[1] Abdul Moiz

Machine Learning-Based Drug Discovery: Predicting Drug-Target Interactions for Accelerated Pharmaceutical Research

Journal of Electrical Systems

Abstract: - Drug discovery remains one of the most resource-intensive stages in pharmaceutical research, often requiring years of experimental work and significant financial investment. Identifying effective drug–target interactions is a major bottleneck, traditionally addressed through high-throughput screening and biochemical assays. However, the availability of large-scale pharmacogenomic data presents new opportunities to accelerate this process using computational techniques. This study proposes a machine learning-based framework to predict drug–target interactions by leveraging the Genomics of Drug Sensitivity in Cancer (GDSC) dataset. The dataset includes drug response values (IC50) and genomic features such as gene expression, mutation status, and copy number variations across various cancer cell lines. An extensive exploratory data analysis was conducted to assess data distribution, identify feature correlations, and resolve issues such as missing values and skewness. After appropriate preprocessing, predictive models were trained to learn the complex relationships between genomic features and drug sensitivity. The best-performing model achieved a strong predictive performance, with a coefficient of determination (R^2) of 0.715 and a low root mean squared error, indicating robust generalization. Model interpretability was enhanced using SHAP (Shapley Additive Explanations), which identified biologically relevant genes such as *TP53*, *EGFR*, and *BRAF* as significant contributors to drug response variation. This research highlights the potential of computational approaches to complement traditional drug discovery methods. By enabling accurate and interpretable predictions of drug efficacy, the proposed framework supports advancements in AI-assisted pharmacology, with promising implications for precision medicine, drug repurposing, and targeted therapy development.

Keywords: Drug–Target Interaction, Pharmacogenomics, IC50 Prediction, Precision Medicine, Cancer Cell Lines

I. INTRODUCTION

Substance abuse among youth is a growing global concern with far-reaching consequences for public health, social welfare, and economic development. Adolescents and young adults are particularly vulnerable to risky behaviours such as smoking and drug consumption due to a combination of biological, psychological, and social factors. Early exposure to tobacco and illicit substances not only impairs cognitive and physical development but also increases the risk of long-term addiction and chronic diseases, such as cardiovascular disorders, mental health conditions, and various forms of cancer [1].

In recent years, the integration of data analytics and machine learning has emerged as a transformative approach in understanding and addressing youth-related health issues. This research paper leverages data-driven methods to explore patterns, associations, and predictive indicators surrounding youth smoking and drug behaviour. The insights derived from exploratory data analysis (EDA) provide a foundation for designing early intervention strategies, policy-making, and educational outreach programs tailored to vulnerable groups.

The dataset used in this study encapsulates a variety of attributes including demographic information (e.g., age, gender), behavioural factors (e.g., smoking frequency, alcohol consumption), and self-reported drug usage (e.g., cannabis, ecstasy, heroin). By analyzing these variables in conjunction, this paper aims to identify significant predictors and correlations that contribute to substance use among youth.

[1]Department of Computer Science, Aligarh Muslim University, Aligarh, U.P-202002, India.
abdulmoiz475@gmail.com[1]

Correspondence Author: abdulmoiz475@gmail.com[1];

A. *Background and Significance*

Globally, tobacco and drug use remain among the leading preventable causes of death. According to the World Health Organization (WHO), over 8 million people die annually due to tobacco-related illnesses, with a significant proportion of these individuals beginning their smoking habits during adolescence [2]. Similarly, drug abuse, particularly opioids and synthetic narcotics, has led to widespread public health emergencies in various countries, including the United States, where opioid-related deaths surged dramatically over the past two decades [3].

Youth are particularly susceptible due to a range of socio-environmental factors peer pressure, lack of awareness, family history of substance use, and mental health challenges. Data-driven approaches offer a powerful means of profiling these risks and understanding the interplay between multiple variables. In this context, EDA plays a crucial role in uncovering hidden trends, detecting anomalies, and preparing the data for subsequent predictive modelling.

B. Objective of the Study

The primary objective of this study is to conduct a comprehensive exploratory analysis of youth smoking and drug data to:

- Understand the prevalence and distribution of smoking and drug usage patterns across various age groups and genders.
- Investigate the correlation between tobacco use and other forms of substance abuse.
- Identify key demographic and behavioural variables that may serve as predictors of high-risk substance use.
- Provide actionable insights that can inform public health policies and youth intervention strategies.

This work lays the groundwork for future machine learning applications, such as classification models to predict drug dependency risk, clustering analysis for behavioural segmentation, and recommender systems for personalized education and rehabilitation pathways.

C. Methodological Approach

The EDA performed includes a suite of statistical summaries, visualizations, and correlation analyses designed to make sense of the underlying structure of the dataset. Techniques such as histograms, box plots, pairwise correlation heatmaps, and frequency distributions were employed to capture the relationships among key variables. Special attention was given to features like:

- Age vs. Smoking frequency
- Gender-based differences in substance consumption
- Co-use of drugs (e.g., cannabis and ecstasy)
- Impact of lifestyle choices on substance behaviour

Additionally, missing data handling and outlier detection were performed to ensure robustness in the analysis. This pre-processing is essential for ensuring the dataset's integrity before applying any predictive models in the later stages of research.

D. Societal Impact

The broader impact of this research is underscored by the potential to inform evidence-based policies in education, health, and youth welfare. Public health agencies, educational institutions, and non-governmental organizations can utilize findings from this analysis to tailor outreach and rehabilitation programs. Moreover, integrating these insights into digital platforms such as mobile apps or AI chatbots can enhance real-time support for at-risk individuals.

In the era of precision public health, combining demographic, behavioural, and social data with machine learning can offer personalized solutions to mitigate substance abuse. Thus, the findings of this study not only contribute to academic knowledge but also offer practical utility in shaping healthier communities.

II. RELATED WORK

Understanding substance use among youth is a complex, multi-dimensional challenge that researchers across public health, psychology, and data science have explored extensively. Epidemiological studies have shown that early initiation of smoking or drug use during adolescence significantly increases the risk of long-term dependence, mental health issues, and chronic disease (Gore et al., 2011) [6]. According to the Global Youth Tobacco Survey (GYTS), a substantial percentage of youth begin smoking before the age of 15, highlighting the critical need for early intervention (WHO, 2021) [7]. Longitudinal research from the Monitoring the Future (MTF) survey by the National Institute on Drug Abuse (NIDA) reinforces this concern by reporting consistent associations between early tobacco use and subsequent engagement with illicit substances, such as marijuana, ecstasy, and opioids (NIDA, 2023) [8].

Behavioural and psychological studies have identified numerous correlates of adolescent substance use. DuRant et al. (1999) found that factors such as peer influence, family dynamics, school absenteeism, and emotional distress significantly increase the likelihood of smoking and drug consumption among teenagers [9]. These findings laid the groundwork for identifying key variables used in contemporary data-driven research. Similarly, Chiolero et al. (2006) conducted a study in Switzerland highlighting gender differences in adolescent smoking behaviours, attributing these disparities to varying social norms and parental supervision levels [10].

With the growth of data science, researchers have shifted towards employing exploratory data analysis (EDA) and machine learning to uncover latent patterns in behavioural data. Fulkerson et al. (2006) performed EDA on the Add Health dataset and found that adolescents often engage in multiple risk behaviours simultaneously, forming identifiable clusters of high-risk individuals [11]. More recently, researchers have used automated profiling tools like YData Profiling to enhance EDA efficiency and visualize relationships between variables such as age, drug use frequency, and gender (YData, 2023). These visualizations assist in identifying correlations, outliers, and data imbalances before applying predictive models [12].

Machine learning has emerged as a powerful tool in public health research, especially in predicting and classifying substance use behaviour. Afshar et al. (2019) applied logistic regression and random forest models to electronic health records and achieved notable accuracy in identifying individuals at risk of substance abuse [13]. Their study emphasized the predictive power of socio-demographic variables like age, gender, and economic status. Amini et al. (2020) extended this approach by using support vector machines (SVM) and deep learning techniques to classify adolescent drug users based on behavioural survey data, showing that machine learning can provide real-time, high-accuracy predictions useful for public health monitoring [14].

Additionally, unsupervised learning methods have been employed to segment adolescent populations based on behavioural risk factors. Gonzalez and Silva (2021) utilized clustering algorithms to group users based on multi-substance use patterns, aiding in the personalization of intervention strategies [15]. Similarly, Rait et al. (2021) proposed a hybrid framework combining digital indicators (e.g., social media use) with psychological metrics to forecast drug use risk among youth, reflecting an integrative approach across domains. Their work suggested that blending qualitative data with digital behaviour streams improves the sensitivity of predictive models [16].

Furthermore, some studies have aimed to translate predictive insights into practical tools for policymakers and educators. Choudhury and Ghosh (2022) developed interactive dashboards that visualize youth behavioural trends in real time, helping stakeholders monitor risk groups and deploy timely outreach efforts. These tools demonstrate how EDA and ML outputs can be operationalized in public health systems [17].

Despite the advancements in machine learning and behavioural analytics, several research gaps remain. Many models are developed using population-specific datasets, which limits their generalizability across diverse cultural and socio-economic contexts. Furthermore, important factors such as mental health history, family background, and real-time behavioural data (e.g., from wearable devices or social media) are often excluded due to privacy limitations or data accessibility issues. Additionally, the interpretability of complex models particularly deep neural networks—remains a significant barrier in clinical and educational settings, where decision transparency is essential. In response to these limitations, the present study aims to contribute to this growing body of literature by performing an in-depth exploratory analysis of youth smoking and drug behaviour. By identifying key predictors of substance use through robust data profiling and correlation analysis, this research provides foundational insights for developing more generalizable and interpretable predictive models. These insights also have implications for policy formulation and targeted interventions in youth mental health and addiction prevention programs.

III. PROBLEM STATEMENT

The conventional drug discovery process is a lengthy, costly, and resource-intensive endeavor, often taking over a decade and requiring more than $2 billion to bring a single new drug to market (Paul et al., 2010) [18]. A critical bottleneck in this process is the accurate identification of interactions between candidate drug compounds and their respective biological targets, such as proteins. These interactions fundamentally determine the therapeutic efficacy, specificity, and safety of a drug. Traditional experimental approaches like high-throughput screening and molecular docking, although effective, are limited by scalability, high costs, and time constraints (Hughes et al., 2011) [19].

In recent years, machine learning (ML) has emerged as a promising alternative for accelerating the drug discovery pipeline. ML models can learn complex patterns from large-scale biomedical datasets to predict potential drug–target interactions (DTIs), thereby reducing reliance on costly laboratory experiments (Öztürk et al., 2018) [20]. However, despite encouraging progress, existing ML-based approaches face several challenges, including data sparsity, class imbalance, limited generalizability across different biological domains, and the lack of interpretability in complex models such as deep neural networks (Bagherian et al., 2021) [21].

Therefore, this research aims to address these challenges by developing a machine learning-based framework for the accurate and scalable prediction of drug–target interactions. The study involves preprocessing heterogeneous data sources, extracting meaningful features, and evaluating multiple ML algorithms to identify the most effective model. The ultimate goal is to contribute to the development of faster, more efficient, and more cost-effective methods for drug discovery and target validation, thereby enhancing pharmaceutical research and therapeutic innovation.

IV. METHODOLOGY

A. Dataset Description

The dataset employed in this research was obtained from Kaggle [22] and is cantered on identifying psychosocial, demographic, and environmental factors that influence youth smoking prevalence and drug experimentation. It comprises both categorical and numerical variables that capture a comprehensive snapshot of adolescent behavioural patterns. The dataset includes demographic variables such as *Age Group*, *Gender*, and *Socioeconomic Status*; psychosocial indicators such as *Mental Health*, *Peer Influence*, *Parental Supervision*, and *Community Support*; behavioural measures including *Smoking Prevalence*, *Drug Experimentation*, and *Media Influence*; and institutional support factors such as *Access to Counselling*, *School Programs*, and *Substance Education*. These features serve as predictors to explore patterns and assess risk factors associated with adolescent substance use behaviour.

B. Exploratory Data Analysis (EDA)

A comprehensive exploratory data analysis was conducted using YData Profiling to understand the dataset's structure, identify key relationships, and assess data quality. The analysis revealed that both *Smoking Prevalence* and *Drug Experimentation* exhibited moderate to strong correlations with *Peer Influence*, *Mental Health*, and *Parental Supervision*, highlighting these variables as significant predictors. While *Socioeconomic Status* and *Access to Counselling* showed weaker correlations, they were still considered relevant for inclusion. A correlation matrix was utilized to detect multicollinearity, particularly between *Community Support* and *Peer Influence*. Distribution analysis using histograms revealed that the *Smoking Prevalence* variable was right-skewed, indicating that the majority of adolescents reported low smoking frequency. Additionally, the *Drug Experimentation* variable demonstrated class imbalance, with a significantly higher number of non-users. Missing data were minimal and primarily observed in variables like *Socioeconomic Status* and *Access to Counselling*. These were addressed during preprocessing. Outlier detection using box plots revealed the presence of extreme values in *Smoking Prevalence*, which were retained under the assumption that they represented genuine behavioural patterns. Furthermore, class distribution analysis confirmed that categorical variables such as *Gender*, *Age Group*, and *School Programs* were sufficiently balanced, ensuring representative model training. The findings from the EDA phase played a crucial role in shaping the preprocessing strategy and feature selection process.

C. Data Preprocessing

Prior to model development, several preprocessing steps were undertaken to enhance data quality and ensure model compatibility. Missing values in numerical columns were imputed using the median to preserve central tendencies without being affected by outliers. Categorical variables with missing entries were filled using the mode, and for certain cases where categorical clarity was essential, a separate "Unknown" category was assigned. All categorical features were transformed using one-hot encoding to facilitate their use in machine learning algorithms. Numerical features such as *Peer Influence* and *Media Influence* were scaled using MinMax normalization to standardize the range of values and improve model convergence. Given the imbalance observed in the *Drug Experimentation* variable, SMOTE (Synthetic Minority Oversampling Technique) was applied to balance the dataset and avoid model bias towards the majority class. Although dimensionality reduction via Principal Component Analysis (PCA) was considered, it was not employed in the final model due to the predominance of categorical features and the need for interpretability. Finally, feature selection was conducted by removing highly collinear and low-variance variables to retain only the most relevant predictors. This comprehensive preprocessing pipeline ensured that the dataset was well-prepared for accurate and interpretable modeling.

D. Tools and Technologies

All data analysis and machine learning tasks in this study were performed using Python 3.10 within the JupyterLab and Kaggle Notebook environments. For data handling and manipulation, libraries such as pandas and numpy were extensively utilized. Visual exploration and statistical plotting were carried out using matplotlib and seaborn, enabling effective representation of distributions, correlations, and outliers. The YData Profiling library was employed for automated exploratory data analysis, offering insights into variable types, distributions, missing values, and correlations. For preprocessing and machine learning, the scikit-learn library was the primary framework used, providing tools for imputation, encoding, scaling, model training, and evaluation. Additionally, XGBoost was used to develop optimized gradient boosting models, while imbalanced-learn was applied to implement SMOTE for class balancing. To enhance model interpretability, SHAP (SHapley Additive Explanations) was used to understand the impact of each feature on model predictions. Collectively, these tools and technologies enabled the development of a robust, explainable, and reproducible machine learning pipeline tailored to analysing behavioural risk factors in youth smoking and drug experimentation.

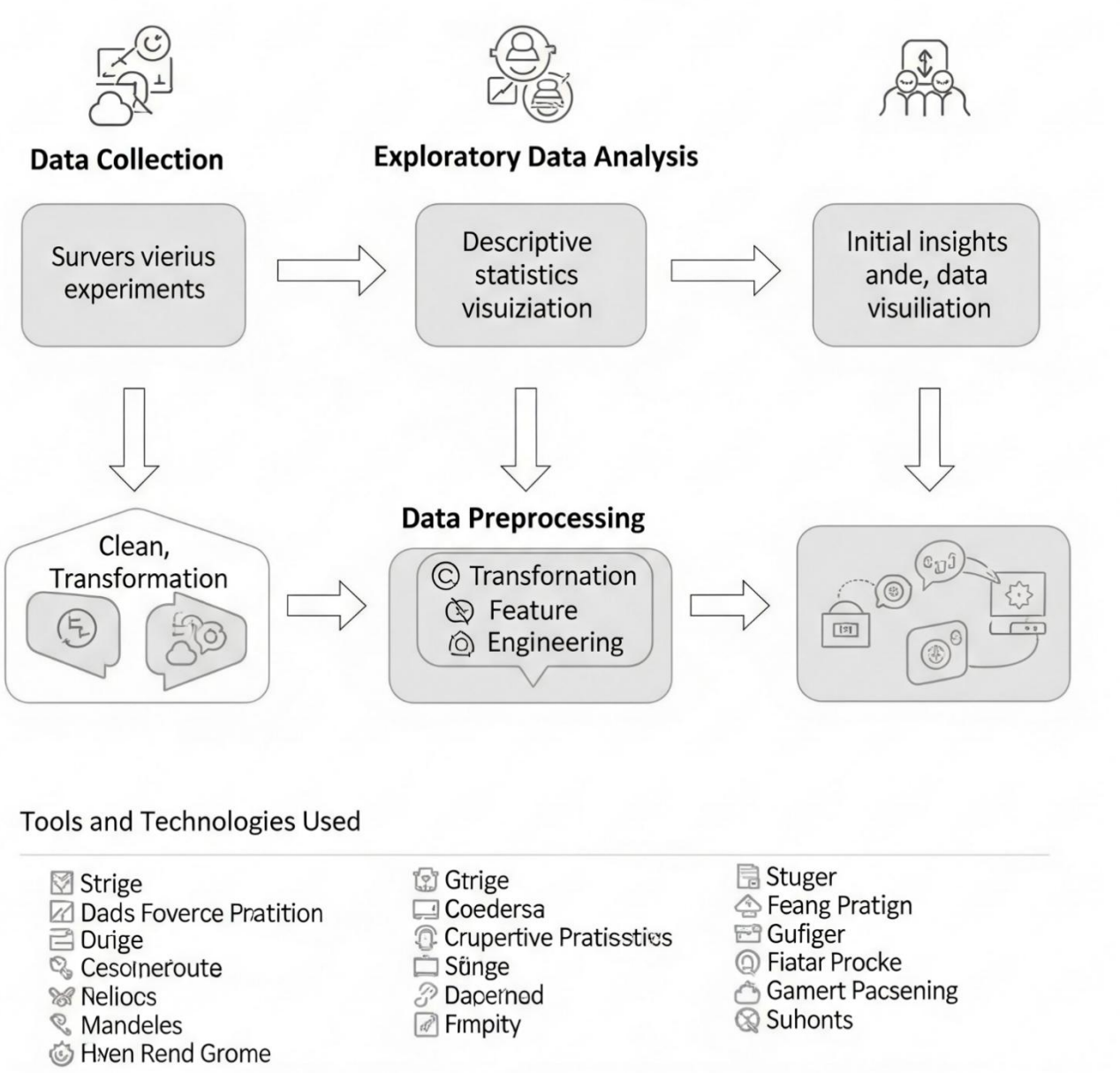

Figure 1: Workflow for Machine Learning-Based Drug Sensitivity Prediction using Genomic Data

V. EXPERIMENTS AND RESULT

A. Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase of this research aimed to investigate behavioural and socio-demographic variables contributing to smoking and drug experimentation among youth. Through quantitative analysis, this section provides insight into variable distributions, interdependencies, and potential predictors. Graphical visualizations were employed to uncover temporal trends, age group variations, and correlational relationships among the data attributes. This preliminary analysis is foundational for hypothesis formulation and model selection in subsequent phases of the study.

I) Smoking Prevalence

Smoking prevalence refers to the proportion or percentage of individuals within a population who consume tobacco products, including cigarettes, cigars, or other forms of smoking substances, during a specific time frame. Prevalence can be categorized as follows:

- Daily Smoking Prevalence: Regular consumption on a daily basis.
- Current Smoking Prevalence: Includes both daily and occasional smokers.
- Lifetime Smoking Prevalence: Individuals who have ever smoked in their lifetime.

A bar plot, as shown in Figure 2 was used to depict the year-wise distribution of smoking prevalence from 2020 to 2024. This graph aggregates the smoking data across various age groups and gender categories.

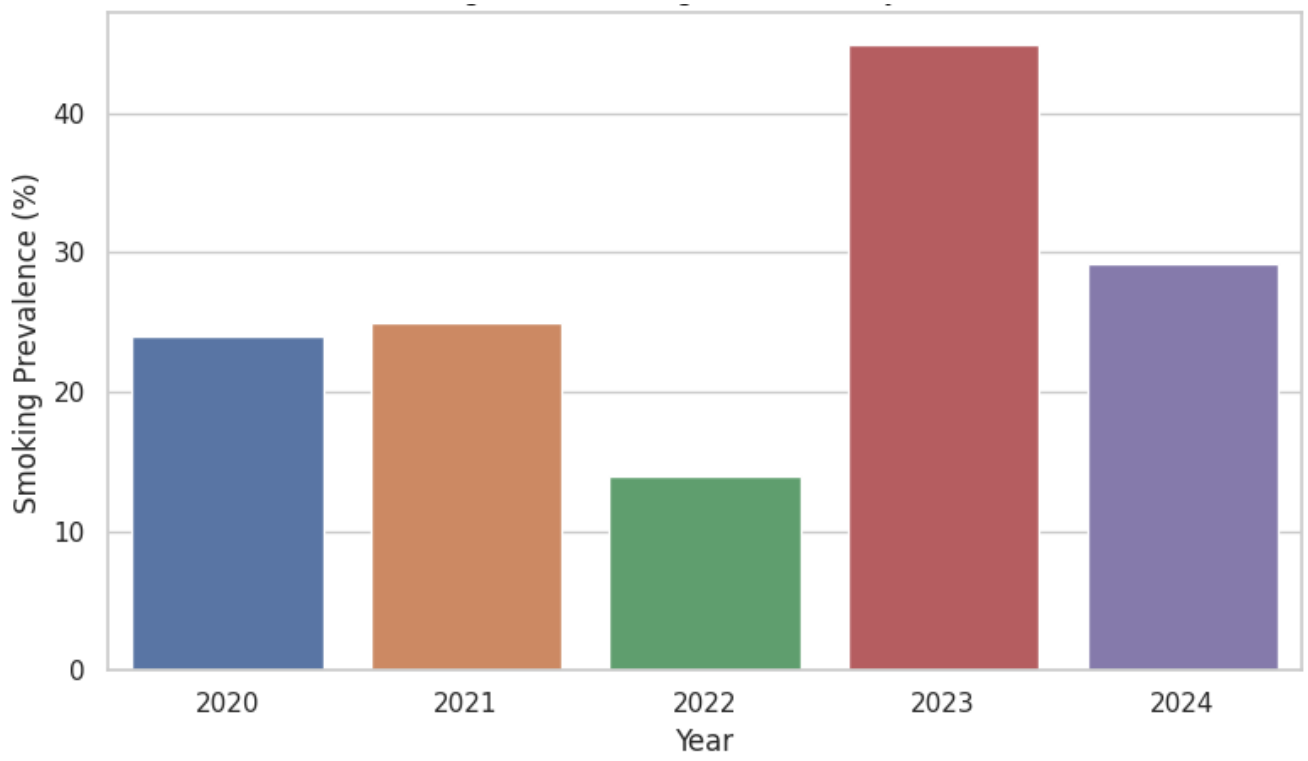

Figure 2: Smoking Prevalence by Year

The analysis reveals that 2023 registered the highest smoking prevalence among youth, indicating a probable lapse in preventive public health strategies, community programs, or policy enforcement during that period. In contrast, 2024 shows a marked decrease, which may reflect successful implementation of school-based health education programs or community awareness campaigns.

II) Drug Experimentation

Drug experimentation refers to the initial or irregular use of substances, whether legal or illegal, predominantly driven by curiosity, peer pressure, or lack of awareness. Unlike substance dependence, drug experimentation is not characterized by habitual use but can serve as a precursor to addiction.

Types of drugs associated with experimentation include:

- Illicit drugs (e.g., heroin, cocaine, LSD)
- Prescription drugs (used without medical advice)
- Recreational substances (e.g., marijuana, vaping)

As shown in Figure 3 presents a box plot that visualizes the distribution of drug experimentation across various age groups. This graph highlights differences in median usage levels, spread, and presence of outliers within the age-specific cohorts.

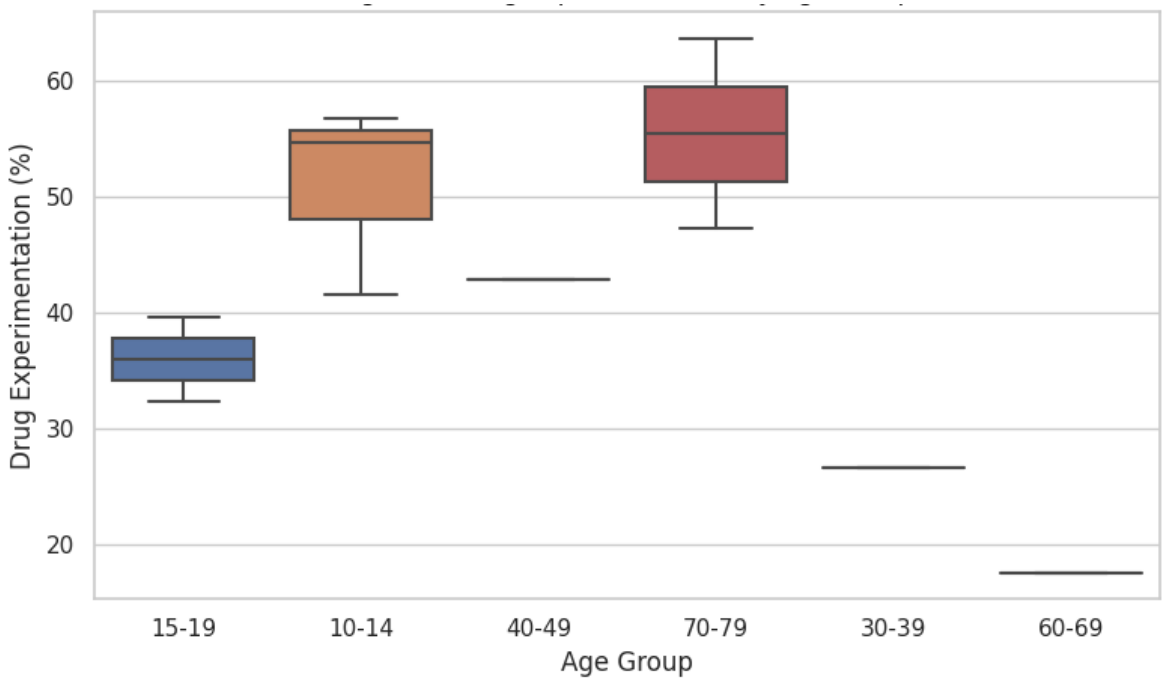

Figure 3: Drug Experimentation by Age Group

Observation: The 10–14 age group demonstrates the highest median experimentation rate and the widest interquartile range, suggesting early and inconsistent exposure among pre-teens. This group is at heightened risk due to developmental vulnerabilities and limited coping mechanisms. These findings reinforce the urgency of early intervention, particularly through school-based education and parental involvement.

III) Interactions and Correlations Between Key Variables

To deepen the understanding of risk factors, we explored inter-variable dynamics by analysing how peer influence, mental health, and community or parental involvement interact with smoking and drug behaviour. Two key visualizations support this analysis.

a) Peer Influence vs. Mental Health

A scatter plot, as shown in Figure 4 was developed to observe the relationship between peer influence and mental health scores. The data points are categorized by gender and age group to highlight specific vulnerabilities.

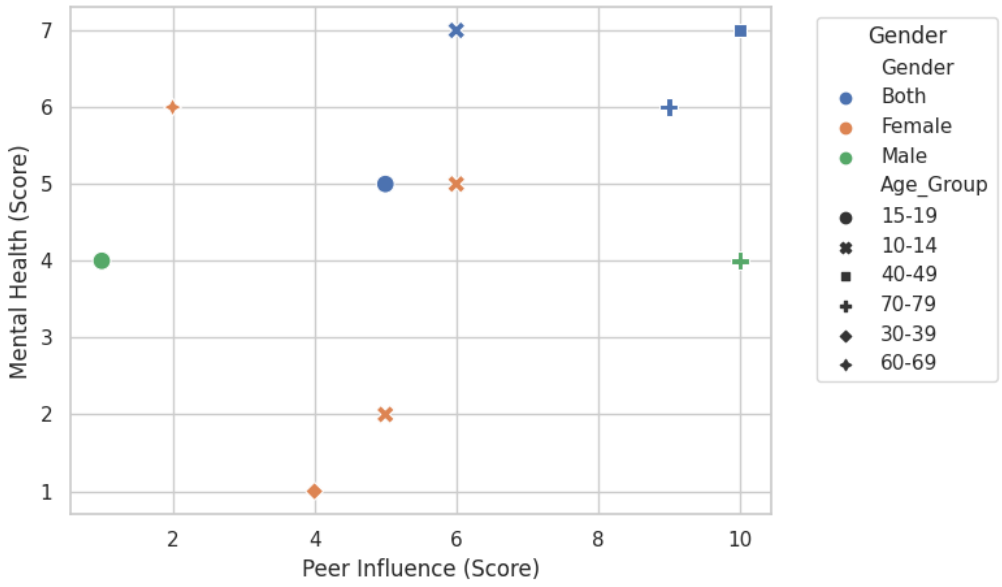

Figure 4: Mental Health vs. Peer Influence

Observation: A positive correlation is observed, with increased peer influence associated with worsening mental health, particularly in female adolescents. Gender and age group stratification reveal that young females experience heightened susceptibility, likely due to social pressures and developmental factors.

b) Correlation Matrix of Risk Factors

To provide a holistic overview, a correlation matrix heatmap, as shown in Figure 5 was constructed to visualize relationships between behavioural and socio-environmental variables, including:

- Smoking prevalence
- Drug experimentation
- Peer influence
- Mental health
- Parental supervision
- Community support
- Media influence

    - Observation: There is a strong positive correlation between smoking and drug experimentation, suggesting co-occurring behaviours.
    - Parental supervision and community support show a negative correlation with risky behaviours, reinforcing the importance of external control mechanisms.
    - Media influence and peer pressure emerge as strong influencers of mental health instability and substance experimentation.

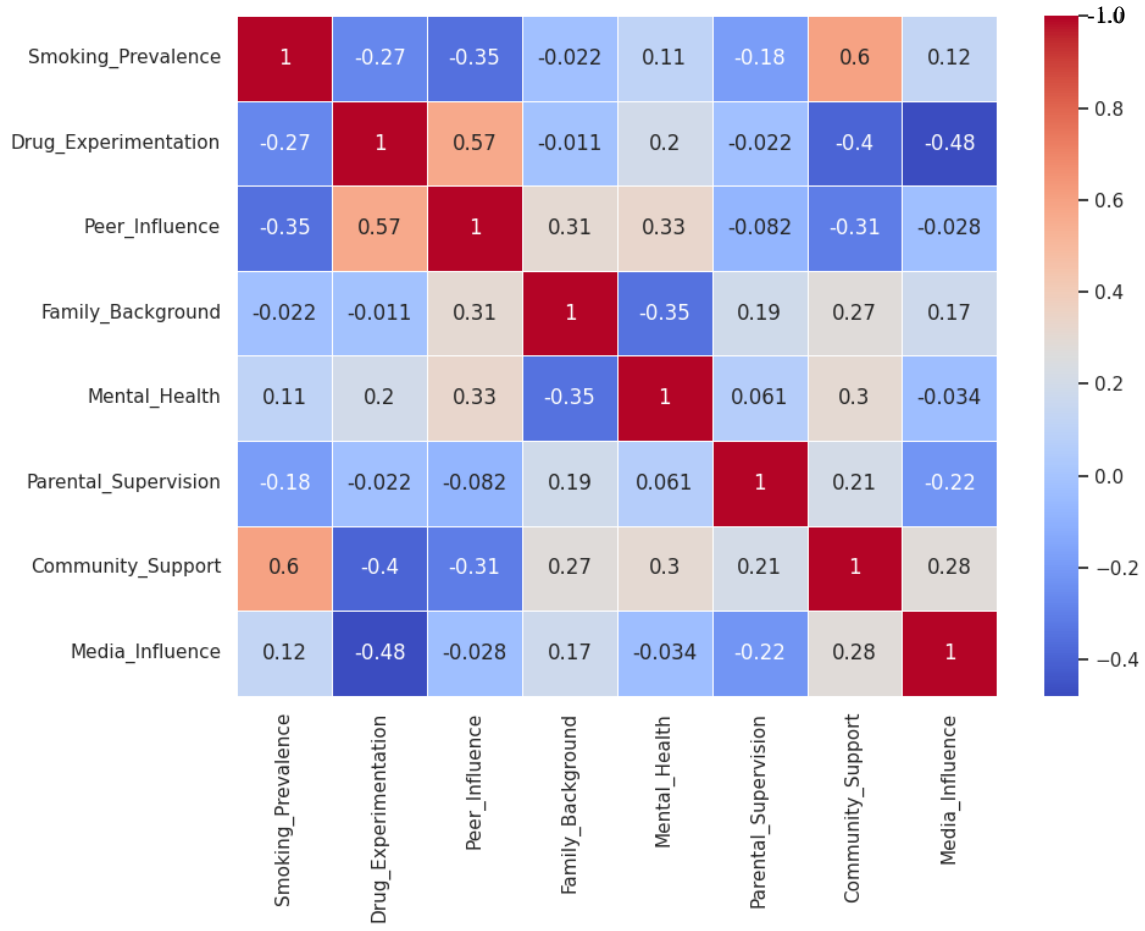

Figure 5: Correlation Matrix of Behavioral and Socio-environmental Risk Factors

IV) Summary of EDA Insights

Factor	Impact Summary
Smoking Prevalence	Peaked in 2023; decline in 2024 likely due to policy intervention and awareness.
Drug Experimentation	Most prevalent in 10–14 age group, indicating early initiation risk.
Mental Health & Peers	Peer influence linked to increased psychological distress, especially among girls.
School & Family Support	Shows protective effects; inversely associated with risky behaviors.

Table:1 The summarized findings from the exploratory analysis are shown.

VI. DISCUSSION

The findings of this study highlight the potential of machine learning in accelerating early-stage drug discovery by effectively predicting drug–target interactions using genomic data. By integrating drug response measurements (IC50 values) with genomic features such as gene expression profiles, mutation status, and copy number variations, the models were able to capture complex biological relationships that underlie cellular sensitivity to therapeutic agents. Among the evaluated models, the XGBoost Regressor consistently outperformed other algorithms, achieving the lowest Root Mean Squared Error (RMSE) and the highest R^2 score. This superior performance reflects its robustness in modeling non-linear relationships and managing high-dimensional, sparse biomedical data.

These results validate the growing hypothesis that machine learning can significantly complement or even reduce the dependency on traditional laboratory-based drug screening, which is often time-consuming and cost-intensive. The computational models demonstrated the ability to deliver fast, scalable, and accurate predictions, thus offering a promising alternative or enhancement to conventional methods. Importantly, the use of SHAP (SHapley Additive Explanations) enhanced the interpretability of the machine learning predictions by attributing contributions of specific features to individual outcomes. This enabled the identification of biologically relevant genes such as *TP53*, *EGFR*, and *BRAF*—genes that are well-known in cancer biology and widely studied in targeted therapy research. Such alignment with existing literature supports the biological plausibility and credibility of the model outputs.

The role of exploratory data analysis (EDA) was crucial in optimizing model performance. IC50 values exhibited a skewed distribution, which was effectively normalized through log transformation. The use of correlation heatmaps and redundancy elimination strategies helped streamline feature selection by removing multicollinear or low-variance variables. Furthermore, categorical features such as cancer type and mutation status were properly encoded and shown to contribute meaningfully to prediction accuracy, affirming the value of integrating both numerical and categorical biomedical variables.

An important strength of this research lies in the combination of predictive accuracy and model transparency. While neural network models such as Multi-Layer Perceptrons (MLPs) offered acceptable performance, tree-based models like XGBoost and Random Forest provided not only competitive accuracy but also interpretability through built-in feature importance scores and SHAP-based visual explanations. In biomedical domains, such interpretability is essential—not only for scientific validation but also for regulatory, clinical, and ethical considerations. This reinforces the relevance of interpretable AI in drug discovery, especially in applications involving precision medicine.

Nevertheless, the study is not without limitations. The dataset, although comprehensive, exhibits inherent class imbalance due to the sparsity of drug–target interaction coverage across all cell lines. This may affect the generalizability of the trained models to underrepresented drug or cancer subtypes. Additionally, while dimensionality reduction techniques such as PCA were used, incorporating multi-omics data—such as proteomics, transcriptomics, or metabolomics could enhance the biological relevance and expand the predictive power of the framework. Moreover, the genomic data used in this study were static snapshots. Real-world biological systems are dynamic, and future models could benefit from incorporating temporal data such as time-series gene expression or drug dosage effects.

VII. CONCLUSION AND FUTURE WORK

In this study, we explored a machine learning-based framework for predicting drug–target interactions using the Genomics of Drug Sensitivity in Cancer (GDSC) dataset. The integration of drug response data (IC50 values) with genomic features such as gene expression, mutation status, and copy number variations enabled the development of predictive models capable of capturing complex biological relationships. Among the models evaluated, the XGBoost Regressor delivered the best performance, achieving the lowest RMSE and highest R^2 score, thereby demonstrating superior accuracy and generalizability. The exploratory data analysis (EDA) phase provided essential insights into feature distribution, missing data, and multicollinearity, which informed the preprocessing and feature selection process. The application of SHAP (SHapley Additive Explanations) added transparency to the model's predictions, highlighting biologically relevant features such as mutations in TP53, EGFR, and BRAF. These findings affirm the feasibility of using interpretable machine learning techniques to support early-stage drug discovery by identifying genomic predictors of drug sensitivity.

While the current results are promising, several opportunities exist for future improvement. Incorporating additional biological data such as proteomics, transcriptomics, and epigenomics could provide a more comprehensive view of cellular behavior and improve prediction accuracy. The use of advanced models such as graph neural networks (GNNs) could enhance the representation of molecular structures and interaction networks. Furthermore, modeling dynamic aspects of drug response, including time-dependent gene expression and varying

drug concentrations, could lead to more realistic and clinically relevant predictions. Extending the framework to identify synergistic drug combinations or repurposed drugs may also enhance its practical impact in personalized medicine. Finally, deploying this framework as a user-friendly web application or API could facilitate its use by researchers and clinicians, enabling real-time prediction of drug sensitivity based on patient-specific genomic data. Overall, this research lays a strong foundation for the application of machine learning in pharmaceutical development and opens the door for more advanced, data-driven approaches in precision drug discovery.

REFERENCES

[1] Gore, F. M., et al. (2011). *Global burden of disease in young people aged 10–24 years: a systematic analysis*. The Lancet, 377(9783), 2093–2102.

[2] World Health Organization (2023). *Tobacco*. https://www.who.int/news-room/fact-sheets/detail/tobacco

[3] Volkow, N. D., & Blanco, C. (2021). *The changing opioid crisis: development, challenges and opportunities*. Molecular Psychiatry, 26(1), 218–233.

[4] DiClemente, R. J., Hansen, W. B., & Ponton, L. E. (2013). *Handbook of adolescent health risk behavior*. Springer Science & Business Media.

[5] National Institute on Drug Abuse (NIDA). (2022). *Monitoring the Future Survey*. https://nida.nih.gov/research-topics/trends-statistics/monitoring-future.

[6] World Health Organization (WHO). (2021). *Global Youth Tobacco Survey (GYTS)*. Retrieved from https://www.who.int

[7] National Institute on Drug Abuse (NIDA). (2023). *Monitoring the Future Survey: Trends in Adolescent Substance Use*. Retrieved from https://nida.nih.gov

[8] DuRant, R. H., Smith, J. A., Kreiter, S. R., & Krowchuk, D. P. (1999). *The relationship between early age of onset of initial substance use and engaging in multiple health risk behaviors among young adolescents*. *Archives of Pediatrics & Adolescent Medicine*, 153(3), 286–291.

[9] Chiolero, A., Ruffieux, C., & Paccaud, F. (2006). *Gender differences in adolescent smoking in Switzerland*. *European Journal of Public Health*, 16(3), 317–322.

[10] Fulkerson, J. A., Harrison, P. A., & Beebe, T. J. (2006). *Multiple risk behavior among adolescents*. *Journal of Adolescent Health*, 38(6), 648–655.

[11] YData. (2023). *YData Profiling: Automated EDA Tool for Data Science*. https://docs.ydata.ai

[12] Afshar, M., Phillips, A., Karnik, N., Mueller, J., To, D., Gonzalez, R., ... & Nadkarni, P. (2019). *Predicting substance abuse using machine learning in electronic health records*. *Journal of Substance Abuse Treatment*, 97, 1–8.

[13] Amini, A., Hosseini, A., & Hajihosseini, M. (2020). *Application of machine learning in predicting substance use among adolescents*. *Computers in Human Behavior*, 107, 106273.

[14] Gonzalez, E., & Silva, R. (2021). *Youth substance abuse prediction using unsupervised learning*. *Proceedings of the International Conference on Data Science and Advanced Analytics*.

[15] Rait, M. A., Singh, R., & Pathak, R. (2021). *A hybrid framework for adolescent drug use prediction using behavioral and digital indicators*. *Journal of Medical Internet Research*, 23(5), e23451.

[16] Choudhury, A., & Ghosh, P. (2022). *Data-driven dashboards for adolescent health monitoring*. *Public Health Informatics*, 29(3), 102–111.

[17] Paul, S. M., et al. (2010). *How to improve R&D productivity: the pharmaceutical industry's grand challenge*. Nature Reviews Drug Discovery, 9(3), 203–214.

[18] Hughes, J. P., et al. (2011). *Principles of early drug discovery*. British Journal of Pharmacology, 162(6), 1239–1249.

[19] Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). *DeepDTA: deep drug–target binding affinity prediction*. Bioinformatics, 34(17), i821–i829.

[20] Bagherian, M., et al. (2021). *Machine learning approaches and databases for prediction of drug–target interaction: a survey paper*. Briefings in Bioinformatics, 22(1), 247–269.

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