Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning for Gene-Expression-Based Prediction of Alcohol Use Disorder with an Integrate Web-Based Dashboard

Hanan Alotaibi

hmalotaibi@kau.edu.sa

King Abdulaziz University

Alaa Bafail

King Abdulaziz University

Dania Aljeaid

King Abdulaziz University

Doaa Sinnari

King Abdulaziz University

Hebah Albatati

King Abdulaziz University

Mona Alkhozae

King Abdulaziz University

Rania M. Alhazmi

King Abdulaziz University

Joud Alharabi

Farah Najjar

Shahad Almasabi

Aisha Alsobhi

King Abdulaziz University

Research Article

Keywords: Alcohol Use Disorder (AUD), Gene Expression, Machine Learning, Elastic Net Logistic Regression, Feature Selection, Interpretable Machine Learning, RNA Sequencing

Posted Date: October 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11096122/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Machine Learning for Gene-Expression-Based Prediction of Alcohol Use Disorder with an Integrate Web-Based Dashboard

Hanan Alotaibi[1], Alaa Bafail[2], Dania Aljeaid[3], Doaa Sinnari[1], Hebah Albatati[1], Mona Alkhozae[1], Rania M. Alhazmi[1], Joud Alharabi, Farah Najjar, Shahad Almasabi, Aisha Alsobhi[3]

Abstract

Alcohol use disorder (AUD) is a complex condition associated with molecular alterations in brain regions involved in reward, motivation, and behavioral regulation. Gene-expression profiling can provide valuable molecular information for studying AUD; however, the high dimensionality and correlation structure of transcriptomic data present challenges for predictive modeling and interpretation. This study develops a machine-learning approach for predicting AUD status from gene-expression profiles and introduces GENO, a web-based dashboard designed to support laboratory users in interpreting the resulting predictions. Publicly available bulk RNA-sequencing data from postmortem brain tissue of 114 individuals, including 58 AUD cases and 56 non-AUD controls, were obtained from the Gene Expression Omnibus repository (GSE253155). Gene-expression data were standardized, and dimensionality was reduced using ANOVA F-test and mutual-information-based feature selection. An Elastic Net logistic regression model was then developed to manage high-dimensional and correlated predictors while maintaining model interpretability. On the held-out test set, the model achieved an accuracy of 86%, an F1-score of 0.865, and a receiver operating characteristic area under the curve (ROC-AUC) of 0.895. For each input profile, the model generates a probability-based prediction of membership in the AUD class and estimates gene-level contributions to the prediction. GENO integrates this predictive pipeline into a web-based interface that supports data upload, automated analysis, result visualization, and structured report generation. A preliminary task-based usability assessment with laboratory users further demonstrated the feasibility of completing the platform's primary workflow. These findings provide preliminary evidence that interpretable regularized machine learning can predict AUD status from postmortem gene-expression profiles within the analyzed dataset.

Keywords: Alcohol Use Disorder (AUD); Gene Expression; Machine Learning; Elastic Net Logistic Regression; Feature Selection; Interpretable Machine Learning; RNA Sequencing.

1. Introduction

Alcohol use disorder (AUD) is a complex and persistent condition associated with substantial health, social, and economic consequences. It is characterized by impaired control over alcohol consumption and continued use despite adverse consequences. Although behavioral and environmental factors contribute significantly to its development and progression, AUD is increasingly understood as a multifactorial disorder arising from interactions among genetic susceptibility, neurobiological processes, environmental exposures, and individual life experiences (Deak & Johnson, 2021; Hancock et al., 2018; Volkow et al., 2016).

Genetic and molecular studies have contributed to a better understanding of the biological mechanisms associated with AUD. Genetic factors influence individual susceptibility to alcohol-related behaviors and disorders; however, AUD is highly polygenic and cannot be explained by a small number of genetic variants alone (Deak & Johnson, 2021; Hancock et al., 2018; Saunders et al., 2022). In addition to inherited variation, prolonged alcohol exposure may produce molecular adaptations in brain regions involved in reward, motivation, decision-making, and emotional regulation. These adaptations include changes in gene expression, epigenetic regulation, and cellular signaling pathways (Anderson & Taniguchi, 2022; Volkow et al., 2016). Consequently, studying AUD-associated gene-expression patterns can provide insights into the molecular correlates of AUD and help identify candidate pathways for further biological investigation.

Hanan Alotaibi
hmalotaibi@kau.edu.sa

[1] Information Technology Department. Faculty of Computing and Information Technology. King Abdulaziz University Jeddah, KSA

[2] Computer Science Department. Faculty of Computing and Information Technology. King Abdulaziz University Jeddah, KSA

[3] Information Systems Department. Faculty of Computing and Information Technology. King Abdulaziz University Jeddah, KSA

RNA sequencing (RNA-Seq) has become an important tool for investigating transcriptomic alterations associated with complex disorders. It enables the large-scale quantification of gene expression and supports the examination of coding transcripts, non-coding RNAs, transcript isoforms, and regulatory patterns (Chen et al., 2023; Smail & Montgomery, 2024). In AUD research, transcriptomic analyses of brain and peripheral tissues have revealed alterations in pathways related to synaptic signalling, neuroimmune responses, inflammation, and cellular metabolism(Jang et al., 2024; Shetty et al., 2024). However, gene-expression differences identified in postmortem brain tissue should be interpreted as molecular patterns associated with diagnosed AUD and prior alcohol exposure rather than as prospective indicators of future AUD development.

The analysis of RNA-Seq data presents considerable computational and methodological challenges. Transcriptomic datasets typically contain substantially more molecular features than biological samples and may be affected by correlated gene expression, technical noise, batch effects, tissue heterogeneity, and clinical or demographic confounders (Chen et al., 2023; Luecken & Theis, 2019). These characteristics increase the risk of overfitting and reduce model reliability when preprocessing, feature selection, and validation are not performed carefully. Machine-learning approaches can assist in identifying informative expression patterns within high-dimensional datasets, but their performance must be evaluated using validation procedures that minimize information leakage and account for the limited sample sizes commonly encountered in transcriptomic studies (Hwang et al., 2024; Pudjihartono et al., 2022).

Among machine-learning approaches, penalized regression methods are particularly suitable for high-dimensional biomedical data. Elastic Net logistic regression combines L1 and L2 regularization, allowing the model to reduce the influence of less informative variables while distributing predictive weight across correlated features (Friedman et al., 2010; Zou & Hastie, 2005a). Compared with more complex black-box approaches, Elastic Net also provides directly interpretable coefficients that can be used to examine how individual gene-expression features contribute to model predictions. Nevertheless, these model-derived contributions should be interpreted as predictive associations within the analysed dataset and not as evidence of causal relationships or independently validated biological biomarkers.

Previous studies have applied machine learning to substance-use research using genetic variants, clinical records, behavioural measures, and transcriptomic data. These studies include models for predicting incident opioid use disorder from administrative health records (Lo-Ciganic et al., 2020), classifiers based on candidate genetic variants (Davis et al., 2025; Donaldson et al., 2021), and transcriptomic analyses seeking molecular signatures associated with alcohol dependence and other substance use disorders (Ellis et al., 2025; Phan et al., 2024; Rosato et al., 2019). Although these studies demonstrate the potential of computational methods, much of the existing research focuses either on biological discovery or isolated model development. Comparatively less attention has been given to integrating interpretable transcriptomic predictive workflows into user-oriented research platforms that allow laboratory users to upload appropriately formatted data, execute standardized analyses, review probability-based predictions, and generate structured reports.

To address this methodological and software-oriented gap, this study presents GENO, an interpretable web-based research platform for predicting AUD status from gene-expression profiles. GENO uses publicly available bulk RNA-Seq data from postmortem brain tissue of individuals with AUD and non-AUD controls obtained from the Gene Expression Omnibus dataset GSE253155 (National Center for Biotechnology Information}, 2024; Willis et al., 2025). The analytical pipeline combines gene-expression preprocessing and feature selection with an Elastic Net logistic regression model. For each input profile, the model generates a probability-based prediction of membership in the AUD class and estimates the contribution of individual gene-expression features to that prediction. The platform integrates data upload, automated processing, probability-based prediction, gene-level contribution analysis, result visualization, and report generation within a unified workflow intended to support laboratory-oriented research.

The study contributions are as follows: the study investigates the feasibility of predicting AUD from high dimensional gene-expression profiles. Also, it applies an interpretable regularized predictive modelling approach that can accommodate correlated transcriptomic features and provide model-based gene contribution estimates. Moreover, it translates the predictive analytical workflow into a web-based proof-of-concept platform designed to improve the accessibility and usability of transcriptomic machine-learning analysis for laboratory researchers. GENO is presented as a predictive research platform rather than a clinically validated diagnostic tool or a system for prospectively predicting future AUD development.

The remainder of this paper is organized as follows. Section 2 reviews related research on transcriptomic profiling and machine-learning applications in substance use disorders. Section 3 describes the dataset, preprocessing procedures, model development, evaluation strategy, and platform implementation. Section 4 presents the experimental and system

evaluation results. Section 5 discusses the findings and limitations, and Section 6 concludes the paper and outlines directions for future work.

2. Literature Reviews

Drug and substance addiction results from the interaction of multiple factors, including environmental circumstances, developmental stage, and genetic susceptibility (Volkow et al., 2016). Addiction has been described as progressing through three repeating stages, each accompanied by changes in brain circuits governing reward, decision-making, and emotional regulation. The three stages are: binge and intoxication, withdrawal and negative affect, and preoccupation and anticipation" (Volkow et al., 2016). These stages show changes in the brain's reward and regulatory circuits, influenced by genetic and drug-related factors in addiction. The activity of genes within the brain's reward circuitry can be modified when genetic factors interact with the neurobiological impacts of repeated drug use, and this results in reinforcing patterns of compulsive drug-seeking and contributing to persistent addictive behaviour (Anderson & Taniguchi, 2022).

Studies show that substance use disorders have a moderate-to-high heritability. However, life experiences, drug exposure, and interactions between genes and the environment also influence the risk of addiction. Genetic differences have a role in determining what makes some people more vulnerable to addiction than others (Hancock et al., 2018). Moreover, the genetic basis of substance use disorders is complex. Many different genes are involved, as some of the genetic variants linked to substance use are also associated with characteristics such as impulsivity and mood regulation (Deak & Johnson, 2021). Others affect how the body metabolizes drugs, how fast a substance is processed and how long a substance stays active (Hancock et al., 2018). Genetics aside, repeated drug use itself can to biological changes in the brain. These changes include gene expression, DNA methylation, chromatin structure, and non-coding RNA activity, all of which shift within brain regions tied to reward and motivation, and these changes can raise relapse risk even long after someone stops using (Anderson & Taniguchi, 2022). Understanding these early genetic and epigenetic shifts that occur after drug exposure could give researchers usable biomarkers for addiction risk. This could help develop preventive interventions that may help with earlier detection and treatment before substance use disorder becomes entrenched (Anderson & Taniguchi, 2022; Hancock et al., 2018; Volkow et al., 2016).

Gene expression data is essential in uncovering the molecular mechanisms behind addiction (Ajmeriya et al., 2025). Next-generation sequencing (NGS) methods enable analysis at high resolution: DNA sequencing captures inherited genetic variation, and RNA-Seq captures dynamic transcriptional changes (Jang et al., 2024). RNA-Seq in particular has become a fundamental approach for studying gene expression. It identifies and quantifies transcripts, enables isoform-level analysis, and characterises non-coding RNA. For this, statistical methods were developed specifically for its discrete count data rather than the continuous intensities of earlier microarray platforms (Chen et al., 2023). Recent studies using both bulk and single-cell transcriptomics in brain and peripheral tissues show consistent changes in gene expression following exposure to alcohol, drugs, and stimulants (Jang et al., 2024). These changes often affect pathways related to synaptic signalling, immune and inflammatory responses, and cellular metabolism, providing insight into underlying biological mechanisms and identifying potential peripheral biomarkers (Jang et al., 2024).

Longitudinal studies that examine groups across different doses and experimental sequences can detect dose-specific and context-dependent changes in gene expression related to exposure (Shetty et al., 2024). Moreover, applying RNA-Seq to peripheral blood may help identify potential molecular markers in accessible tissues, offering a less invasive alternative to sampling central tissues (Shetty et al., 2024). High-quality RNA-Seq data provide the most reliable insights, particularly when collected over time in longitudinal studies (Shetty et al., 2024). It can be challenging to analyse RNA-Seq data, whether from bulk or single-cell sources, because of the complexity of the data, as datasets have very large numbers of genes which may be influenced by noise and technical error (Luecken & Theis, 2019). Also, variation can occur due to different experimental conditions (batch effects) and, in single-cell data, from heterogeneity in cell types within a sample (Luecken & Theis, 2019). Thus, the analytical method used can strongly influence the results, which makes transparent, well-documented pipelines and validation on independent datasets essential (Chen et al., 2023; Luecken & Theis, 2019). These approaches ensure that observed changes in gene expression, such as those triggered by drug treatment, can be translated into dependable markers for early detection and prevention (Jang et al., 2024; Shetty et al., 2024). Using traditional statistical methods on RNA-Seq data can face challenges with the results, particularly pronounced in high-dimensional, low-sample-size datasets (Chen et al., 2023). Also, genes may interact in non-linear and complex ways that traditional statistical methods capture poorly (Chen et al. 2023; Luecken and Theis 2019). Thus, alternative methods based on machine learning (ML) and deep learning (DL) have been introduced to address these (Chen et al., 2023). Such approaches are

applied not only for prediction but also for feature selection, narrowing high-dimensional data before modelling. Over the past few years, there has been an increase in the use of machine learning (ML) approaches on high-dimensional genomic data to identify patterns relevant to complex disease risk, including substance use disorders (Pudjihartono et al., 2022). Classical ML algorithms such as support vector machines, random forests, K-nearest neighbours, and artificial neural networks are commonly used to illustrate the relation between genetic features and phenotypes (Pudjihartono et al., 2022). Tree-based algorithms, including random forests and gradient boosting, are categorised as embedded methods because feature selection is integrated directly into model training, with feature importance ranked using metrics such as Mean Decrease Impurity (Pudjihartono et al., 2022). Penalised regression methods such as LASSO and elastic net work with linear classifiers by shrinking the coefficients of features that do not meaningfully contribute to the model, thereby reducing high-dimensional input to a smaller, more informative feature set prior to downstream prediction (Pudjihartono et al., 2022)

Deep learning has the ability to identify complex patterns within gene expression data, including interactions that traditional statistical tests overlook (Erfanian et al., 2023; Hwang et al., 2024). RNA-seq data in its various forms, which are bulk, single-cell, and spatial, have become the main foundation for these deep learning models (Erfanian et al., 2023; Smail & Montgomery, 2024). Convolutional neural networks, autoencoders, and transformer techniques are applied to compress, reconstruct, and re-represent transcriptomic data in ways that maintain biologically meaningful structure (Erfanian et al., 2023; Hwang et al., 2024; Smail & Montgomery, 2024). This carries great weight in addiction and substance use research, where the neurobiological signals tend to be subtle and scattered across many genes rather than concentrated in just a few (Bogdan et al., 2023). Building a reliable genomic AI pipeline for complex disease research is far from straightforward, as it involves a series of decisions, and getting any one of them wrong can compromise the results. The process typically starts with quality filtering of the data through normalization and correcting for batch effects in RNA-seq inputs. Then, features that do not carry meaningful information are removed (Smail & Montgomery, 2024). Filtering data at this initial stage is crucial to avoid generating misleading results.

Consequently, prioritizing stability must be treated as the primary objective during the process of feature selection. This implies running the process repeatedly with bootstrap samples or permutation schemes, and checking that the same features keep showing up rather than trusting a single pass (Pudjihartono et al., 2022). This issue is critical in high-dimensional genomic datasets where the number of features can exceed the number of available samples, and that can raise the risk of encountering misleading associations that are hard to detect. Model training is carried out within nested cross-validation loops, with external replication used to detect overfitting that internal validation alone can miss (Hwang et al., 2024; Pudjihartono et al., 2022). Once the model is developed, methods such as SHAP values, LIME explanations, and pathway enrichment analyses are applied to examine what has actually been learned, helping distinguish genuine biological signal from statistical noise (Hwang et al., 2024). Following these steps carefully can yield genetic or transcriptomic signatures that effectively help us categorise individuals based on their disease risk, recognise early molecular responses to exposures, and highlight biomarkers with real potential for clinical use (Hwang et al., 2024; Pudjihartono et al., 2022).

Despite this potential, there are many challenges to using metadata fragmentation in public repositories, as the lack of consistent benchmarks inhibits progress; clinical-grade applications require rigorous external validation and transparent model explanations. Pudjihartono et al. conducted a study that highlights the benefits of using ML-based approaches (Pudjihartono et al., 2022). They found ML-based approaches can capture complex SNP-SNP interactions that standard genome-wide association approaches typically miss. Moreover, the study emphasises the value of a ensemble approach for enhancing predictive performance in genetic studies, demonstrating that ensemble feature selection methods, which combine outputs from multiple approaches rather than relying on a single technique, are effective.

In addition, RNA sequencing has emerged as a powerful approach for discovering biomarkers across many clinical and disease contexts (Pudjihartono et al., 2022). The transcriptome profiling approach is uniquely sensitive to both genetic and environmental determinants of gene regulation, as it captures more than which genetic variants are present. It shows how they are expressed across different tissues and cellular environments (Smail & Montgomery, 2024). This functional dimension of genomic data uncovers disease progression trends, subtypes, and treatment responses that standard genetic data alone may fail to catch(Smail & Montgomery, 2024). Additionally, when we combine different types of data, we often end up with models that are more reliable and effective than those that rely on just one source. For instance, Hwang et al. study demonstrated that multimodal deep learning models combining genomic, transcriptomic, and proteomic data deliver superior output than single-modality models in disease classification and biomarker identification

(Hwang et al., 2024). These results highlight how research is moving away from isolated analyses and transitioning toward models that connect different datasets and methodologies. For these models to be clinically relevant, they must demonstrate accuracy, replicate across diverse groups, identify biologically plausible pathways, and account for confounding factors like population differences and coexisting health conditions.

2.1. Machine Learning and Gene Expression Analysis in Addiction Research.

Over the past decade, combining transcriptomics with machine learning has improved the analysis of noisy RNA-seq data. In addiction research, these methods help identify gene expression patterns linked to substance use, support early risk detection, and highlight candidate genes associated with addictive behaviours.

Lo-Ciganic et al. found that ML approaches are effective in predicting opioid use disorder (OUD) (Lo-Ciganic et al., 2020). To predict incident OUD, they designed and assessed algorithms using four popular machine-learning approaches, namely: elastic net, random forests, gradient boosting machine, and deep neural network. The study revealed that machine learning significantly exceeded existing Centres for Medicare & Medicaid Services (CMS) opioid safety measures, which missed over 70% of incident OUD cases. The results demonstrate that structured clinical data can be used to predict OUD risk across large populations. Adding biological or molecular indicators could further strengthen personalized risk prediction by providing information not captured in prescription and diagnosis data.

A study conducted by Sullivan et al. identified key gene modules and potential therapeutic targets associated with opioid use disorder (OUD) (Sullivan et al., 2025). They used a network-based machine-learning approach and integrated multiple types of biological data, including GWAS, transcriptomic, proteomic, and epigenomic datasets from the dorsolateral prefrontal cortex of individuals who died from opioid overdose. Starting with 404 candidate genes identified across several datasets, their Gene set Refinement through Interacting Networks (GRIN) method narrowed the list to 211 highly connected genes involved in important signaling pathways such as Akt, BDNF, and ERK/MAPK. From there, the researchers built a drug–gene interaction network linking 48 of these genes to 414 approved or investigational drugs. The network already included several existing OUD and alcohol use disorder treatments, such as buprenorphine, methadone, naltrexone, acamprosate, and topiramate, suggesting real potential for repurposing these drugs for new uses (Sullivan et al., 2025). These results demonstrate how multi-omics analyses can guide future laboratory research.

Another study, with finer resolution, conducted by Phan et al. identified cell-type-specific transcriptional alterations in opioid use disorder (OUD) using single-nucleus RNA-seq across human and non-human primate striata (Phan et al., 2024). The study analyzed approximately 98,848 nuclei collected from the caudate and putamen of 12 individuals, enabling the identification of major neuronal populations as well as less common cell types, including interneurons. To identify cell type-specific regulatory programs, they used the SCENIC (pySCENIC) machine learning framework. This includes GRNBoost2, which infers transcription factor-gene relationships, followed by cisTarget, which filters for direct targets based on promoter binding motifs. AUCell then scores the activity of these regulatory modules in individual cells. The findings revealed significant changes in gene expression related to neurodegeneration, interferon signalling, and DNA damage within striatal neurons from individuals with OUD (Phan et al., 2024) .

A study by Rosato et al. provides valuable evidence that peripheral molecular markers may help identify substance use disorders, although the sample size was relatively small (Rosato et al., 2019). The researchers aim to discover whether salivary microRNAs (miRNAs) could distinguish individuals with alcohol dependence from healthy controls through small-RNA sequencing from saliva samples. Rosato et al. used Random Forest, a machine-learning model. The study included 56 African American and 64 European American participants and identified several miRNAs that were differentially expressed between groups. Three miRNAs, miR-10a-5p, miR-1290, and miR-4488, showed consistent expression changes across both populations. Relying on the five most informative miRNAs identified through the Gini index, the best-performing model reached classification accuracies of 79.1% in African Americans and 72.2% in European Americans. Salivary miRNAs may carry clinically meaningful signals tied to alcohol dependence, though the expression differences observed here didn't hold up after correcting for multiple testing(Rosato et al., 2019). Some studies have utilized human transcriptomic data with direct experimental work to back up their claims about candidate genes. For instance, a study by Ellis et al. employed machine learning to analyze RNA-sequencing data from the orbitofrontal cortex (OFC) of human heroin users alongside control subjects (Ellis et al., 2025). They identified the gene SHISA7 as the top predictor of heroin use, and researchers overexpressed

Shisa7 in the OFC of rats to validate its role. The analysis of proteomic and transcriptomic data reveals that Shisa7 overexpression in rats led to gene expression and protein interactions similar to those of people who use heroin, especially the effects on neurodegenerative and neuroimmune pathways. The study models achieved predictive performance, with an AUROC of 1.0 using Gini and Shapley methods and 0.912 via the ANOVA F-test. This indicates that OFC Shisa7 is an essential factor in OUD research (Ellis et al., 2025).

Advanced AI and ML methods are drawing attention these days, but classical statistical approaches still hold their place in addiction research, as predictive models have an effective statistical foundation; precision and specificity matter, and so does careful thought about the study population. Also, the quality of the biological data used to train a model is essential. An important finding in biomarker research is that recent work using spatial transcriptomics and genotype-informed approaches shows that molecular responses to opioid exposure can vary substantially across brain regions and cell types (Xie et al., 2025). Xie et al. (2025) for instance, found that gene-expression changes in Oprm1 A118G mice depended on both the genetic variant involved and the specific cell type examined. Genetic background and cellular context clearly matter when identifying biomarkers, since accounting for them helps researchers zero in on the targets most likely to hold up under further experimental testing.

Preclinical studies have employed ML and systems pipelines to sort candidate genes from RNA-seq that was included in cocaine self-administration models, and they aim to combine differential gene expression with conservation, brain-specific coding, and cross-dataset stability to recommend promising therapeutic targets (Vannan et al., 2023). Bubier et al. used the Gene Weaver platform to pull together a wide range of addiction genomics resources, QTLs, expression studies, sequence variation, and functional screens, and applied convergent analysis to narrow a large QTL interval down to a single strong candidate, Rab3b, which they then validated behaviourally in knockout mice (Bubier et al., 2020). Their work shows how aggregating heterogeneous data can directly guide mechanistic follow-up in the lab. A similar method was applied effectively to a cocaine addiction study where researchers first identified and refined candidate genes through a systems-level analysis, and then confidently confirmed their functions in living organisms. Engeln et al. profiled the mouse ventral pallidum (VP) transcriptome after cocaine self-administration, combining differential expression analysis with transcriptional-regulator prediction and projection-specific RiboTag sequencing (Engeln et al., 2022). This method revealed a selective upregulation of Nr4a1 (Nurr77) in VP neurons projecting to the mediodorsal thalamus (MDT). Together, these studies provide evidence that combining transcriptomic data with predictive frameworks could give us a better shot at identifying people vulnerable to substance use disorders before the condition takes hold.

Overall, the existing literature emphasizes that machine learning and AI are powerful tools for uncovering the molecular mechanisms behind substance abuse across different tissues and drug types. Building on these findings, this study investigates the use of machine learning to evaluate addiction risk via gene expression profiles. A core objective of this study is to translate these capabilities into an interactive, lab-oriented web platform that optimize predictive accuracy. Despite these encouraging results, several limitations should be considered. Many reported performance measures are derived from relatively small post-mortem datasets, which may increase the risk of overfitting and limit generalizability. In addition, some of the highest AUROC values reported in the literature are related to network reconstruction tasks rather than direct disease classification. Other studies have identified potentially relevant molecular signatures but were unable to develop classifiers with sufficient predictive performance. Furthermore, only a limited number of biomarkers have been validated in independent cohorts or assessed in accessible biofluids, which remains a critical requirement for future diagnostic applications.

3. Methodology

This section presents the methodological procedures followed to develop the proposed GENO system. The methodology includes several key stages: dataset acquisition and preparation, data preprocessing and splitting, model architecture, model training, and performance evaluation.

3.1. Dataset

GENO uses the gene expression dataset obtained from the Gene Expression Omnibus repository, which is a public database hosted by the National Center for Biotechnology Information (NCBI) [22]. The original dataset contained 194 samples. After matching the gene expression data with the corresponding metadata using the sample IDs, 114 matched samples were retained for analysis. The resulting dataset comprised samples from 114 individuals, including 58 individuals with alcohol use disorder classified as high-risk, and 56 individuals with non-alcohol use disorder classified as low-risk. Each sample was assigned a binary class labelled as 0 for low addiction risk and 1 for high addiction risk. Each input sample consists of a set of gene expression features

arranged in a consistent gene order to ensure uniformity during both model training and inference

3.2. Data Preprocessing and Splitting

To prepare the raw gene expression data for machine learning, a preprocessing pipeline was employed to improve data quality, reduce dimensionality, and enhance classification performance. The pipeline comprised sequential stages, as illustrated in Figure 1. First, gene expression values were normalized using standard scaling to ensure that all gene expression features were represented on a comparable numerical scale. This step is important for regularized linear models because coefficients can otherwise be biased toward variables with larger numerical ranges. Second, feature selection was performed to reduce the dimensionality of the gene expression matrix and retain genes with stronger discriminatory power. Two methods were applied: the ANOVA F-test to identify features showing significant between-class variation, and Mutual Information to
capture potentially non-linear dependencies between individual genes and the addiction-risk label. Genes retained by these selection procedures were merged into a consensus feature set, producing the final preprocessed feature matrix for model training. The dataset was randomly split into a training set (70%) and a hold-out test set (30%). The training set was used
for feature selection and model fitting, while the test set was reserved for final performance evaluation. All preprocessing steps were first fitted using the training set and subsequently applied to the hold-out test set. Specifically, the scaling parameters were estimated from the training set, and feature selection was performed using only the training data. The same scaling parameters and selected feature set were subsequently applied to the test set. This procedure helped prevent information leakage and provided a more reliable assessment of model performance.

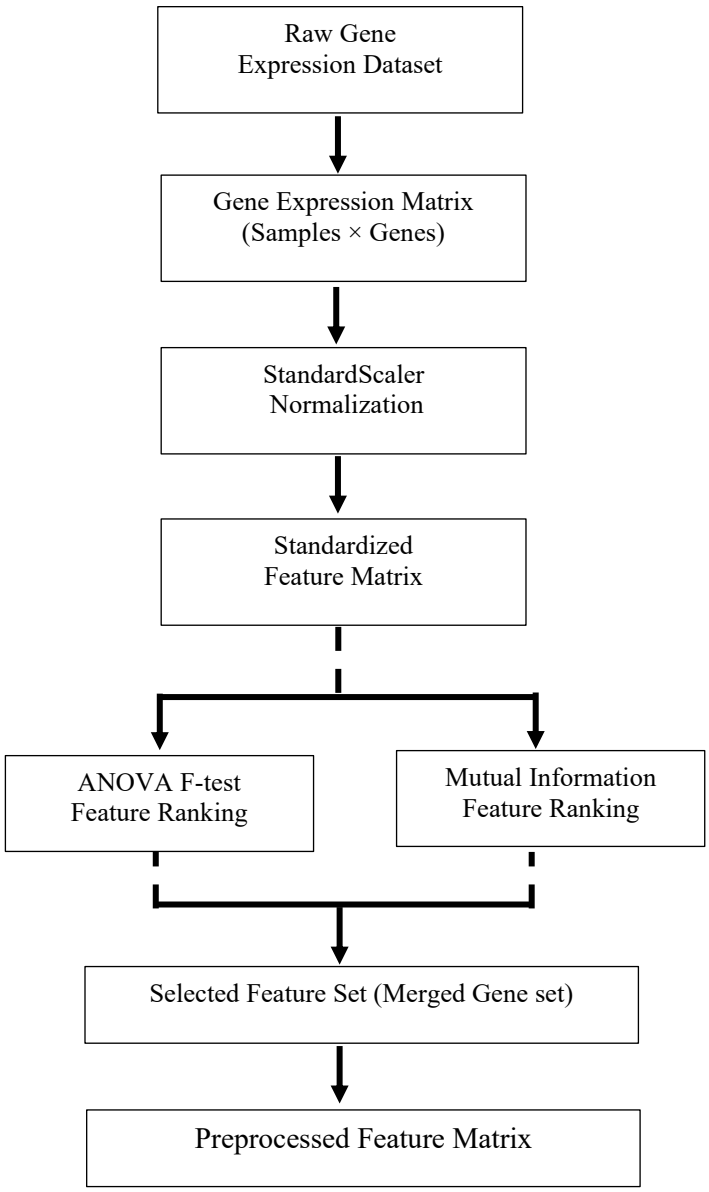


Figure 1 : Data preprocessing pipeline

3.3. Model Architecture

The GENO system uses an Elastic Net Logistic Regression model as its primary AI component to predict addiction risk from gene expression profiles. Figure 2. shows the architecture of the proposed Elastic Net Logistic Regression Classifier. This architecture was selected for its ability to handle high-dimensional genomic data and correlated feature spaces (Zou & Hastie, 2005a) and its use of both L1 and L2 regularizations to control the model complexity. The L1 regularization promotes sparsity by shrinking less informative gene coefficients toward zero, facilitating embedded feature selection, whereas the L2 regularization allocates weight across correlated predictors and enhances coefficient stability, thereby minimizing overfitting (Friedman et al., 2010; Hoerl & Kennard, 1970; Tibshirani, 1996; Zou & Hastie, 2005b). Additionally, the logistic function in. the model outputs a continuous probability score, allowing the GENO system to present addiction risk as a percentage, which is interpretable for laboratory professionals

The Elastic Net regularization penalty is defined as:

$$\lambda(\alpha \sum_{i=1}^{n} |\beta_i| + \frac{1-\alpha}{2} \sum_{i=1}^{n} \beta_i^2) \qquad (1)$$

where λ controls the overall regularization strength and α determines the balance between L1 and L2 regularization.

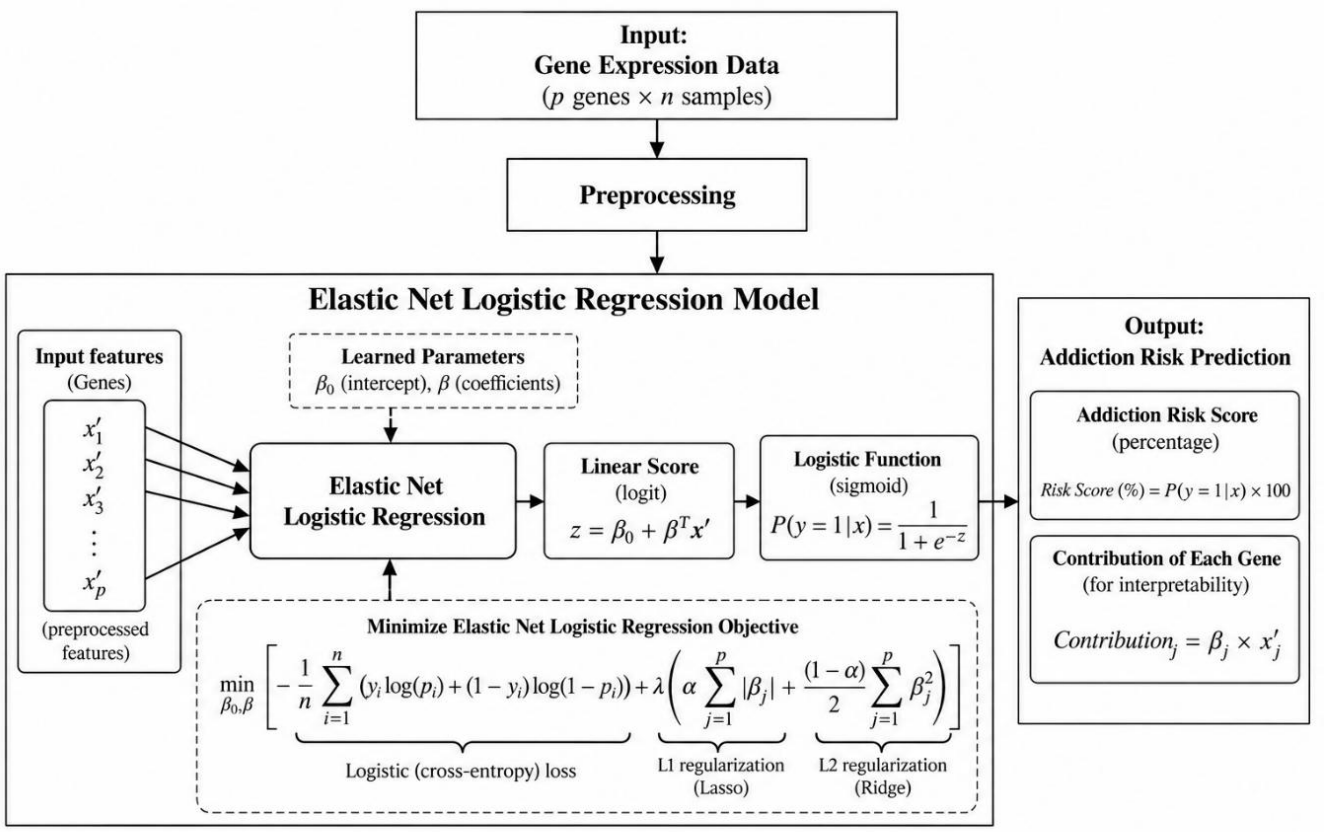


Figure 2 :Architecture of the Proposed Elastic Net Logistic Regression Classifier

The optimization objective minimized during model training can be expressed as:

$$minimize\ \{ L(\beta) + \lambda(\alpha \sum_{i=1}^{n} |\beta_i| + \frac{1-\alpha}{2} \sum_{i=1}^{n} \beta_i^2)\} \quad (2)$$

where L(β) represents the logistic loss function. As α nears 1, the model behaves like Lasso regression, while as α approaches 0, it resembles Ridge regression. In GENO, this balance helps retain biologically informative genes and reduces instability caused by correlated gene expression patterns. In GENO, the sample probability is calculated using equations (3) and (4). The probability of class 1

$$P(y = 1/X) = \frac{e^{\beta_0 + \beta^T x}}{1 + e^{\beta_0 + \beta^T x}} \quad (3)$$

The Probability of class 0

$$P(y = 0/X) = \frac{1}{1 + e^{\beta_0 + \beta^T x}} \quad (4)$$

During inference, the model estimates the probability of addiction risk.

$$\text{Risk Score} = P(y = 1 \mid x) \times 100 \quad (5)$$

To enhance result interpretation, the system calculates each gene's contribution to the final prediction by scaling gene expression values with the model's coefficients (coefficient times scaled expression), as shown in equation (6).

$$Contribution_j = \beta_i * x'_j \quad (6)$$

3.4. Model Training

The classifier model was configured using Elastic Net regularization with the SAGA optimization solver. The regularization balance was controlled by setting the L1 ratio to 0.5, while the inverse regularization strength was set to 0.7. To ensure convergence during optimization, the maximum number of iterations was set to 5000. In addition, the classifier employed balanced class weights to account for potential class imbalance, and a fixed random state of 42 was used to ensure reproducibility of the training process. The key training parameters are summarized in Table 1.

Because the model was trained to make predictions based on the absolute position of an input feature rather than its column name, any variation in feature order can

lead to inconsistent results. Therefore, to ensure prediction consistency during inference, precise alignment of gene features between the training and input data was maintained by storing the feature order used during model training. Incoming input data were dynamically reindexed to conform to the exact gene order expected by the model. This ensures that all the features align perfectly with the training configuration, delivering stability and accuracy. This behaviour is standard in machine learning. g models that process numerical feature arrays.

Table 1: Key Training parameters

Parameter	Value
Penalty	Elastic Net
L1_ratio	0.5
Solver	Saga
Inverse regularization strength	0.7
Max iterations	5000
Random state	42
Class weight	Balanced

3.5. Model Evaluation

Performance was assessed on a separate test set consisting of 30% of the data, using standard classification metrics. Accuracy was the primary summary statistic, indicating the proportion of correct predictions across all test samples.

$$Accuracy = \frac{Correct\ predications}{Total\ predictions} \quad (7)$$

In biomedical datasets, class imbalance often leads to an overestimation of model performance when one class dominates. To address this issue, balanced accuracy was used to ensure equal consideration of both addiction-risk classes.

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (8)$$

Precision, recall, and the F1-score were used to assess the balance between accurately detecting positive cases and missing positives. Precision indicates the percentage of high-risk samples correctly identified as positive, while recall (or sensitivity) shows how many actual high-risk samples are detected. The F1-score, which is the harmonic mean of precision and recall, is particularly valuable when the positive class is the main concern (Chicco & Jurman, 2020; Grandini et al., 2020).

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (9)$$

Discriminative performance was evaluated using the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (ROC-AUC). ROC-AUC measures how well the model distinguishes between positive and negative classes across all thresholds. An AUC of 0.5 indicates random guessing, while 1.0 indicates perfect discrimination. This metric is valuable for binary classification because it assesses the model's discriminative power regardless of the chosen threshold.

$0 < AUC < 1$

AUC = 0.5 indicates no discrimination.

AUC = 1 indicates perfect discrimination between classes.

4. Experimental Results

To evaluate the GEMO model's performance, we used a 30 % hold-out test set from the gene expression dataset. Table 2 summarises the model's overall performance in terms of accuracy, balanced accuracy, F1 score, and AUC metrics. All threshold-dependent metrics were computed at the default decision threshold of 0.5, with the AUD group (Class 1) treated as the positive class, and F1-score reported as the macro-average across the two classes.

Table 2.Overall predictive performance metrics results

Metric	Value
Accuracy	86%
Balanced accuracy	0.865
F1 Score	0.865
AUC	0.895

The model yielded an accuracy of 86%, indicating correct addiction risk classification for the majority of the evaluated samples. Balanced accuracy, given in Equation (8) as the mean of sensitivity and specificity, was 0.865. This value is close to the overall accuracy because the test set was nearly balanced, which indicates that the accuracy figure was not driven by the prevalence of one class. The F1-score of 0.865 indicates a successful balance between precision and recall across the evaluated samples. The system scored an AUC of 0.895, reflecting a strong discriminative capacity to distinguish between addiction-risk classes.

Table 3 demonstrates that the proposed GENO model achieved a highly consistent performance across the class-specific metrics for the non-AUD group (Class 0) and the AUD group (Class 1). For precision, the model achieved scores of 0.84 for Class 0 and 0.89 for Class 1. For recall, the model scored 0.89 for Class 0 and 0.84 for Class 1. Both classes obtained the same F1-score (0.86), which indicates that the proposed model does not favour one case over another. The GENO model also produces continuous much each edge contributed to a

Table 3. Class-specific predictive metrics for addiction risk classification results

Class	Precision	Recall	F1 Score
0	0.84	0.89	0.86
1	0.89	0.84	0.86

given prediction, following Equations (5) and (6). These outputs allow a prediction to be examined in detail rather than reported only as a class label.

To examine the behaviour of the model across decision thresholds, precision–recall and confidence-based performance curves were analysed, as shown in Figure 3. The model achieves an average precision of 0.912 across all classes, indicating that the system retains a high positive predictive value across different sensitivity levels. The metrics reported in Tables 2 and 3 correspond to the default decision threshold of 0.5; the confidence curves, by contrast, report the pooled one-vs-rest F1-score computed at each candidate threshold value. At a threshold of 0.292, the pooled F1-score is 0.79, with a precision of 0.71 and a recall of 0.89 across all classes, as illustrated in Figure 3. These evaluation curves confirm the stability of the Elastic Net logistic regression architecture in handling the genomic input space.

In addition to the predictive model, GENO was developed as a web-based platform with an interactive dashboard to support laboratory users in managing gene expression analyses and reviewing their results. The dashboard provides an overview of previous analyses together with their corresponding risk levels. For each analysis, the generated report includes the predicted risk of alcohol use disorder, the associated risk score, and relevant gene expression information.

The platform was developed primarily in Python, with Flask used for backend operations and PostgreSQL for data storage. The user interface was implemented using the React JavaScript library. Visual Studio Code served as the main development environment, while Google Colab was used to facilitate real-time collaboration among team members. During implementation, the core functionality of GENO was tested, including connectivity between the frontend, backend, and database components. Examples of the implemented interface and dashboard are presented in Figure 4.

The usability of the GENO interface and dashboard was evaluated with six laboratory technicians from the Genome Excellence Centre at King Abdulaziz University. Participants were asked to complete three main tasks: logging into the system, uploading a gene expression profile, and viewing and downloading the

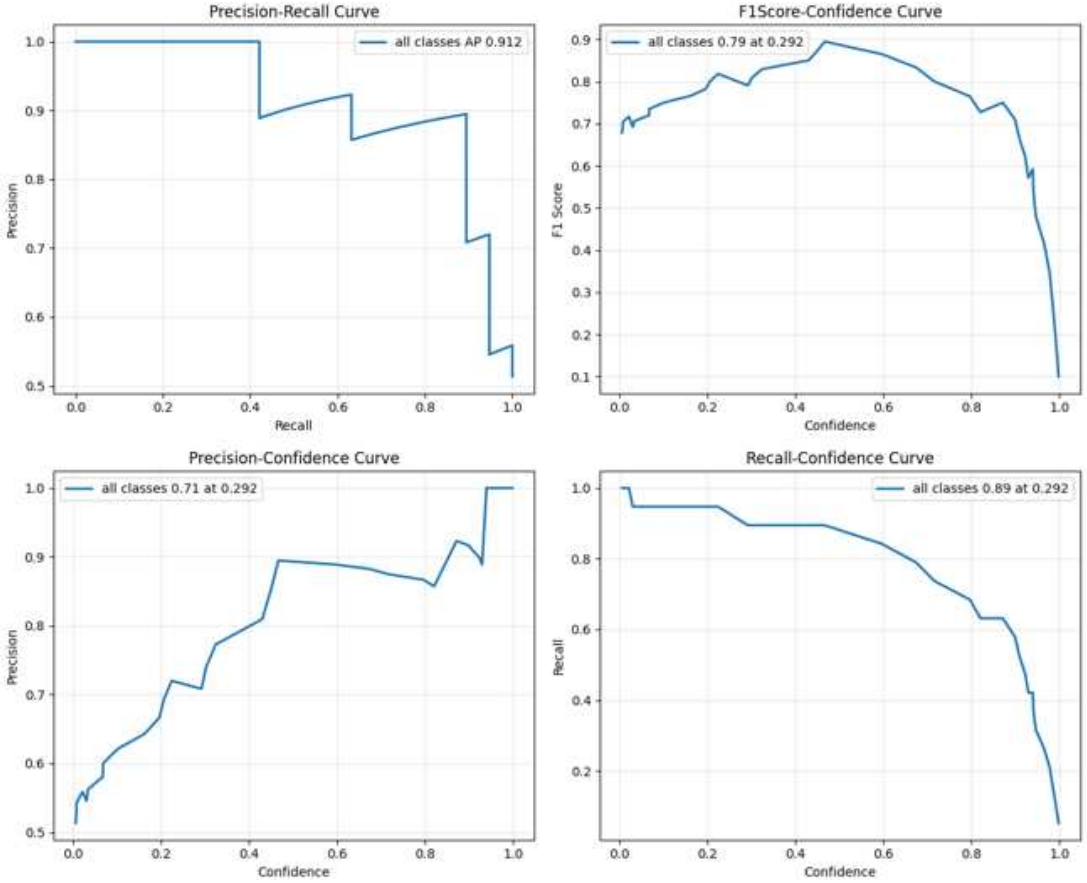


Figure 3: Performance evaluation of the GENO classification model: precision–recall, F1-score–confidence, precision–confidence, and recall–confidence curves.

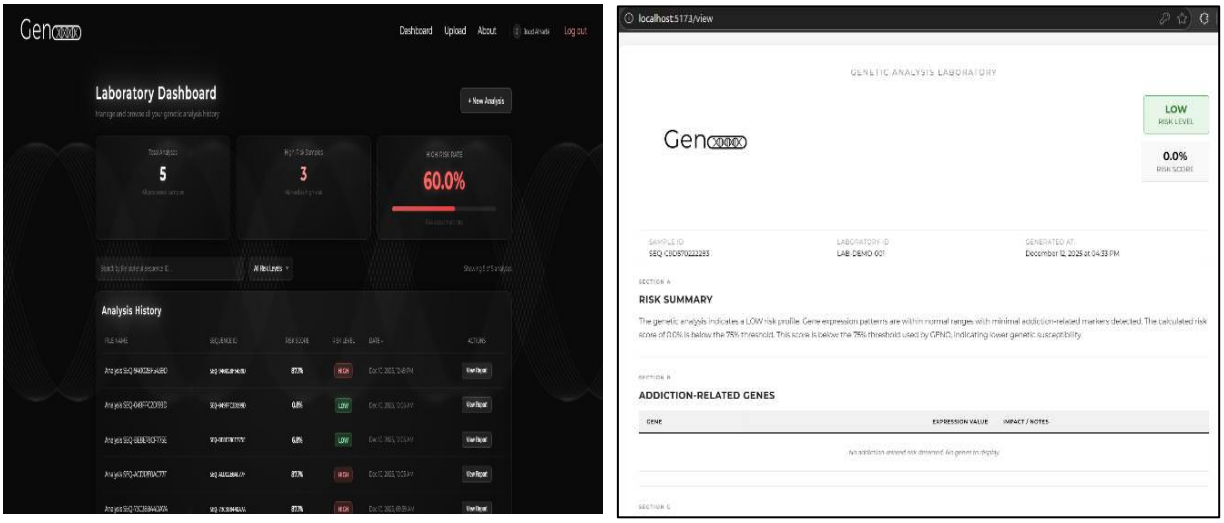


Figure 4. GENO web-based platform interfaces: (a) laboratory dashboard and analysis history; (b) generated addiction-risk analysis report.

generated report. All participants completed the assigned tasks successfully and without errors. The mean completion times were 14.5 seconds for logging in, 19 seconds for uploading the gene expression profile, and 16.5 seconds for viewing and downloading the report.

After completing the tasks, participants were invited to provide qualitative feedback on their experience with the platform. They were asked whether any part of the process was confusing and whether the system was clear, easy to use, and straightforward to navigate. Participants indicated that the tasks were easy to understand and that navigation through the platform was smooth. The overall user experience received a rating of 10/10. The visual design was rated 8/10, with participants suggesting the addition of a light mode and a shortcut button providing direct access to the control panel.

5. Discussion and Limitations

The findings of this study support the feasibility of combining transcriptomic data with an interpretable machine-learning approach for addiction-risk classification. The suitability of Elastic Net for GENO is consistent with existing research such as Pudjihartono et al., (Pudjihartono et al., 2022), who have similarly demonstrated the value of regularization and feature-selection methods for reducing dimensionality and improving model interpretability. By combining L1 and L2 regularization, Elastic Net controls model complexity and reduces the influence of less informative features while improving stability among correlated predictors. One of the key advantages of the GENO is that its analysis is not limited to categorical predictions. The model provides a probability-based risk score for addiction risk and identifies the genes that contribute most significantly to the prediction. These results offer laboratory users additional insights into the model's predictions, rather than simply presenting the expected risk category. Accordingly, the gene contributions generated by the GENO should be considered model-based indicators of predictive effect rather than biological biomarkers.

To contextualize the proposed platform, a comparison study was conducted to validate the GENO with similar studies related to addiction risk prediction and genetic analysis in term of methodology and performance. The selection criteria include (1) prediction or identifying addiction risk using computational approaches, (2) leveraging machine learning to improve the understanding of substance use disorder. Donaldson et al. carried an experiment to predict the risk of opioid use disorder (OUD) based on a set of 15 single nucleotide polymorphisms (SNPs). The finding showed a sensitivity of 82.5% and a specificity of 79.9%, indicating high performance (Donaldson et al., 2021). However, a subsequent large-scale evaluation of a set of 15 SNPs in another study showed an accuracy of only 52.8%, raising concerns about the generalizability of prediction based on a limited number of SNPs (Davis et al., 2025). Saunders et al. managed to identify around 4,000 genetic variants associated with tobacco and alcohol use. The study showcased the complexity genetics related to substance use behaviors without developing an individual-level risk classifier (Saunders et al, 2022*)*. Yi Xu1 et al. revealed that when combining genetic indicators with other variables, such as demographic, will positively improve predictive

performance. This emphasizes the importance of integrating multiple data sources rather than using genetic information alone, highlighting the importance of integrating multiple data sources (Saunders et al, 2022).

When comparing the forementioned studies to GENO, the platform achieved 86% accuracy and an area under the curve (ROC) of 0.89 using gene expression data and logistic regression. Contrasting to SNP-based approaches that rely on a predetermined set of markers, GENO analyses genome-wide gene expression patterns delivering a comprehensive molecular profile. Furthermore, beyond the predictive model, one of GENO's practical contributions is the integration of the analytical workflow into a web-based platform designed for laboratory use. Most current work in addiction genomics focuses solely on identifying molecular correlations or developing computational models without considering providing practical and user-friendly system to translate analytical processes. Fortunately, GENO addresses this gap by integrating the workflow cycle starting from data submission to report generation as illustrated in Figure 5. This integration represents an initial step toward translating genomic machine-learning techniques from computational experiments into tools that support laboratory-oriented analysis. A brief summary of related studies and comparison are presented in Table 4.

Several limitations should be considered when interpreting the current GENO application. First, the system currently supports gene expression data only in CSV format and does not support other commonly used genomic data formats. Additionally, the analysis relies on a limited set of gene expression features, which are adequate for predicting addiction risk. However, they may not be capable of capturing more complex gene expression patterns and biological interactions related to substance addiction. The current model is also limited to alcohol-related addiction and handles gene expression information without incorporating other factors, for example demographic, clinical, behavioural, or environmental. Consequently, focusing on one factor might marginally impair the overall risk assessment. Finally, the platform currently relies on standard credential authentication and does not implement advanced security mechanisms such as multi-factor authentication (MFA) or role-based access control (RBAC). These limitations promote opportunities to improve the model, data inputs, and system capabilities in future developments.

Several limitations should be considered when interpreting the current GENO application. First, the system currently supports gene expression data only in CSV format and does not support other commonly used genomic data formats. Additionally, the analysis relies on a limited set of gene expression features, which are adequate for predicting addiction risk. However, they may not be capable of capturing more complex gene expression patterns and biological interactions related to substance addiction. The current model is also limited to alcohol-related addiction and handles gene expression information without incorporating other factors, for example demographic, clinical, behavioural, or environmental. Consequently, focusing on one factor might marginally impair the overall risk assessment. Finally, the platform currently relies on standard credential authentication and does not implement advanced security mechanisms such as multi-factor authentication (MFA) or role-based access control (RBAC). These limitations promote opportunities to improve the model, data inputs, and system capabilities in future developments.

6. Conclusion And Future Work

This paper introduced GENO, an AI-based system that integrates gene-expression analysis with Elastic Net Logistic Regression to support addiction-risk assessment. The results demonstrate the feasibility of using gene expression data for addiction risk classification while providing a user-friendly workflow for laboratory users that translates machine learning-

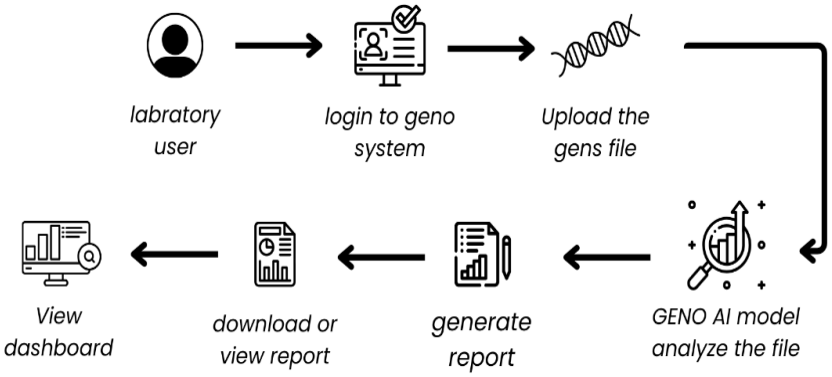


Figure 5: GENO web-based workflow for gene-expression-based addiction risk analysis.

Table 4: Comparison of GENO with Existing Addiction-Risk and Genetic Studies

Study	Core Focus	Methodology/Data Used	Key Findings & Performance	Limitation
(Donaldson et al., 2021)	Validation of an Opioid Use Disorder (OUD) risk classifier.	ML applied to 15 specific single nucleotide polymorphisms (SNPs) in 385 patients	Claimed high efficacy: 82.5% sensitivity and 79.9% specificity in predicting OUD risk	Highly criticized for a small sample size that failed real-world application
(Davis et al., 2025)	Independent audit and critique of the FDA-approved OUD test validated in study	Large-scale replication study testing the 15-SNP panel on 453,664 veterans	The test is ineffective with an accuracy of 52.8%	Confirms that a restricted 15-SNP panel cannot reliably predict a complex disorder
(Saunders et al., 2022)	Broad discovery of genetic variants linked to tobacco and alcohol use	Massive multi ancestry genome-wide association Study (GWAS) of 3.4 million people	Identified nearly 4,000 genetic variants associated with substance use behaviours across diverse populations	Highlights that addiction genetics are highly complex and spread across thousands of genes not just a few
(Yi Xu1 et al., 2020)	Predicting smoking behaviour using genetic variant and ML	ML models applied to SNPs and demographic variables	Highest predictive accuracy can be achieved by combining genetic markers with demographic variables	Genetic data alone had lower predictive power, environmental and behavioural factors were crucial
GENO (Proposed Platform)	Web platform for automated, efficient detection of addiction-related genetic markers in laboratory samples	Elastic net logistic regression model trained on gene-expression data	Achieved 86% accuracy in expression analysis with an ROC-AUC score of 0.89	Currently restricted to genetic expression data. It lacks integration of behavioural and environmental variables

based gene-expression analysis into practical laboratory applications. The current application provides a foundation for further development rather than a complete solution for addiction risk assessment. In particular, the results highlight the importance of considering addiction as a complex condition that cannot be represented by gene-expression information alone. Therefore, future development should broaden the model's scope beyond its current focus on alcohol-related addiction and incorporate larger and more diverse datasets to improve its predictive power and applicability.

For future development, a comprehensive assessment will be conducted on GENO platform to integrate other contributing factors such as demographic, clinical, behavioural and environmental. These factors will positively influence the prediction accuracy for risk addiction. Furthermore, it is essential to explore various file formats other than CSV and support complex data structure like JASON. It is worth to improve scalability and compatibility of the proposed platform by deploying secure cloud-based to facilitate laboratory workflows securely. These enhancements could strengthen GENO as a research platform for investigating addiction risks and support the ongoing development of interpretable AI methods for gene-expression analysis.

7. References

Ajmeriya, S., Chatterjee, B., & Karmakar, S. (2025). Identification of variants from gene expression data of opioid-addicted patients. *Journal of Applied Genetics*. https://doi.org/10.1007/s13353-025-00989-x

Anderson, E. M., & Taniguchi, M. (2022). Epigenetic effects of addictive drugs in the nucleus accumbens.

Frontiers in Molecular Neuroscience, *15*, 828055. https://www.frontiersin.org/journals/molecular-neuroscience/articles/10.3389/fnmol.2022.828055/full

Bogdan, R., Hatoum, A. S., Johnson, E. C., & Agrawal, A. (2023). The Genetically Informed Neurobiology of Addiction (GINA) model. *Nature Reviews Neuroscience*, *24*(1), 40–57. https://doi.org/10.1038/s41583-022-00656-8

Bubier, J. A., Philip, V. M., Dickson, P. E., Mittleman, G., & Chesler, E. J. (2020). Discovery of a role for Rab3b in habituation and cocaine induced locomotor activation in mice using heterogeneous functional genomic analysis. *Frontiers in Neuroscience*, *14*, 721. https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2020.00721/full

Chen, J.-W., Shrestha, L., Green, G., Leier, A., & Marquez-Lago, T. T. (2023). The hitchhikers' guide to RNA sequencing and functional analysis. *Briefings in Bioinformatics*, *24*(1), bbac529. https://doi.org/10.1093/bib/bbac529

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Davis, C. N., Jinwala, Z., Hatoum, A. S., Toikumo, S., Agrawal, A., Rentsch, C. T., Edenberg, H. J., Baurley, J. W., Hartwell, E. E., Crist, R. C., Gray, J. C., Justice, A. C., Gelernter, J., Kember, R. L., Kranzler, H. R., VA Million Veteran Program, Muralidhar, S., Moser, J., Deen, J. E., … Yen, A. W. (2025). Utility of Candidate Genes From an Algorithm Designed to Predict Genetic Risk for Opioid Use Disorder. *JAMA Network Open*, *8*(1), e2453913. https://doi.org/10.1001/jamanetworkopen.2024.53913

Deak, J. D., & Johnson, E. C. (2021). Genetics of substance use disorders: A review. *Psychological Medicine*, *51*(13), 2189–2200. https://www.cambridge.org/core/journals/psychological-medicine/article/genetics-of-substance-use-disorders-a-review/B3BAE9D2DCF78C7C4833A8DB4420F5B2

Donaldson, K., Cardamone, D., Genovese, M., Garbely, J., & Demers, L. (2021). Clinical Performance of a Gene-Based Machine Learning Classifier in Assessing Risk of Developing OUD in Subjects Taking Oral Opioids: A Prospective Observational Study. *Annals of Clinical and Laboratory Science*, *51*(4), 451–460.

Ellis, R. J., Ferland, J.-M. N., Rahman, T., Landry, J. L., Callens, J. E., Pandey, G., Lam, T., Kanyo, J., Nairn, A. C., & Dracheva, S. (2025). Machine learning analysis of the orbitofrontal cortex transcriptome of human opioid users identifies Shisa7 as a translational target relevant for heroin seeking leveraging a male rat model. *Biological Psychiatry*, *98*(1), 23–33. https://www.sciencedirect.com/science/article/pii/S0006322324018158

Engeln, M., Fox, M. E., Chandra, R., Choi, E. Y., Nam, H., Qadir, H., Thomas, S. S., Rhodes, V. M., Turner, M. D., Herman, R. J., Calarco, C. A., & Lobo, M. K. (2022). Transcriptome profiling of the ventral pallidum reveals a role for pallido-thalamic neurons in cocaine reward. *Molecular Psychiatry*, *27*(10), 3980–3991. https://doi.org/10.1038/s41380-022-01668-7

Erfanian, N., Heydari, A. A., Feriz, A. M., Iañez, P., Derakhshani, A., Ghasemigol, M., Farahpour, M., Razavi, S. M., Nasseri, S., Safarpour, H., & Sahebkar, A. (2023). Deep learning applications in single-cell genomics and transcriptomics data analysis. *Biomedicine & Pharmacotherapy*, *165*, 115077. https://doi.org/10.1016/j.biopha.2023.115077

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22. https://www.jstatsoft.org/article/view/v033i01

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: An Overview* (arXiv:2008.05756). arXiv. https://doi.org/10.48550/arXiv.2008.05756

Hancock, D. B., Markunas, C. A., Bierut, L. J., & Johnson, E. O. (2018). Human Genetics of Addiction: New Insights and Future Directions. *Current Psychiatry Reports*, *20*(2), 8. https://doi.org/10.1007/s11920-018-0873-3

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hwang, H., Jeon, H., Yeo, N., & Baek, D. (2024). Big data and deep learning for RNA biology. *Experimental & Molecular Medicine*, *56*(6), 1293–1321. https://doi.org/10.1038/s12276-024-01243-w

Jang, W.-J., Lee, S., & Jeong, C.-H. (2024). Uncovering transcriptomic biomarkers for enhanced diagnosis of methamphetamine use disorder: A comprehensive review. *Frontiers in Psychiatry*, *14*, 1302994. https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2023.1302994/full

Lo-Ciganic, W.-H., Huang, J. L., Zhang, H. H., Weiss, J. C., Kwoh, C. K., Donohue, J. M., Gordon, A. J., Cochran, G., Malone, D. C., Kuza, C. C., & Gellad, W. F. (2020). Using machine learning to predict risk

of incident opioid use disorder among fee-for-service Medicare beneficiaries: A prognostic study. *PLOS ONE*, *15*(7), e0235981. https://doi.org/10.1371/journal.pone.0235981

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: A tutorial. *Molecular Systems Biology*, *15*(6), MSB188746. https://doi.org/10.15252/msb.20188746

National Center for Biotechnology Information}. (2024). *GEO Accession GSE253155: Gene expression differences of alcohol use disorder in human brain* [Dataset]. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE253155

Phan, B. N., Ray, M. H., Xue, X., Fu, C., Fenster, R. J., Kohut, S. J., Bergman, J., Haber, S. N., McCullough, K. M., Fish, M. K., Glausier, J. R., Su, Q., Tipton, A. E., Lewis, D. A., Freyberg, Z., Tseng, G. C., Russek, S. J., Alekseyev, Y., Ressler, K. J., … Logan, R. W. (2024). Single nuclei transcriptomics in human and non-human primate striatum in opioid use disorder. *Nature Communications*, *15*(1), 878. https://doi.org/10.1038/s41467-024-45165-7

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics*, *2*, 927312. https://doi.org/10.3389/fbinf.2022.927312

Rosato, A. J., Chen, X., Tanaka, Y., Farrer, L. A., Kranzler, H. R., Nunez, Y. Z., Henderson, D. C., Gelernter, J., & Zhang, H. (2019). Salivary microRNAs Identified by Small RNA Sequencing and Machine Learning as Potential Biomarkers of Alcohol Dependence. *Epigenomics*, *11*(7), 739–749. https://doi.org/10.2217/epi-2018-0177

Saunders, G. R. B., Wang, X., Chen, F., Jang, S.-K., Liu, M., Wang, C., Gao, S., Jiang, Y., Khunsriraksakul, C., Otto, J. M., Addison, C., Akiyama, M., Albert, C. M., Aliev, F., Alonso, A., Arnett, D. K., Ashley-Koch, A. E., Ashrani, A. A., Barnes, K. C., … Vrieze, S. (2022). Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature*, *612*(7941), 720–724. https://doi.org/10.1038/s41586-022-05477-4

Shetty, A. C., Sivinski, J., Cornell, J., McCracken, C., Sadzewicz, L., Mahurkar, A., Wang, X.-Q., Colloca, L., Lin, W., & Pilli, N. (2024). Peripheral blood transcriptomic profiling of molecular mechanisms commonly regulated by binge drinking and placebo effects. *Scientific Reports*, *14*(1), 10733. https://www.nature.com/articles/s41598-024-56900-x

Smail, C., & Montgomery, S. B. (2024). RNA Sequencing in Disease Diagnosis. *Annual Review of Genomics and Human Genetics*, *25*(1), 353–367. https://doi.org/10.1146/annurev-genom-021623-121812

Sullivan, K. A., Kainer, D., Lane, M., Cashman, M., Miller, J. I., Garvin, M. R., Townsend, A., Quach, B. C., Willis, C., & Kruse, P. (2025). Multiomic network analysis identifies dysregulated neurobiological pathways in opioid addiction. *Biological Psychiatry*, *98*(1), 11–22. https://www.sciencedirect.com/science/article/pii/S0006322324017815

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. https://academic.oup.com/jrsssb/article-abstract/58/1/267/7027929

Vannan, A., Dell'Orco, M., Perrone-Bizzozero, N. I., Neisewander, J. L., & Wilson, M. A. (2023). An approach for prioritizing candidate genes from RNA-seq using preclinical cocaine self-administration datasets as a test case. *G3: Genes, Genomes, Genetics*, *13*(10), jkad143. https://doi.org/10.1093/g3journal/jkad143

Volkow, N. D., Koob, G. F., & McLellan, A. T. (2016). Neurobiologic Advances from the Brain Disease Model of Addiction. *New England Journal of Medicine*, *374*(4), 363–371. https://doi.org/10.1056/NEJMra1511480

Willis, C., White, J. D., Minto, M. S., Quach, B. C., Han, S., Tao, R., Shin, J. H., Deep-Soboslay, A., Hyde, T. M., & Mayfield, R. D. (2025). Gene expression differences associated with alcohol use disorder in human brain. *Molecular Psychiatry*, *30*(4), 1617–1626. https://www.nature.com/articles/s41380-024-02777-1

Xie, Y., Guessoum, O., Schug, J., Jo, A., Cullen, D. K., Kaestner, K. H., & Blendy, J. A. (2025). Spatial transcriptomics reveals distinct cell type dynamics following opioid dependence in mice with the common human variant in the μ-opioid receptor, Oprm1 A118G. *Research Square*, rs-3. https://pmc.ncbi.nlm.nih.gov/articles/PMC12393594/

Yi Xu1, , Liyu Cao1, , Xinyi Zhao, , Yinghao Yao, , Qiang Liu, , Bin Zhang, Yan Wang, , Jennie Z. Ma, Ming D. Li, & Lanjuan Li. (2020). Prediction of Smoking Behavior From Single Nucleotide Polymorphisms With Machine Learning Approaches. *Frontiers in Psychiatry 11*, *11*(416).

Zou, H., & Hastie, T. (2005a). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. https://academic.oup.com/jrsssb/article-abstract/67/2/301/7109482

Zou, H., & Hastie, T. (2005b). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. https://academic.oup.com/jrsssb/article-abstract/67/2/301/7109482