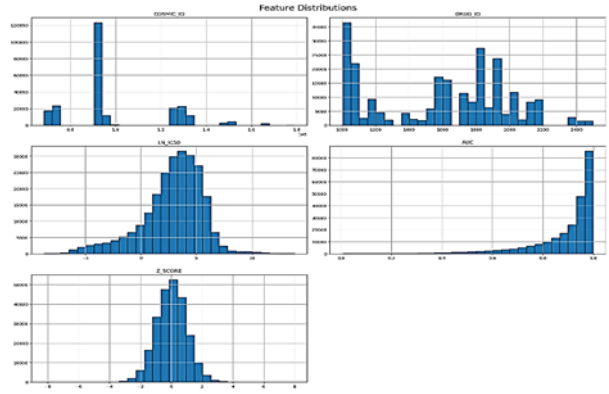

Fig:2 Distribution of GDSC Dataset Features

This plot shown in Figure 2 the distribution of important features in the GDSC dataset, including COSMIC_ID, DRUG_ID, LN_IC50, AUC, and Z_SCORE. LN_IC50 and Z_SCORE exhibit near-normal distributions, while AUC is highly right-skewed, indicating a concentration of values toward the lower end. COSMIC_ID and DRUG_ID show distinct, non-continuous spikes, representing their categorical nature as unique identifiers for cell lines and drugs.

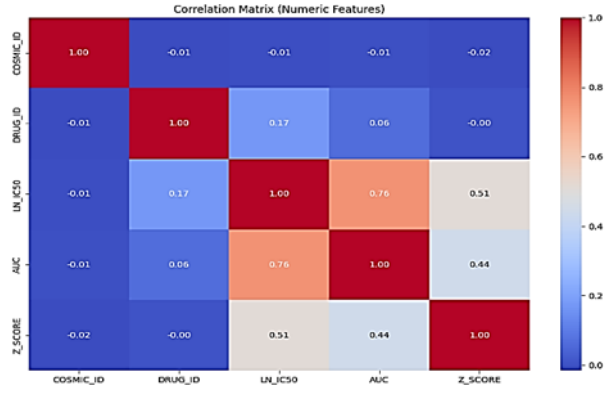

Fig:3 Correlation Heatmap of Numeric Features in GDSC Dataset

The correlation heatmap illustrates relationships among numeric features in the GDSC dataset in Figure 3. A strong positive correlation exists between AUC and LN_IC50 (0.76), suggesting similar patterns in drug response. Z_SCORE also shows moderate correlation with both LN_IC50 and AUC. COSMIC_ID and DRUG_ID demonstrate near-zero correlations, confirming their categorical nature and lack of direct linear relationship with the continuous features.

Data Preprocessing

In order to design and implement ML models, data preparation is essential. It guarantees the input data's accuracy and consistency, and suitably structured, enabling the model to identify meaningful patterns and make reliable predictions on unseen data. The following steps were performed during the preprocessing phase:

Data Cleaning

Effective data cleaning is essential to ensure the reliability and accuracy of downstream analysis, particularly in biomedical datasets that often contain missing values, noise, and outliers. The data preprocessing steps involved in this study are as follows:

Removed missing and outlier data from gene expression and mutation datasets using statistical thresholds and quality control filters.

Imputed missing gene expression and mutation values and domain-specific methods to ensure data completeness and consistency.

One-Hot Encoding

The preprocessing technique of one-hot encoding involves normalizing categorical data to numerical form suitable for ML models. It makes a binary column to all different categories where 1 is added to the corresponding category and 0 to the rest. This does not make any ordinal assumptions about the data. Although one-hot encoding does not transform the categorical data into other types, it may also result in high dimensionality in case there are a lot of different categories. Notwithstanding this, it is an essential approach to data preprocessing used in classification and regression.

Data Normalization

Normalization is a technique used to rescale data from its original range to a new, defined range. It reduces variations within the dataset, making the values more consistent and comparable, and ensuring they behave in a more uniform manner [15]. The range of independent variables or data features can be rescaled using the feature scaling approach known as min-max normalization. This method transforms features to a fixed range, typically [0, 1] or [-1, 1], preserving the relationships among the original data values. The transformation is defined by the following expressed in Equation 1:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

Where X is the initial value of the data, and X_{min} and X_{max} He feature's minimum and maximum values, respectively.

Data Splitting

The dataset was divided into two segments: 30% was set aside for testing in order to assess the model, and 70% was used for training the model in order to discover underlying patterns. The usefulness and generalize of the model to previously unseen data.

Proposed Random Forest Model

As an appropriate tool, the Random Forest method has already developed into a common data analysis tool for high-throughput data. Since the RF approach is straightforward, interpretable,

and adaptable to a wide range of predictor factors, it is helpful in molecular biology research. RF's ability to accurately determine the role of each variable used in response prediction is one of its most important characteristics [16]. Additionally, it performs exceptionally well even in situations with more than two classes, when the majority of the predictive variables contain noise, and when there are significantly more variables than observations.

The random forest's output is the one that the majority of trees choose for classification tasks. Decision trees are typically outperformed by them. The characteristic that has to be separated is selected in order to construct a decision tree. A split measure is used to determine the best split [17]. As a measure of a node's impurity, Equation (2) shows how to calculate the Gini index, one such split measure. If a value-based split were to be performed, how distributed would each split dataset be:

$$gini\ index = 1 - \sum_{i=1}^{n}(P_i)^2 \qquad (2)$$

This metric for measuring the computational efficiency of a node's impurity is significantly higher than that of entropy. The split results in a subset of the dataset where each split node has one fewer feature and fewer columns. This procedure is carried out repeatedly to create a tree and then a forest. A decision tree with two nodes is always produced. To determine the Gini significance, use Equation (3):

$$GI_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \quad (3)$$

Where C_j is the node's impurity value, GI_j is its significance, w_j is the weighted number of samples that reach node j, $left(j)$ is the child node from the left split on node j, and $right(j)$ is the child node from the right split on node j. Each feature vector's significance is determined using the formula in Equation (4).

$$f_i = \frac{\sum_j GI_j}{\sum_{k \in all\_nodes} GI_j} \qquad (4)$$

where f_iHat is the importance of feature i, GI_j is the significance of node j. The tree is constructed based on this metric, after which the subsequent set of input vectors (bag) is chosen and the procedure is repeated. A group of created trees makes up the final forest.

Performance Metrics

Performance indicators are crucial for assessing predictive models' reliability and accuracy, particularly in classification tasks. These metrics shed light on a model's ability to differentiate between several classes. Performance indicators are crucial for assessing predictive models' reliability and accuracy, particularly in classification tasks. These metrics shed light on a model's ability to differentiate between several classes. Some of the key terms used in classification performance evaluation are: These include the number of false positives (FP), true negatives (TN), false negatives (FN), and true positives (TP). The efficacy of the models was assessed using the performance measures listed below:

Accuracy

Accuracy, which is the proportion of properly detected labels in the whole population, is another indicator of the overall robustness of the model. [18]. It can be presented in Equation (5)

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \qquad (5)$$

Precision

A predicted true label's likelihood of being true is known as the model's precision, or positive predictive value. It is defined as follows Formula (6) was used to express the accuracy:

$$Precision = \frac{TP}{TP+FP} \qquad (6)$$

Recall

Recall, also known as Sensitivity or the TPR, is defined by Equation (7) and may be thought of as the proportion of true class labels that the model correctly detected as true.

$$Recall(Rc) = \frac{TP}{TP+FN} \qquad (7)$$

F1 score

The significance of TP and TN is taken into account by the F1-score or F-measure. Equation (8) indicates that the F1-score is just the accuracy and recall harmonic means:

$$F1\ score(F1) = 2 \times \frac{Precision \times Recall}{Precision+Recall} \qquad (8)$$

To determine the reliability of the model, its predictions on the test dataset were analyzed using various performance metrics.

RESULT ANALYSIS AND DISCUSSION

This study evaluates the effectiveness of a Random Forest-based predictive analytics framework for advancing drug development and identifying genetic biomarkers. The experiments were conducted using Python 3.9 and scikit-learn, with training executed on a high-performance computing environment equipped with an NVIDIA Tesla (16 GB VRAM) to efficiently handle large-scale genomic and biomedical data. As detailed in Table II, the Random Forest model demonstrated excellent performance across key evaluation metrics, achieving an accuracy of 97.7%, a precision, recall and F1-score of 98.4%. The combined results demonstrate that the Random Forest model can be an excellent tool in finding meaningful patterns in the biomedical data and providing accurate and reliable detection of the possible genetic biomarkers.

Evaluation Metrics	**Random Forest**
Accuracy	97.7
Precision	98.4
Recall	98.4
F1-score	98.4

Table II: Predictive Performance of Random Forest for Drug Discovery and Biomarker Identification

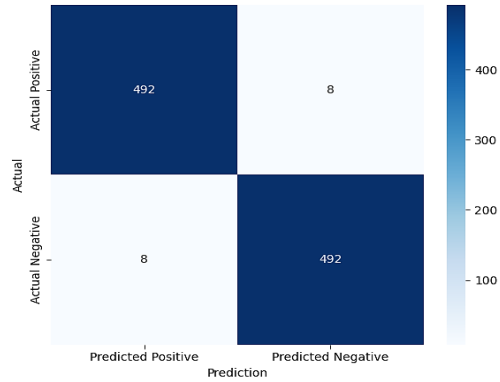

Fig. 4. Confusion Matrix of Random Forest Model for Drug Development and Genetic Biomarker Discovery

The forest model is used in drug development and genetic biomarker discovery. An accurate prediction was seen with 492 true positives and 492 true negatives demonstrating the high predictive accuracy shown in Figure 4. Its sensitivity is high with only 8 false positives, and the specificity is good with only 8 false negatives, which testify to the high reliability of the model and minimal misclassification of the relevant genetic or drug-related biomarkers.

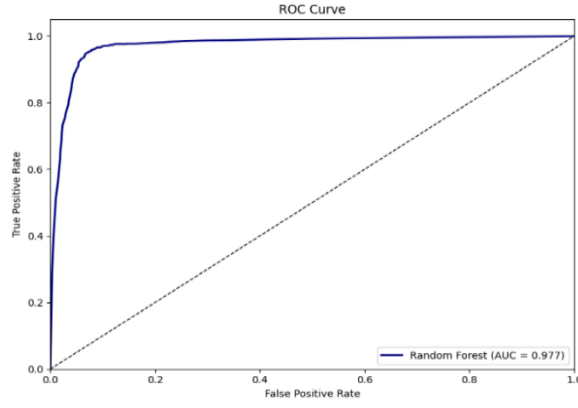

Fig:5 ROC Curve for Random Forest Model in Drug Development and Genetic Biomarker Discovery

In Figure 5, the ROC Curve shows the performance of a Random Forest model in Drug Development and Genetic Biomarker Discovery. A graph illustrating the TPR and FPR at various levels is displayed. With an AUC of 0.977, the model has an outstanding discriminative power, which means that the model was highly accurate in identifying the relevant or non-relevant cases of biomarkers or drug response.

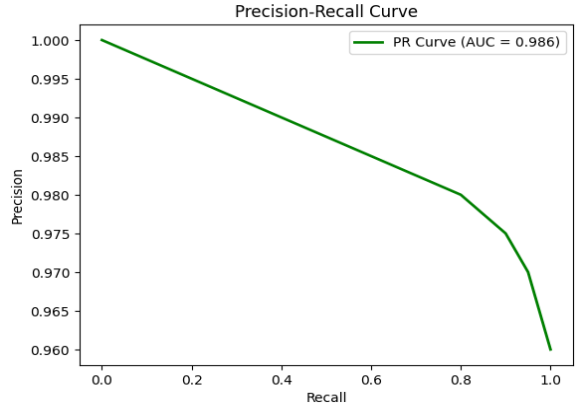

Fig:6 Precision-Recall Curve for Random Forest Model in Drug Development and Genetic Biomarker Discovery

Figure 6 shows the Precision-Recall Plot of a Random Forest model, in the field of Drug Development and Genetic Biomarker Discovery. The curve shows the tradeoff between precision and recall with different threshold settings. The Area Under the Curve (AUC) of the model is a high 0.986, which means that it is a highly effective model with its capacity to identify the presence of relevant positive instances and reduce the number of false positives, particularly in such an essential usage.

Comparative Study

In this section, the analysis of ML models is provided on a comparative basis. Table III represents the correctness of various algorithms in predicting the useful biomarkers and drug response. Random Forest had the highest accuracy of 97.7% emphasizing its effectiveness in working with high-dimensional genetic data, but also producing interpretable results that are necessary in clinical implications. Support Vector Machines (SVM) were close behind with an accuracy of 95%. The Bidirectional Long Short-Term Memory networks (BiLSTM) achieved an accuracy of 80%, while Graph Attention Networks version 2 (GATv2) achieved an accuracy of 77.9% with a new method of modeling relational structures in genetic networks. The results highlight the importance of Random Forest as a principal model to be adopted in AI-based frameworks in order to hasten achievements in personalized medicine and biomarker discovery.

Model	**Accuracy**
Random Forest	97.7
SVM[19]	95
BiLSTM[20]	80
GATv2[21]	77.9

Table: III Comparative Results of Machine Learning Models For Genetic Biomarker Discovery and Drug Response Prediction

The comparative analysis of the model's performance according to the measure of accuracy indicates that the RF model has the best accuracy of 97.7% and therefore is better able to manage the task of classification as indicated in Table 3. SVM came next with an acceptable accuracy of 95 percent which implied that it is very powerful yet just below Random Forest. However, the

deep learning models such as BiLSTM and GATv2 had a relatively lower performance measuring 80% and 77.9 %, respectively. This indicates that more complicated deep learning networks did not perform as well as the more basic models of traditional machine learning in the given scenario, perhaps because of the nature of the dataset, or the representation of its features.

Conclusion and Future Scope

The study proposes an effective machine learning model for predictive drug response modeling and genetic biomarker identification on the GDSC dataset. The use of Random Forest learner with powerful preprocessing methods ensured impressive results, such as an accuracy of 97.7%, large precision, recall, and F1-scores of 98.4%. These metrics show that the model provides an opportunity to reveal complex relationships among genomic data and effectively forecast the therapeutic outcomes. In comparison with SVM, BiLSTM, and GATv2, Random Forest always yielded a better performance in terms of finding relevant biomarkers even though it is still interpretable and computationally meaningful. These accomplishments notwithstanding, it seems that there are still some constraints. The model is based on one set of data which might limit its applicability in wider genomic settings. Besides, although the framework is good with regards to classification, it lacks use of time-series biological data or relational biological data which would add value to even more predictions. Future studies are recommended to work on combination of multi-omics data to amplify prediction accuracy and biological significance. Explainable AI techniques are also possible additions that may bring greater insight to the model decisions leading to clinical trust. Also, the implementation of the model in real-time clinical structures and an assessment of its effectiveness in different types of cancer would guarantee its use in a wider range of cases. These advancements will become the precursor of a data-driven, patient-specific treatment approach in contemporary precision medicine.

References

[1] W. Danter, “Advancing Drug Development with AIHumanoid Simulations: A Virtual Phase 1 Comparative Study of Standard Chemotherapy versus Standard Chemotherapy plus COTI-2 for Pancreatic Adenocarcinoma,” Sep. 10, 2023, Cold Spring Harbor Laboratory Press. Doi: 10.1101/2023.09.08.23295256.

[2] R. Spreafico, L. B. Soriaga, J. Grosse, H. W. Virgin, and A. Telenti, “Advances in Genomics for Drug Development,” Genes (Basel)., vol. 11, no. 8, p. 942, Aug. 2020, doi: 10.3390/genes11080942.

[3] A. D. Hingorani et al., “Improving the odds of drug development success through human genomics: modelling study,” Sci. Rep., vol. 9, no. 1, 2019.

[4] Ó. Álvarez-Machancoses, E. J. D. Galiana, A. Cernea, J. F. de la Viña, and J. L. Fernández-Martínez, “On the Role of Artificial Intelligence in Genomics to Enhance Precision Medicine,” Pharmgenomics. Pers. Med., vol. Volume 13, pp. 105–119, Mar. 2020, doi: 10.2147/PGPM.S205082.

[5] A. Mishra, A. Majumder, D. Kommineni, C. Anna Joseph, T. Chowdhury, and S. K. Anumula, “Role of Generative Artificial Intelligence in Personalized Medicine: A Systematic Review,” Cureus, vol. 17, no. 4, Apr. 2025, doi: 10.7759/cureus 82310.

[6] F. Ahmad, “Optimizing Treatment: The Role of Pharmacology, Genomics, and AI in Improving Patient Outcomes,” Drug Dev. Res., vol. 86, no. 3, p. e70093, 2025.

[7] S. Pandya, “Predictive Modeling for Cancer Detection Based on Machine Learning Algorithms and AI in the Healthcare Sector,” TIJER – Int. Res. J., vol. 11, no. 12, 2024.

[8] S. Pandya, “Integrating Smart IoT and AI-Enhanced Systems for Predictive Diagnostics Disease in Healthcare,” Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 10, no. 6, pp. 2093–2105, Dec. 2024, doi: 10.32628/CSEIT2410612406.

[9] S. Ahmad, S. N. A. Shah, R. Parveen, and K. Raza, "Machine Learning for Genomic Profiling and Drug Discovery in Personalised Lung Cancer Therapeutics," J. Drug Target., vol. 0, no. ja, pp. 1–29, 2025, doi: 10.1080/1061186X.2025.2530656.
[10] J. Mehta, I. Singla, T. Mehra, H. Garg, N. Yadav, and A. K. Goyal, "Leveraging Machine Learning for Biomarker Discovery and Risk Prediction in Autoimmune Disease Genetics," in 2025 10th International Conference on Signal Processing and Communication (ICSC), 2025, pp. 464–469. doi: 10.1109/ICSC64553.2025.10967918.
[11] T. N. Kalyani, M. A. Lakshmi, V. Yamuna, M. Rizvana, V. Shoba, and A. Athiraja, "AI-Driven Predictive Modeling for Accelerated Drug Discovery and Personalized Medicine Development," in 2024 IEEE 1st International Conference on Green Industrial Electronics and Sustainable Technologies (GIEST), 2024, pp. 1–7. doi: 10.1109/GIEST62955.2024.10959983.
[12] H. Zhou, B. Mao, and S. Guo, "Mathematical Modeling of Tumor Growth in Preclinical Mouse Models with Applications in Biomarker Discovery and Drug Mechanism Studies," Cancer Res. Commun., vol. 4, no. 8, pp. 2267–2281, Aug. 2024, doi: 10.1158/2767-9764.CRC-24-0059.
[13] A. Spooner, G. Mohammadi, P. S. Sachdev, H. Brodaty, and A. Sowmya, "Ensemble feature selection with data-driven thresholding for Alzheimer's disease biomarker discovery," BMC Bioinformatics, vol. 24, no. 1, pp. 1–24, 2023, doi: 10.1186/s12859-022-05132-9.
[14] K. Dwivedi, A. Rajpal, S. Rajpal, M. Agarwal, V. Kumar, and N. Kumar, "An explainable AI-driven biomarker discovery framework for Non-Small Cell Lung Cancer classification," Comput. Biol. Med., vol. 153, p. 106544, 2023, doi: https://doi.org/10.1016/j.compbiomed.2023.106544.
[15] M. Çalışkan and K. Tazaki, "AI/ML advances in non-small cell lung cancer biomarker discovery," Front. Oncol., vol. 13, p. 1260374, Dec. 2023, doi: 10.3389/fonc.2023.1260374.
[16] M. Ram, A. Najafi, and M. T. Shakeri, "Classification and biomarker genes selection for cancer gene expression data using random forest," Iran. J. Pathol., vol. 12, no. 4, p. 339, 2017.
[17] L. C, P. S, A. H. Kashyap, A. Rahaman, S. Niranjan, and V. Niranjan, "Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers," Cancer Inform., vol. 22, p. 11769351231167992, Jan. 2023, doi: 10.1177/11769351231167992.
[18] A. Korotcov, V. Tkachenko, D. P. Russo, and S. Ekins, "Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets," Mol. Pharm., vol. 14, no. 12, pp. 4462–4475, 2017.
[19] C.-Y. Lu, Z. Liu, M. Arif, T. Alam, and W.-R. Qiu, "Integration of gene expression and DNA methylation data using MLA-GNN for liver cancer biomarker mining," Front. Genet., vol. 15, Dec. 2024, doi: 10.3389/fgene 2024.1513938.
[20] S. A. Saeedinia, M. R. Jahed-Motlagh, A. Tafakhori, and N. K. Kasabov, "Diagnostic biomarker discovery from brain EEG data using LSTM, reservoir-SNN, and NeuCube methods in a pilot study comparing epilepsy and migraine," Sci. Rep., vol. 14, no. 1, p. 10667, 2024.
[21] Y. Inoue, H. Lee, T. Fu, and A. Luna, "drGAT: Attention-Guided Gene Assessment of Drug Response Utilizing a Drug-Cell-Gene Heterogeneous Network.," ArXiv, no. 2017, 2024.

Nur Mohammad, Mani Prabha, Sadia Sharmin, Rabeya Khatoon, & Md Ahsan Ullah Imran. (2024). COMBATING BANKING FRAUD WITH IT: INTEGRATING MACHINE LEARNING AND DATA ANALYTICS. The American Journal of Management and Economics Innovations, 6(07), 39–56. https://doi.org/10.37547/tajmei/Volume06Issue07-04

Al Wahid, S.A., Mohammad, N., Islam, R., Faisal, Md.H. and Rana, Md.S. (2024) Evaluation of Information Technology Implementation for Business Goal Improvement under Process Functionality in Economic

Development. Journal of Data Analysis and Information Processing, 12, 304-317. doi: 10.4236/jdaip.2024.122017.

Mohammad, N. , Khatoon, R. , Nilima, S. , Akter, J. , Kamruzzaman, M. and Sozib, H. (2024) Ensuring Security and Privacy in the Internet of Things: Challenges and Solutions. Journal of Computer and Communications, 12, 257-277. doi: 10.4236/jcc.2024.128016.

Bhuyan, Md K., et al. "Convolutional Neural Networks Based Detection System for Cyber-attacks in Industrial Control Systems." Journal of Computer Science and Technology Studies, vol. 6, no. 3, 7 Aug. 2024, pp. 86-96, doi:10.32996/jcsts.2024.6.3.9.

B. Biswas, N. Mohammad, M. Prabha, R. M. Jewel, R. Rahman and A. Ghimire, "Advances in Smart Health Care: Applications, Paradigms, Challenges, and Real-World Case Studies," 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS), Cox's Bazar, Bangladesh, 2024, pp. 1-7, doi: 10.1109/COMPAS60761.2024.10796606.

M. K., Khatoon, R. ., Al Mahmud, M. A. ., Tiwari, A. ., ., M. S., Hosain, M. S. ., ., N. M., & Johora, F. T. . (2025). Enhancing Regulatory Compliance in the Modern Banking Sector: Leveraging Advanced IT Solutions, Robotization, and AI. Journal of Ecohumanism, 4(2), 2596–2609. https://doi.org/10.62754/joe.v4i2.6672

Sharmin, S. , Prabha, M. , Johora, F. , Mohammad, N. and Hossain, M. (2024) Open Banking and Information Service: A Strategic Relationship in the FinTech World. Open Journal of Business and Management, 12, 3743-3758. doi: 10.4236/ojbm.2024.125186.

Joy, M. S. A. ., Alam, G. T. ., & Bakhsh, M. M. . (2024). Transforming QA Efficiency: Leveraging Predictive Analytics to Minimize Costs in Business-Critical Software Testing for the US Market. Journal of Artificial Intelligence General Science (JAIGS) ISSN:3006-4023, 7(01), 77–89. https://doi.org/10.60087/jaigs.v7i01.297

Bakhsh, M. M. ., Joy, M. S. A. ., & Alam, G. T. . (2024). Revolutionizing BA-QA Team Dynamics: AI-Driven Collaboration Platforms for Accelerated Software Quality in the US Market. Journal of Artificial Intelligence General Science (JAIGS) ISSN:3006-4023, 7(01), 63–76. https://doi.org/10.60087/jaigs.v7i01.296

Bakhsh, M. M., Alam, G. T. ., & Nadia, N. Y. . (2025). Adapting Agile Methodologies to Incorporate Digital Twins in Sprint Planning, Backlog Refinement, and QA Validation. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 4(2), 67-79. https://doi.org/10.60087/jklst.v4.n2.006

Gazi Touhidul Alam; Md Ismail Jobiullah; Arannita Saha Suspee; Mohammed Majid Bakhsh; Abu Saleh Muhammad Saimon; Syed Mohammed Muhive Uddin. “Creating a Knowledge Hub: AI-Powered Learning Management Systems for BA-QA Training.” Volume. 10 Issue.4, April-2025 International Journal of Innovative Science and Research Technology (IJISRT), 3111-3118, https://doi.org/10.38124/ijisrt/25apr1081

Alam, G. T., Bakhsh, M. M. ., Nadia, N. Y. ., & Islam, S. A. M. . (2025). Predictive Analytics in QA Automation:: Redefining Defect Prevention for US Enterprises. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 4(2), 55-66. https://doi.org/10.60087/jklst.v4.n2.005.