ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

Recent advances in genetic feature marker discovery through differential expression and biostatistical analysis

Ankita Saha[1,2], Shibakali Gupta[3], Chyan Paul[4], Saurav Mallik[5,6]*, and Korhan Cengiz[7]*

[1]Department of Computer Science, Swami Vivekananda University, Barrackpore, West Bengal, India

[2]Department of Science and Management, ABS Academy of Management and Health Science, Durgapur, West Bengal, India

[3]Department of Computer Science and Engineering, University Institute of Technology, Burdwan University, West Bengal, India

[4]Department of Computer Science and Engineering, Swami Vivekananda University, Barrackpore, West Bengal, India

[5]Department of Biostatistics, University of Miami, Florida, United States of America

[6]College of Pharmacy, University of Arizona, Tucson, Arizona, United States of America

[7]Department of Electrical Engineering, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia

Abstract

Genetic feature discovery is essential for understanding complex diseases and traits. This comprehensive review provides an in-depth comparison of differential expression analysis methods and statistical hypothesis tests—such as Student's *t*-test, Chi-square test, analysis of variance, Empirical Bayes methods, and Significant Analysis of Microarrays—used in genetic feature marker discovery. Our analysis highlights the strengths and weaknesses of these approaches in terms of methodologies, applications, performance, and accuracy. While the statistical tests provide straightforward interpretation, machine learning techniques provide superior capabilities for handling high-dimensional data and complex biological interactions. We conducted two mini-experiments: (i) Identification of differentially expressed genes, upregulated genes and downregulated genes using statistical tools (i.e., Student's t-test and Welch's t-test) under different conditions (normalization methods and *p*-value correction strategies) using the GSE31699 dataset from the NCBI Gene Expression Omnibus, and (ii) gene set enrichment analysis—covering Kyoto Encyclopedia of Genes and Genomes pathways and Gene Ontology terms like Biological process, Cellular component and Molecular function—using the GSE30760 dataset with the DAVID 2021 tool. Furthermore, we discussed the potential of hybrid approaches combining statistical tests with machine learning and optimization techniques for enhanced feature discovery. Future work will focus on multi-omics data integration, the development of explainable AI methods, and scalable algorithms. This review aims to serve as a comprehensive guide for researchers involved in genetic marker identification, highlighting both statistical and computational perspectives on differential expression and gene set enrichment studies.

Keywords: Genetic feature discovery; Statistical tests; KEGG pathway analysis; Gene set enrichment analysis

***Corresponding authors:**
Saurav Mallik (sauravmtech2@gmail.com);
Korhan Cengiz (kcengiz@pmu.edu.sa)

Citation: Saha A, Gupta S, Paul C, Mallik S, Cengiz K. Recent advances in genetic feature marker discovery through differential expression and biostatistical analysis. *Artif Intell Health*. 2026;3(1):54-70. doi: 10.36922/AIH025180036

Received: April 28, 2025

Revised: July 17, 2025

Accepted: August 1, 2025

Published online: September 9, 2025

1. Introduction

Genetic feature discovery is a crucial step in understanding complex diseases and traits. It involves identifying important gene variants associated with specific phenotypes. The rapid progress in high-throughput genomics techniques has led to an exponential growth in genomic information, rendering traditional statistical methods inadequate for analyzing these vast datasets. Consequently, machine learning approaches have emerged as powerful tools in genetic feature discovery, offering improved performance and accuracy. However, the choice between statistical tests and machine learning methods remains a subject of debate among researchers. Statistical tests—such as t-tests, analysis of variance (ANOVA), and Chi-square tests—have been the cornerstone of genetic feature discovery for decades. These methods provide straightforward interpretation and hypothesis testing, making them appealing for identifying significant associations between genetic variants and traits. Nevertheless, their limitations become apparent when dealing with high-dimensional data, complex interactions, and multiple testing corrections.

Machine learning approaches, including random forests, support vector machines, and neural networks, have revolutionized genetic feature discovery by handling complex relationships and high-dimensional data. These methods excel in identifying patterns and interactions that may elude traditional statistical tests. However, the "black box" nature of many machine learning models often obscures interpretability, making it challenging to understand the underlying biological mechanisms. The integration of statistical tests and machine learning methods has emerged as a promising strategy for leveraging the strengths of both approaches. Hybrid methods can combine the hypothesis-driven framework of statistical tests with the pattern-recognition capabilities of machine learning, leading to improved feature discovery and biological interpretation.

This review aims to provide a comprehensive comparison of statistical tests and machine learning approaches in genetic feature discovery. We examine the methodologies, applications, advantages, limitations, and future directions of both paradigms, while highlighting the potential of hybrid methods and emerging trends in multi-omics integration, explainable AI, and scalable algorithms. By bridging the gap between statistical and machine learning methods, this review seeks to serve as a valuable resource for advancing genetic feature discovery and unravelling the complexities of diseases and trait etiology.

2. Fundamentals of the central dogma of molecular biology and drug discovery

Biomedical research is a magnificent field of science that strives to uncover pathways to confining and treating diseases that cause morbidity and death in living things. This experimental domain covers numerous scientific disciplines and relies on rigorous exploration by scientists, chemists, and biologists. The discovery of new drugs and treatments requires robust scientific testing and thorough evaluation. Researchers in this field have the responsibility to conduct their work in a beneficent, prudent, and proper manner.[1] To address difficult biomedical problems, researchers use bioinformatics—an interdisciplinary field that integrates computational tools for analyzing biomedical data.[2]

Bioinformatics has emerged from the convergence of several disciplines, including computer science, biology, mathematics, statistics, and others. Alongside related fields like computational biology and biochemistry, bioinformatics has expanded significantly in recent years, driven by the growing need to understand complex biological systems. Defining these emerging disciplines has posed a challenge to researchers and educators alike. Among them, bioinformatics has had a particularly profound impact on the medical field. It also plays an essential role in areas such as space exploration, agriculture, and more. Broadly defined, bioinformatics is the integration of computer science, statistics, biology, and mathematics to collect, organize, analyze, and interpret biological data. This integration enables the development of software applications for analyzing DNA sequences, proteins, evolutionary genetics, biomolecular interactions, and biological networks, as well as managing datasets derived from genomic, proteomic, and post-genomic studies.[3-5]

2.1. Central dogma of molecular biology

In molecular biology, the term "central dogma" plays a pivotal role in biomarker discovery and hub gene selection. It comprises three basic processes: transcription, translation, and replication. The process of converting DNA (deoxyribonucleic acid) to RNA (ribonucleic acid) is termed transcription, whereas the process of transforming RNA to protein is called translation (Figure 1). The process of duplicating DNA is denoted as DNA replication.

There are different kinds of RNAs, such as messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). Among these, mRNA plays a major role in the detection and prognosis of various diseases and disorders like tissue-specific cancer, Alzheimer's disease, and other neurodegenerative diseases.[6,7] Aberrant gene expression,

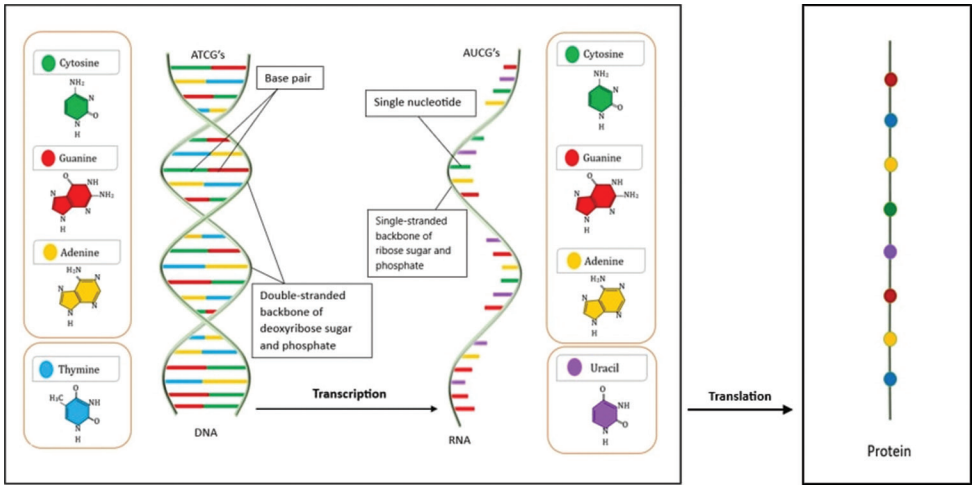

Figure 1. Central dogma of molecular biology. Image created by the authors.

whether significantly increased (up-regulation) or decreased (down-regulation), can contribute to disease pathogenesis in both humans and model organisms. The statistical significance of differential gene expression is typically assessed using *p*-values derived from hypothesis testing. A *p*-value represents the probability of observing an effect purely by chance.[8,9] The magnitude of gene expression changes is usually quantified using the fold change (FC) method.[10,11] There are two kinds of statistical hypothesis tests: parametric tests and non-parametric tests. Parametric tests (described in Section 2.4.1) are suitable for normally distributed data, whereas non-parametric tests (described in Section 2.4.2) are more appropriate when the assumption of normality is violated.[12,13]

2.2. Pre-test analysis

Before performing statistical analysis, it is important to apply pre-filtering techniques, normalize the data, and test for normality. These steps reduce analytical error and improve the reliability of results.

2.2.1. Pre-test filtering procedure

High-dimensional datasets typically require several statistical tests, which can reduce statistical power and inflate error rates if not properly controlled. Various filtering methods are utilized to evaluate differential gene expression before implementing a formal statistical test. A basic approach is to examine the all-inclusive variance of every gene (row-wise) and retain only genes with minimal variance. For example, the "genevarfilter" function in MATLAB allows users to exclude genes with low variance, based on user-defined percentiles (e.g., 5th, 30th, or 45th percentile). Genes with low variance may yield seemingly significant *p*-values, despite lacking true biological relevance. Pre-filtering such genes can help reduce false positives. In the Limma package, a variance-based filtering approach is integrated into the t-test framework to minimize spurious detections, especially in studies with small sample sizes.[14] Proper filtering is beneficial only when the false positive rate is adequately controlled. However, improper filtering—especially if not aligned with class labels—can adversely affect the control of Type I errors.

Consider a data matrix of dimension $m \times n$, where m indicates total number of genes and n indicates the number of samples. Let the expression data for gene i be denoted by $Y_i = (Y_{i1},\ldots,Y_m)^t$. If $Y_{i1},\ldots,Y_m$ are independently and normally distributed for every $i \in H_0$ (i.e., under the null hypothesis), then the test statistics before and after filtering are marginally independent. This implies that unadjusted *p*-values remain valid after filtering and applying the two-sample t-test. Thus, the un-adjusted *p*-value will be correct after applying two test statistics and filtering. When the sample size is bigger, the implementation of an experimental null distribution helps ensure accurate estimation of conditional effects introduced by filtering.

Indeed, if the null hypothesis is false, the test statistics and the filtering criterion are not necessarily independent. Filtering enhances detection power only if the test statistic and filtering criterion are positively correlated under the alternative hypothesis.

FC represents another filtering approach. Tools like "volcanoplot[1]" integrate FC thresholds with test statistics to identify significantly differentially expressed genes. Genes with FC values exceeding a lower threshold (up-regulated) or below an upper threshold (down-

[1] More information is available online at: http://www.mathworks.in/help/bioinfo/ref/mavolcanoplot.html

regulated) are considered as having passed the filter. The Bioconductor package "genefilter[2]" provides additional tools for implementing pre-filtering strategies.

Another commonly used method involves filtering based on intensity variation or the highest within-class mean. Consider two classes (i.e., disease vs. control). If the dataset follows a normal distribution (Gaussian distribution) with known common variance σ_2, the within-class variance for the mean is $\overline{\sigma}_2 = 2\sigma_2 / n$. Genes for which the maximum of the average expression across the two groups exceeds a threshold u^* (i.e., $\max\{\overline{Y}_{i,1_g}, \overline{Y}_{i,2_g}\} > u^*$ are retained for further analysis.

2.2.2. Data normalization techniques

Once the pre-filtering approach is complete, gene expression data must be normalized to bring measurements from various scales to a common scale. Gene-wise normalization techniques such as zero mean normalization, median normalization, and min-max normalization are commonly applied.[15,16] Other standard techniques include statistical column normalization,[15] variance stabilizing normalization,[17,18] and quantile normalization.[17,19]

2.2.3. Normality tests (NT) in data analysis

Following normalization, it is critical to apply NT[20] to each gene's expression data to assess whether the dataset conforms to a normal distribution, which may affect the accuracy of findings. Confirming normality is important for ensuring the assumptions underlying parametric statistical tests are met. Several methods are available for normality testing,[21] including the Jarque-Bera (JB) test,[22,23] Shapiro-Wilk test,[24] Anderson-Darling (AD) test,[24] Kolmogorov–Smirnov (KS) test,[24] and Lilliefors test.[24] Based on the results of these tests, parametric tests may be used for data that follow a normal distribution, whereas non-parametric tests are more appropriate for non-normally distributed data. For datasets with very small sample sizes (e.g., 1–5 samples), statistical testing may not be meaningful. In such cases, only FC methods can be applied to assess gene expression differences.

The JB test is used to evaluate whether a dataset exhibits characteristics of a normal distribution based on skewness and kurtosis. Skewness measures asymmetry, while kurtosis quantifies the "tailedness" or sharpness of the distribution peak.[25] This test does not require prior calculation of mean or standard deviation to be methodically implemented.[26] First proposed by Carlos Jarque and Anil Bera in 1980, it has become a standard method for testing normality in statistical research. The JB test statistic is defined in Equation I.

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \quad \text{(I)}$$

where n, *S*, and *K* denote sample size, sample skewness, and sample kurtosis, respectively.

For 2,000 or more sample sizes, the test statistic is compared with the Chi-squared distribution[27] with 2 degrees of freedom. If the computed statistics are larger than the critical Chi-squared value, normality is rejected. Chi-squared estimation demands a large sample size for accurate results.[28] Thus, simulation-based methods are used when sample sizes are below 2000. Typically, 100,000 normally distributed samples—generated with similar mean value and standard deviation (SD) as the original data—are used to estimate a reference distribution for the JB test statistic.

2.3. FC

FC is a basic yet widely used method for determining gene expression patterns.[29] According to the literature, two definitions of FC are available.[11] For real-time expression values, the FC for gene g is calculated as the ratio of average expression values between two groups, as shown in Equation II.

$$FC = \frac{\overline{x}_{1_g}}{\overline{x}_{2_g}} \quad \text{(II)}$$

Where $\overline{x}_{1_g}$ represents the average expression value of gene g in the experimental (case) group, and $\overline{x}_{2_g}$ represents the average expression value in the control (normal) group.

2.4. Statistical test

A statistical test is a method used to draw conclusions regarding the larger population by examining a smaller set of samples.[12] It involves using statistical techniques to analyze the data and determine whether the results are due to chance or if they are statistically significant.

2.4.1. Parametric distributions in statistics

Parametric testing methods, known as traditional or classical statistical tests, assume that the samples are drawn from normally distributed populations with similar variances across groups.[13] If the data do not fit these assumptions, nonparametric statistical tests are applied. Parametric tests are typically based on the mean expression value of genes.[30,31] Various commonly used parametric tests are briefly discussed below.

One of the most common statistical tests is the Student's *t*-test,[32] particularly the two-sample t-test, which is used to assess differences between the means of two independent

[2] The tool can be accessed at: http://www.bioconductor.org/packages/2.12/bioc/html/genefilter.html

groups. This test calculates a *p*-value using the cumulative distribution function of the t-distribution. The *p*-value measures the probability of observing a t-value as extreme as, or more extreme than, the observed one under the null hypothesis. According to conventional statistical thresholds, a $p<0.05$ indicates a statistically significant difference.

Let us assume, for a given gene g: group 1: n_1 experimental samples, $mean = \bar{x}_{1_g}$, and $SD = s_{1_g}$; and group 2: n_2 control samples, $mean = \bar{x}_{2_g}$, and $SD = s_{2_g}$. The t-statistic (t) can be calculated using Equation III, and the standard error of the difference in means (se_g) is calculated using Equation IV. The pooled standard deviation is computed using Equation V.

$$t = \frac{(\bar{x}_{1_g} - \bar{x}_{2_g})}{se_g} \tag{III}$$

$$\left(se_g = spooled * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \tag{IV}$$

$$\left(spooled = \sqrt{\frac{(n_1 - 1) * s_{1_g}^2 + (n_2 - 1) * s_{2_g}^2}{(n_1 + n_2 - 2)}} \right) \tag{V}$$

where $\bar{x}_{1_g}$ and $\bar{x}_{2_g}$ are the means of group 1 and group 2; s_{1_g} and s_{2_g} are the SDs of group 1 and group 2; and n_1 & n_2 are the sample sizes of group 1 and group 2.

The degrees of freedom for this test are given by $n_1 + n_2 - 2$. In this method, the populations are assumed to have equal variances. Testing or estimating variability of two different groups is called as "Behrens-Fisher" problem.[33,34] This problem arises when comparing means from two normally distributed but heteroscedastic populations. To address this issue, it is important to test whether the variance of each population is equivalent to the others. If they are unequal, the Welch's t-test should be used, which does not assume equal population variances. In this case, the t can be calculated using Equation VI, which provides an unpooled estimate of the population standard deviation.[35]

$$t = \left((\bar{x}_{1_g} - \bar{x}_{2_g}) \Big/ \sqrt{(s_{1_g}^2 / n_1) + (s_{2_g}^2 / n_2)} \right) \tag{V}$$

where $\bar{x}_{1_g}$ and $\bar{x}_{2_g}$ are the means of group 1 and group 2; s_{1_g} and s_{2_g} are the SDs of group 1 and group 2; and n_1 & n_2 are the sample sizes of group 1 and group 2.

The assumption of normality assumption (NA) may not always hold in gene expression data. Let each gene g have a common variance σ^2, such that the class-specific mean variance becomes $\bar{\sigma}^2 = 2\sigma^2 / n$. To pre-filter genes, one might compute a test statistic such as: $U_g^I = max\{\bar{x}_{1_g}, \bar{x}_{2_g}\}$ and retain genes for which U_g^I exceeds the user-defined threshold u^*. Alternatively, the test statistic, which resembles a standardized t-statistic with known variance, can be used: $U_g^{II} = (\bar{x}_{1_g} - \bar{x}_{2_g}) / \sqrt{2}\bar{\sigma}$. This allows for more robust detection of differentially expressed genes while explicitly incorporating variance assumptions.

2.4.2. Primary non-parametric statistical techniques

Non-parametric methods, also known as distribution-free methods, do not assume any specific underlying data distribution. These techniques are particularly useful when data violate the assumptions of parametric tests, such as normality or equal variance. Below are some commonly used non-parametric statistical methods for identifying differentially expressed genes.

The Wilcoxon rank-sum test (RST) is widely used for small sample sizes or when the data are not normally distributed. In such cases, the t-test may not be reliable, and RST provides a robust alternative. This test ranks all values from both groups together and then calculates the sum of ranks for each group.

Let $W_1 = \sum ranks_{group1}$ and $W_2 = \sum ranks_{group2}$. If the sample sizes n_1 and n_2 of group 1 and group 2 are equal, the test statistics T is defined as $T = min(T_1, T_2)$, where T_1 and T_2 are the rank sums of each group. If the sizes are not equal, the statistic is computed as follows: $T_2 = n_1 (n_1 + n_2 + 1) - T_1$. A significantly lower value of T suggests rejecting the null hypothesis of equal sample means. For small samples, critical values of T are tabulated. The z-score for each gene is computed using Equations VI, VII, and VIII.

$$z = \frac{(\left| T - mean_{w_1} \right| - 0.5)}{\sqrt{var_{w_1}}} \tag{VI}$$

$$var_{w_1} = n_2 * mean_{w_1} / 6 = n_1 * n_2 * (n_1 + n_2 + 1) / 12 \tag{VII}$$

$$mean_{w_1} = n_1 * (n_1 + n_2 + 1) / 2 \tag{VIII}$$

The RST and Mann-Whitney U test are mathematically equivalent, but RST is computationally slower.[36,37] The test statistic is given in Equation IX.

$$z = \frac{\left(u_1 - mean_{u_1} \right)}{\sqrt{var_{u_1}}} \tag{IX}$$

where $var_{u_1} = n_1 * n_2 * (n_1 + n_2 + 1) / 12$ and $mean_{u_1} = n_1 * n_2 / 2$. Here, $u_1 = T_1 - n_1 * (n_1 + 1)/2$ and $T1 = \sum ranks_{group1}$.

One limitation of the *t*-test is its sensitivity to small standard errors, especially in low-expression genes, which can yield artificially large test statistics. To overcome this, SAM introduces a small positive constant s_0 (also called a "fudge factor") to stabilize variance.[38] The SAM statistics, as proposed by Tusher *et al.*,[39] are described in Equation X.

$$t_{sam} = \frac{\bar{x}_{1_g} - \bar{x}_{2_g}}{se_g + s_0} \quad (X)$$

where $se_g = sPooled * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ is the standard error of the group means value

2.4.3. Other non-parametric tests

Introduced by Kruskal and Wallis, the KW test is utilized to determine whether multiple independent samples follow the same distribution.[40-42] It is the non-parametric equivalent of one-way ANOVA.

Another common test is the ideal discriminator (ID) method,[42] which is a resampling-based technique used to identify genes that are maximally expressed in one group and minimally expressed in another.[43] The technique justifies significance by selecting the genes (or miRNAs) that have the maximum Pearson's correlation coefficient (PCC) relative to the ID. Significance is determined by comparing the observed PCC obtained from permutations of 50,000 random columns of data.

The KS test[43,44] is one of the useful non-parametric tests utilized to assess equality of continuous, one-dimensional probability distributions.[44,45] It compares either an empirical distribution with a reference cumulative distribution, or two empirical distributions. The test statistic represents the maximum deviation between the two distributions and is sensitive to differences in both location and shape. The KS test is particularly useful for evaluating distributional differences between two samples. Its asymptotic *p*-values are reliable when: $(n_1 * n_2)/(n_1 + n_2)4$, where n_1 and n_2 are the sample sizes of the two groups.

2.5. Testing errors and performance metrics

This section describes different types of hypothesis testing errors in statistics, followed by performance metrics used to evaluate statistical methods. Two popular types of errors in statistics are type I and type II errors (Figure 2),[45,46] which describe specific flaws in experimental inference.[46,47]

Let us assume, m is the number of null hypotheses (e.g., genes), n is the number of conditions (samples), and R is the number of rejected null hypotheses. A Type I error occurs when the null hypothesis H_0 is incorrectly rejected, even though it is actually true. The probability of

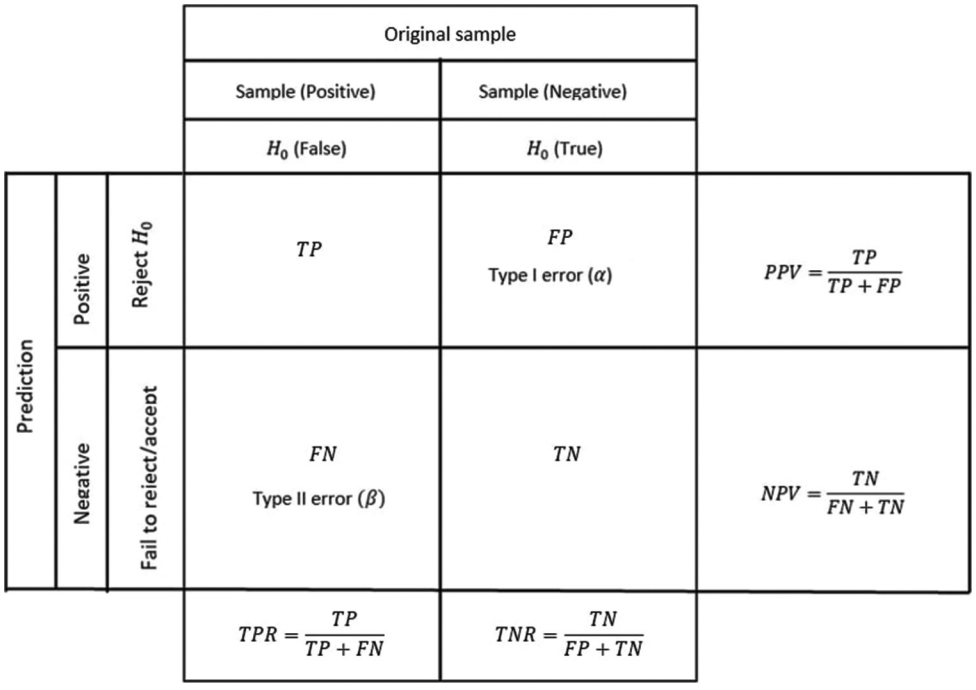

Figure 2. Relationship between hypothesis test outcome and error types. Image created by the authors.
Abbreviations: α: Type I error; β: Type II error; H_o: Null hypothesis; FN: False negative; FP: False positive; NPV: Negative predictive value; PPV: Positive predictive value; TN: True negative; TP: True positive; TNR: True negative rate; TPR: True positive rate.

committing a Type I error is referred to as the statistical significance level, denoted by α. Type I error rates can be characterize via four ways:[47] (1) Per comparison error rate (*PCER*), which is the expected proportion of false positive value (*E*) of Type I error (*EP*) is divided by the total number across all hypothesis (*m*) ($PCER = E(FP)/m$); (2) Per-family error rate (*PFER*), which is the expected number (*E*) of *FP* in the set $PFER = E(FP)$; (3) Family-wise error rate (*FWER*), which measures the probability (P) of one or more *FP* ($FWER = P(FP \geq 1)$); and (4) False discovery rate (*FDR*), introduced by Benjamini and Hochberg, is the expected proportion (E) of *FP* among the rejected hypothesis ($FDR = E(EP/(TP+FP)), R > 0$).[48]

Fundamentally, in multiple testing process, $PCFR \leq FWER \leq PFER$. Therefore, the *PFER* method is more conventional than the *FWER* technique, while *FWER* practice is more conventional the *PCER*. *PFER* produces more false positive errors than the *PWER*, while *FWER* produces more false positive errors than *PCER*.

A type II error occurs when the alternative hypothesis H_1 is rejected even though it is actually true. The probability of this error is denoted by β, and the power of the test is defined as $1-\beta$, representing the probability of selecting H_1 when it is true. In statistical classification—particularly in medicine and bioinformatics—there are two key performance indicators: (1) Sensitivity (true positive, TP, rate): proportion of correctly identified TPs. (2) Specificity (true negative, TN, rate): proportion of correctly identified TNs. These metrics are intrinsically linked to type I and type II errors and are summarized along with related metrics in Table 1.

In 1975, a biochemist named Brian W. Matthews introduced the Matthews correlation coefficient (MCC), which is a quality-centric measurement of binary classifications. It is especially suitable for imbalanced datasets, providing a single-valued metric that summarizes the confusion matrix.[49] MCC stands out as a robust and balanced metric, particularly well-suited for handling datasets of varying sizes and uneven class distributions, making it an effective tool for evaluating performance in such scenarios. MCC ranges from -1 (complete disagreement between prediction and actual outcome) and +1 (perfect prediction). An MCC value of 0 indicates an arbitrary prediction of the actual value. This statistic is also called the phi coefficient. The MCC[50] can be computed using Equation XI as follows.[51]

$$MCC = \frac{\left(TP * TN\right) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad \text{(XI)}$$

where TP is the true positives, TN is the true negatives, FP is the false positives, and FN is the false negatives.

2.6. Post-test method: Various *p*-value corrections

Multiple testing correction refers to the adjustment of *p*-values when statistical analyses are repeated multiple times on the same dataset. When each test is conducted at a 5% significance level, every individual test carries a 5% probability of committing a type I error, i.e., incorrectly rejecting the null hypothesis.[51] However, when many tests are conducted simultaneously, the FWER—the probability of committing at least one type I error—can exceed 5% and may reach approximately 30% in some scenarios.[52] Therefore, it is essential to control this cumulative error rate to avoid false discoveries.

To maintain the FWER at a desired level (commonly set at $\alpha = 0.05$), the significance threshold for individual tests must be adjusted to be more stringent.[53,54] *p*-value

Table 1. Advantages and disadvantages of parametric and non-parametric tests

Test	Advantages	Disadvantages
FC	• Useful for a small number of samples (e.g., 1–2). • Simpler biological interpretation.	• Prone to high FPs. • Ignore variability and is highly affected by outliers.
Student's t-test	• Performs well with a large sample size from two groups with similar variance.	• Performs poorly with a small number of samples. • FP rate increases with increasing variance. • Not suitable for comparing more than two groups.
Mann–Whitney U test (Non-parametric)	• Robust to outliers and works well for non-normal data. • Useful for ranked or ordinal data. • Detects median difference. • More efficient and fewer chances of mistakes.	• Less powerful than the t-test for normally distributed data. • Calculations are more complicated, especially for small sample sizes. • Not appropriate for comparing more than two groups.
SAM	• Avoids the problem of having small variances. • Suitable for small sample sizes. • Uses permutation to account for gene correlation • Does not rely on parametric assumptions. • Reports local FDR and links expression changes over time.	• Performance is inconsistent for small sample sizes. • The correlation method for sample variance is not model-motivated.

Abbreviations: FC: Fold change; FDR: False discovery rate; FP: False positive; SAM: Significance Analysis of Microarrays.

corrections are applied to account for the increased likelihood of FP resulting from multiple comparisons. In genomic data analysis, each gene or miRNA is typically tested independently, and the FP rate is directly proportional to both the number of comparisons and the chosen *p*-value threshold. GeneSpring and other statistical platforms categorize multiple testing correction methods into four main types.

2.6.1. Bonferroni correction

The Bonferroni correction is a conservative method used to adjust *p*-values when various dependent or independent statistical tests are performed on an individual dataset.[50] This technique controls the FWER by dividing the desired significance level (α) by the number of comparisons (m). Alternatively, it can be implemented by multiplying each unadjusted *p*-value by the number of hypotheses tested: *Adjusted* $p = p \times m$. If the adjusted *p*-value is still less than the chosen significance threshold (e.g., 0.05), the test result is considered statistically significant.[55] In this context, m represents the total number of genes (or miRNAs) tested. The Bonferroni method is particularly effective for strong control of the FWER when many pairwise tests are involved. This method discards null hypothesis, H_g if the unadjusted *p*-value is equal to or lesser than α/m. The single-step Bonferroni-corrected *p*-value is calculated as $\tilde{p}_g = min(m\tilde{p}_g, 1)$. Then, based on Boole's inequality, the FWER is bounded following Equation XII.

$$\begin{aligned} FWER = Pr(FP \geq 1) &= \Pr\left(\bigcup_{g=1}^{m_0}\left\{\tilde{p}_g \leq \alpha\right\}\right) \\ &\leq \sum_{g=1}^{m_0} Pr\left(\tilde{p}_g \leq \alpha\right) \\ &\leq \sum_{g=1}^{m_0} Pr\left(p_g \leq \frac{\alpha}{m}\right) \leq \frac{m_0\alpha}{m} \end{aligned} \quad \text{(XII)}$$

Where m_0 implies the total number of the true null hypothesis and p_g is the unadjusted *p*-value for gene *g*. This final inequality follows from the assumption that under the null hypothesis H_g, the probability $Pr\ (p_g \leq (x|H_g)) \leq x$ for $x \in [0,1]$.[56]

2.6.2. Bonferroni-Holm (step-down) correction

Although similar to Bonferroni correction, the Bonferroni-Holm correction is slightly less stringent, offering improved statistical power.[57] In this method, the unadjusted *p*-value of each gene (or miRNA) is first sorted in increasing sequence.[58] Then, a sequence of comparisons is performed: (1) The smallest *p*-value is multiplied by the total number of hypotheses m; (2) the second smallest is multiplied by $m-1$; (3) the third by $m-2$, and so on.[55] This process continues until a *p*-value fails to meet the significance threshold (e.g., 0.05), at which point the procedure stops, and all remaining hypotheses are not rejected. Let us assume $Pr_1 \leq Pr_2 \leq \ldots Pr_m$ denote the observed unadjusted *p*-values and $Hr_1, Hr_2, \ldots, Hr_m$ indicates the null hypotheses. As stated by Holm (1979), the index can be defined as $g^* = min\left\{g : Pr_g > \frac{\alpha}{m-g+1}\right\}$. and the hypotheses Hr_g, where g = 1,…, g^*-1, are all rejected. If no such g^* exist, then all hypotheses are rejected. Since the correction becomes progressively less stringent as the *p*-value increases, this method is uniformly more powerful than the Bonferroni correction.[59]

2.6.3. Westfall and Young permutation method

Unlike Bonferroni and Holm procedures—which are single-step methods that adjust *p*-values independently—the Westfall and Young permutation method incorporates the dependency structure between tests.[60] This approach is particularly suitable for genomic data such as DNA microarrays, where expression levels of many genes are often highly correlated.[61]

The method follows a step-down process, similar to Holm's, but uses permutations to create resampled datasets. Specifically, the data are randomly partitioned into two artificial groups (e.g., control and treatment), and *p*-values for all genes are calculated within each permuted dataset. This process is repeated many times to generate a null distribution of *p*-values.

The single-step min *p* adjusted *p*-values, for a gene g, are corrected using Equation XII.

$$\tilde{p}_g = \Pr\left(\min_{1\leq l\leq m} P_l \leq p_g \middle| H_0^C\right) \quad \text{(XII)}$$

where P_l refers to the unadjusted *p*-value of the l^{th} hypothesis and H_0^C denotes the complete null hypothesis.

Alternatively, the single-step max T adjusted *p*-value is defined as in Equation XIII.

$$\tilde{p}_g = \Pr\left(\max_{1\leq l\leq m} |T_l| \geq |t_g| \middle| H_0^C\right) \quad \text{(XIII)}$$

where T_l denotes the non-normally distributed test statistic of the l^{th} hypothesis (e.g., t-statistic of l^{th} hypothesis with different degrees of freedom across tests). This permutation-based correction provides one of the most powerful FWER control methods, as it directly accounts for correlation between test statistics. However, due to its computational intensity, it may be impractical for large datasets or high-throughput applications.

2.6.4. Benjamini-Hochberg (BH) false discovery rate (FDR) correction

The BH FDR correction is a less stringent multiple testing correction compared to Bonferroni or Holm. Unlike those approaches which control the FWER, the BH method controls the FDR, which is the expected proportion of FP among all rejected hypotheses. This relaxation allows more discoveries (TP) to emerge.[62,63]

The BH method was introduced by Yoav Benjamini and Yosef Hochberg (1995)[64,65] and it is widely used due to its balance between sensitivity (true discovery) and error control. In comparison to other corrections, the BH procedure is less conservative, allowing more hypotheses to be rejected. It tolerates a small proportion of FP, leading to fewer FN. It also assumes independence or positive dependence among test statistics, though empirical extensions have addressed violations of these assumptions.

To apply the correction, the unadjusted *p*-values are first sorted in ascending order: $Pr_1 \leq Pr_1 \leq Pr_m$. To control the FDR at a chosen level α, the largest index g* is determined using Equation XIV.

$$\max\{g : p_{r_g} \leq (g/m)\alpha\} \quad \text{(XIV)}$$

where Pr_g is the unadjusted *p*-value, m is the total number of hypotheses tested, and α is the chosen significance level. Then, all hypotheses H_{r_g}, where g = 1,…, g*, are rejected. If no such g* exists, then no hypothesis is rejected. The adjusted p values are then calculated using Equation XV.

$$\tilde{p}_{r_g} = \min_{k=g,\ldots,m}\left\{\min\left(\frac{m}{k}p_{r_k},1\right)\right\} \quad \text{(XV)}$$

For large-scale data, the FDR can also be estimated using Empirical Bayes methods.[66,67] These approaches combine frequentist inference with Bayesian shrinkage techniques and provide a multivariate estimation strategy for both effect sizes and error rates.

2.7. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and gene ontology (GO) analysis

In addition to *p*-value-based gene filtering used for hypothesis testing, KEGG pathway and GO filtering can be applied to identify biologically meaningful gene subsets. For instance, KEGG identifiers such as hsa0521 (bladder cancer), hsa05216 (thyroid cancer), hsa05310 (asthma), and hsa05144 (malaria) allow for pathway-level filtering, while GO identifiers like GO:0000122, GO:0005829, and GO:0005515 are used for ontology-based categorization. To perform such filtering, web-based tools like DAVID (Database for Annotation, Visualization, and Integrated Discovery) are commonly used. In DAVID, users can upload a list of gene identifiers along with the species name (e.g., Homo sapiens) to retrieve KEGG Pathway and GO terms associated with the input gene set. Internally, DAVID uses Fisher's Exact Test to determine the statistical significance of gene-term enrichment. It evaluates whether the overlap between the gene set and a specific pathway or GO term occurs more frequently than expected by chance. A *p*-value threshold (commonly $p<0.005$) is used to determine significantly enriched pathways or ontology categories.

GO is a structured, controlled vocabulary representing gene product attributes across all species. It covers three domains: (1) Biological process (BP): describes the broader biological objectives or pathways to which a gene contributes; (2) Cellular component (CC): indicates the subcellular location or structure where the gene product is active; and (3) Molecular function (MF): refers to the biochemical activity of a gene product.

This enrichment-based approach facilitates Gene Set Enrichment Analysis (GSEA) by revealing potential functional implications of gene expression changes and connecting statistically significant genes to known biological contexts.

2.8. Detection of miRNA target

A single miRNA can regulate multiple genes. In general, a miRNA can reduce the expression of its target genes. A miRNA is a non-coding biomolecule are crucial biomarker for various diseases, particularly in cancer diagnosis and prognosis. Various methods have been developed for miRNA target detection, such as bioinformatics tools using high-throughput sequencing data, conserved seed region matching, and hybrid deep learning-based approaches that integrate convolutional and recurrent neural networks.[68] Popular computational tools for miRNA target prediction include miRanda,[1] TargetScan, and DIANA-microT-CDS.

3. Comparative study of statistical tests and other computational tools in genetic feature discovery

3.1. Differential expression analysis

3.1.1. Dataset

In this study, we utilized microarray gene expression data related to uterine leiomyoma retrieved from the NCBI Gene Expression Omnibus (GEO) under the accession number GSE31699. The dataset comprises two kinds of samples: (i) 18 uterine leiomyoma (diseased/experimental)

samples (denoted as UL) and (ii) 18 myometrial (matched normal/control) samples (denoted as MM), all derived from African-American women.

3.1.2. Comparison of different statistical tests

Firstly, we removed the genes with missing values (NA) and then low variance. Thereafter, NT using the JB test was performed to separate the dataset into normally distributed and non-normally distributed subsets. We considered only the matched pairs—16 UL and 16 MM samples—for further analysis.

Given the limitations of parametric tests on non-normally distributed data, our analysis emphasized this more challenging subset. Our analysis emphasized this more challenging subset (i.e., zero-mean normalization, min-max normalization) and thereafter applied statistical hypothesis tests (i.e., Student's two-sample t-test and Welch's two-sample t-test) without *p*-value correction as well as with *p*-value correction (using Bonferroni, BH, FDR, Holm, and Hochberg methods). Finally, we computed the differential expression analysis to determine (i) the number of differentially expressed (DE) genes, (ii) the number of up-regulated genes (UpG), and (iii) the number of down-regulated genes (DownG). We set $p<0.05$ for significance as well as FC ≥1.10 (for UpG) and FC ≤−1.10 (for DownG) as cut-offs.

The results are presented in Table 2. Overall, Student's two-sample t-test and Welch's t-test exhibited only slight differences in the number of DE, UpG, and DownG under different conditions, particularly when applied to non-normally distributed data using different normalization methods and *p*-value correction techniques.

In recent years, several advanced statistical methods and tools have been developed for differential expression analysis in microarray, RNA-Seq, and other omics datasets. Among parametric tools, Limma[14] and DESeq2[69] are widely

Table 2. Comparative analysis of statistical hypothesis tests under varying normalization methods and *p*-value corrections using the non-normally distributed microarray dataset (GSE31699).

ID	Statistical hypothesis test	Normalization method	*p*-value correction	DE genes (*n*)	Up-regulated genes (*n*)	Down-regulated genes (*n*)
CS1	Student's t-test	ZM	Without correction	570	16	12
CS2	Student's t-test	ZM	Bonferroni	3	1	0
CS3	Student's t-test	ZM	BH	111	10	7
CS4	Student's t-test	ZM	FDR	111	10	7
CS5	Student's t-test	ZM	Holm	3	1	0
CS6	Student's t-test	ZM	Hochberg	3	1	0
CS7	Student's t-test	MM	Without correction	570	16	12
CS8	Student's t-test	MM	Bonferroni	3	1	0
CS9	Student's t-test	MM	BH	111	10	7
CS10	Student's t-test	MM	FDR	111	10	7
CS11	Student's t-test	MM	Holm	3	1	0
CS12	Student's t-test	MM	Hochberg	3	1	0
CS13	Welch t-test	ZM	Without correction	559	16	12
CS14	Welch t-test	ZM	Bonferroni	1	0	0
CS15	Welch t-test	ZM	BH	111	10	7
CS16	Welch t-test	ZM	FDR	54	7	5
CS17	Welch t-test	ZM	Holm	1	0	0
CS18	Welch t-test	ZM	Hochberg	1	0	0
CS19	Welch t-test	MM	Without correction	559	16	12
CS20	Welch t-test	MM	Bonferroni	1	0	0
CS21	Welch t-test	MM	BH	54	7	5
CS22	Welch t-test	MM	FDR	54	7	5
CS23	Welc1h t-test	MM	Holm	1	0	0
CS24	Welch t-test	MM	Hochberg	1	0	0

Abbreviations: DE: Differentially expressed; BH: Benjamini-Hochberg; FDR: False discovery rate; MM: Min-max normalization; ZM: Zero-mean normalization.

adopted. Among non-parametric approaches, commonly used methods include the Mann–Whitney U test, Shrink *t*-test, and SAM.[39] For high-throughput sequencing data analysis, the Genome Analysis Toolkit developed by the Broad Institute[70] is widely used. It supports functionalities such as identifying Single-nucleotide polymorphisms, assessing copy number variations, and detecting structural variations, in addition to differential expression analysis.

Deep learning-based frameworks have also emerged. One notable example is DeepDiff,[71] which predicts differential gene expression scores from histone modification data. In parallel, recent studies have proposed improved methodologies to optimize DE analysis. Gomez *et al.*[72] demonstrated a computational drug discovery pipeline using DE signatures. Peng *et al.*[73] developed a high-performance ensemble-based inference framework for proteomics data. Aurelio *et al.*[74] proposed a DE analysis pipeline tailored to non-model species (e.g., *Cedrela odorata*).

For single-cell sequencing, dedicated tools such as DEsingle,[75] Pagoda2,[76] Seurat,[77] and Ascend[78] offer specialized functions for differential expression and methylation analyses. Several other recent well-known cancer diagnosis methods that use machine learning, deep learning, or optimization include mammography-based diagnosis,[79] integrated ultrasound and mammography approaches,[80] and skin lesion classification.[81-83]

3.2. Gene set enrichment study

In addition to the differential expression analysis, we also conducted a traditional GSEA study using the NCBI GEO dataset GSE30760. The analysis was performed using DAVID 2021 (December 2021) with the latest DAVID Knowledgebase v2023q4 enrichment tool.[84]

3.2.1. KEGG pathway analysis

Using a corrected *p*-value threshold of <0.05, we obtained 138 enriched KEGG pathways. Using FDR-corrected $p<0.05$, we identified 120 enriched KEGG pathways. A detailed summary of the top 10 KEGG pathways and the associated statistics is provided in Table 3. Additionally, the complete list of enriched KEGG pathways is available in Supplementary file[3].

3.2.2. GO:BP

In this analysis, we obtained 745 enriched GO:BP terms based on *p*-value correction <0.05. For FDR-corrected $p<0.05$, we identified 182 enriched GO:BP terms. The top five most significantly enriched GO:BP terms include signal transduction, positive regulation of transcription by RNA polymerase II, cell adhesion, positive regulation of DNA-templated transcription, and negative regulation of transcription by RNA polymerase II. A summary of the top 10 GO:BP terms with the associated statistics are provided in Table 4. Complete details of all enriched GO:BPs terms are provided in Supplementary File[4].

3.2.3. GO:CC

This analysis identified 183 enriched GO:CC terms with $p<0.05$ and 91 enriched GO:CC terms when FDR-corrected $p<0.05$ was applied. The top five enriched GO:CC terms are as follows: cytosol, plasma membrane, membrane, cytoplasm, and extracellular exosome. The top 10 GO:CC terms and the associated statistics are provided in Table 5,

[3] Data available at GSE30760_DAVID_allgenes_geneset_encriched_KEGG_path.csv

[4] Data available at GSE30760_DAVID_allgenes_geneset_enriched_GO_BP.csv

Table 3. Top ten enriched KEGG pathways GSE30760 gene expression data using DAVID 2021 and DAVID Knowledgebase v2023q4

KEGG pathway ID & name	Genes (*n*)	*p*	Gene names	FDR
hsa05200: Pathways in cancer	245	6.09E-09	*RB1, SPI1, HHIP, KEAP1, CALML3*	9.09E-07
hsa05205: Proteoglycans in cancer	110	8.15E-09	*IHH, FZD10, FGF2, ELK1, TNF*	9.09E-07
hsa04550: Signaling pathways regulating pluripotency of stem cells	82	3.52E-08	*GSK3B, WNT2B, RIF1, ONECUT1, PIK3CD*	2.62E-06
hsa04015: Rap1 signaling pathway	111	5.68E-08	*ITGA2B, CTNND1, CALML3, CALML4, FGF2*	3.17E-06
hsa04510: Focal adhesion	106	1.40E-07	*MYLK2, ITGA2B, ELK1, ACTB, MYLK*	6.24E-06
hsa04514: Cell adhesion molecules	84	1.25E-06	*CD86, CD40, PTPRS, ITGAM, ITGB2*	3.87E-05
hsa04820: Cytoskeleton in muscle cells	115	1.28E-06	*ITGA2B, ENO3, ACTB, ACTG2, COMP*	3.87E-05
hsa04072: Phospholipase D signaling pathway	80	1.39E-06	*DGKG, DGKE, DGKD, DGKA, PIK3CD*	3.87E-05
hsa04611: Platelet activation	69	3.29E-06	*MYLK2, ITGB3, ITGA2B, PIK3CD, PIK3CB*	8.16E-05
hsa05414: Dilated cardiomyopathy	58	1.60E-05	*ITGB3, ITGA2B, TNF, ACTB, SLC8A2*	3.42E-04

Abbreviation: FDR: False discovery rate.

Table 4. Top ten GO: BP terms from GSE30760 gene expression data using DAVID 2021 and DAVID Knowledgebase v2023q4

GO: BP ID and name	Genes (*n*)	*p*	Gene names	FDR
GO: 0007165: Signal transduction	547	2.27E-30	*CNTFR, GMFB, GMFG, GLDN, CRHBP*	2.07E-26
GO: 0045944: Positive regulation of transcription by RNA polymerase II	508	3.56E-29	*ATF1, RB1, EHF, SPI1, GABPB2*	1.62E-25
GO: 0007155: Cell adhesion	246	3.74E-17	*SLC23A2, APP, SPON1, COL12A1, ICAM2*	1.14E-13
GO: 0045893: Positive regulation of DNA-templated transcription	299	1.66E-16	*TRRAP, GPATCH3, ELK1, ACTB, PSMD9*	3.79E-13
GO: 0000122: Negative regulation of transcription by RNA polymerase II	372	1.72E-14	*ZNF177, RB1, TCERG1, APP, ZNF296, SPI1*	3.13E-11
GO: 0001525:Angiogenesis	131	1.70E-13	*PLXND1, ITGA2B, SERPINE1, UBP1, RORA*	2.58E-10
GO: 0098609: Cell-cell adhesion	102	1.10E-12	*CLSTN3, CTNND2, ITGA2B, CTNND1, ICAM2*	1.44E-09
GO: 0008284: Positive regulation of cell population proliferation	209	9.71E-11	*CNTFR, VIPR1, ACTB, MYC, KDR*	1.11E-07
GO: 0007268: Chemical synaptic transmission	110	1.20E-10	*CHRM1, CHRM4, RPS6KA3, HTR6, HTR7*	1.21E-07
GO: 0048009: Insulin-like growth factor receptor signaling pathway	51	1.46E-10	*DDR1, RET, ALK, FLT1, IRS1, FLT4*	1.33E-07

Abbreviations: BP: Biological process; FDR: False discovery rate; GO: Gene Ontology.

Table 5. Top ten GO: CC terms from GSE30760 gene expression data using DAVID 2021 and DAVID Knowledgebase v2023q4

GO: CC ID & name	Genes (*n*)	*p*	Gene names	FDR
GO: 0005829: Cytosol	1935	2.01E-44	*SCOC, NUP107, TESK1, SLA2, SCP2*	2.70E-41
GO: 0005886: Plasma membrane	1828	1.45E-30	*TFRC, SLA2, HTR6, HTR7, AKT2*	9.75E-28
GO: 0016020: Membrane	1773	4.68E-30	*PGLYRP3, SPI1, NUP107, TFRC, NDST1*	2.10E-27
GO: 0005737: Cytoplasm	1933	8.47E-29	*TSKS, POP7, TESK1, SLA2, ALKBH6*	2.85E-26
GO: 0070062: Extracellular exosome	826	1.46E-28	*TFRC, ISLR, PSMD7, PSMD2, DPYSL2*	3.92E-26
GO: 0005654: Nucleoplasm	1371	6.47E-26	*ATF1, SCOC, POP7, SPI1, PWWP2B*	1.45E-23
GO: 0009986: Cell surface	298	1.75E-24	*APP, SLC46A2, SPARC, TFRC, HHIP*	3.37E-22
GO: 0009897: External side of plasma membrane	190	3.62E-17	*FCN1, CD86, CNTFR, CD84, CSF3R*	6.07E-15
GO: 0048471: Perinuclear region of cytoplasm	303	2.47E-16	*IFITM3, CYFIP1, APP, EIF4A1, TFRC*	3.68E-14
GO: 0000785: Chromatin	422	3.43E-15	*ATF1, RB1, EHF, SPI1, RAX*	4.62E-13

Abbreviations: CP: Cellular process; FDR: False discovery rate; GO: Gene Ontology.

Table 6. Top ten GO: MF terms from GSE30760 gene expression data using DAVID 2021 and DAVID Knowledgebase v2023q4

GO: MF ID & name	Genes (*n*)	*p*	Gene names	FDR
GO: 0005515: Protein binding	4522	1.62E-99	*PGLYRP3, SCOC, NUP107, TFRC, PWWP2B*	5.07E-96
GO: 0042802: Identical protein binding	676	2.92E-19	*ATF1, RB1, GABPB2, TFRC, ACCS*	4.57E-16
GO: 1990837: Sequence-specific double-stranded DNA binding	249	1.17E-14	*ZNF177, ZNF296, GFI1, FOXI1, RAX*	1.22E-11
GO: 0019904: Protein domain specific binding	107	2.73E-10	*FOXA1, APP, PLXND1, ZFYVE9, ZMYND8*	2.13E-07
GO: 0003700: DNA-binding transcription factor activity	245	9.49E-10	*ATF1, ZNF296, SPI1, GFI1, FOXI1*	5.94E-07
GO: 0005178: Integrin binding	83	1.59E-09	*APP, ITGAM, ITGB3, ITGA2B, ITGB2*	8.31E-07
GO: 0140801: Histone H2AXY142 kinase activity	65	3.44E-09	*DDR1, RET, ALK, DYRK4, ITK*	1.35E-06
GO: 0035401: Histone H3Y41 kinase activity	65	3.44E-09	*DDR1, RET, ALK, DYRK4, ITK*	1.35E-06
GO: 0005524: ATP binding	544	8.52E-09	*PI4K2B, TESK1, SMC3, SMC2, MYLK*	2.96E-06
GO: 0001228: DNA-binding transcription activator activity, RNA polymerase II-specific	202	1.02E-08	*ATF1, EHF, FOXI1, RAX, SOX21*	2.98E-06

Abbreviations: FDR: False discovery rate; GO: Gene Ontology; MF: Molecular function.

and the complete list of all the enriched GO:CCs terms is listed in Supplementary File[5].

3.2.4. GO:MF

In this analysis, we identified 208 enriched GO:MF terms with a $p<0.05$. For the FDR-corrected $p<0.05$, the enriched terms were reduced to 91. The top five most enriched GO:MF terms include protein binding, identical protein, sequence-specific double-stranded DNA binding, protein domain-specific binding, and DNA-binding transcription factor activity. The top 10 GO:MF terms and the associated statistics are provided in Table 6. Additionally, all the enriched GO:MFs terms are presented in Supplementary File[6].

4. Conclusion

In recent times, the discovery of genetic and epigenetic features—such as gene and methylation markers—has played an important role in understanding complex diseases and traits. This study provides a comprehensive review and comparative study of various well-known statistical hypothesis testing methods (i.e., Student's *t*-tests, ANOVA, Chi-square tests) in the context of genetic feature discovery and gene set enrichment analysis for microarray or RNA-seq datasets. Our analysis highlights the strengths and weaknesses of each approach, examining their methodologies, applications, performance, accuracy, and future directions. While classical statistical tests offer transparent and interpretable results, machine learning and deep learning techniques demonstrate superior capacity for managing high-dimensional data and modeling intricate biological interactions. We also explore the emerging potential of hybrid strategies that integrate statistical inference with machine or deep learning models to improve the reliability and efficiency of feature discovery. Looking ahead, promising directions include the integration of multi-omics data, the development of explainable AI models, and the advancement of scalable computational frameworks. This review serves as a resourceful guide for researchers aiming to harness the complementary strengths of statistical and machine learning methodologies in genetic and epigenetic biomarker discovery. In future work, we plan to conduct experimental evaluations using publicly available RNA-seq and Illumina DNA methylation datasets to identify robust biomarkers for various biological conditions and disease states.

Acknowledgments

We thank all lab members and researchers from our department at Swami Vivekananda University, Kolkata, India.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Ankita Saha, Shibakali Gupta, Chyan Paul

Visualization: Ankita Saha, Shibakali Gupta, Chyan Paul, Saurav Mallik

Writing–original draft: Ankita Saha, Shibakali Gupta, Chyan Paul

Writing–review & editing: Shibakali Gupta, Chyan Paul, Saurav Mallik, Korhan Cengiz

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

R code and Supplementary Files are available on https://drive.google.com/drive/folders/1oQoQZFwiPB6Zpeb0s7TmYbq WKrY0z_2_?usp=drive_link.

References

1. What is Biomedical Research? *California Biomedical Research Association*. Available from: https://statesforbiomed.org/education/background-on-biomedical-research/what-is-biomedical-research [Last accessed on 2024 Oct 09].
2. Bayat A. Clinical review science, medicine, and the future bioinformatics. *BMJ*. 2002;324:1018-1022.
   doi: 10.1136/bmj.324.7344.1018
3. Chowdhary M, Rani A, Parkash J, Shahnaz M, Dev D. Bioinformatics: An overview for cancer research. *J Drug Deliv Ther*. 2016;6(4):69-72.
   doi: 10.22270/jddt.v6i4.1290
4. Zhang S, Liu K, Liu Y, Hu X, Gu X. The role and application of bioinformatics techniques and tools in drug discovery. *Front Pharmacol*. 2025;16:1547131.
   doi: 10.3389/fphar.2025.1547131
5. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthc J*. 2021;8(2):e188-e194.
   doi: 10.7861/fhj.2021-0095
6. Khan FA, Nsengimana B, Khan NH, *et al*. Differential expression profiles of circRNAs in cancers: Future clinical

[5] Data available at GSE30760_DAVID_allgenes_geneset_enriched_GO_CC.csv

[6] Data available at GSE30760_DAVID_allgenes_geneset_enriched_GO_MF.csv

and diagnostic perspectives. *Gene Protein Dis*. 2022;1(2):138.

doi: 10.36922/gpd.v1i2.138

7. Yeh C, Madison T, Plas K. Exploring the cell-to-cell communication network to better defeat cancer. *Tumor Discov.* 2025;4(2):92.

doi: 10.36922/td.8323

8. Bandyopadhyay S, Mallik S, Mukhopadhyay A. A survey and comparative study of statistical tests for identifying differential expression from microarray data. *IEEE/ACM Trans Comput Biol Bioinform*. 2014;11(1):95-115.

doi: 10.1109/TCBB.2013.147

9. Biomolecule. *Encyclopaedia Britannica*; 2022. Available from: https://www.britannica.com/science/biomolecule [Last accessed on 2023 Mar 15].

10. Morey JS, Ryan JC, Van Dolah FM. Microarray validation: Factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biol Proced Online*. 2006;8(1):175-193.

doi: 10.1251/bpo126

11. Adler M, Alon U. Fold-change detection in biological systems. *Curr Opin Syst Biol*. 2018;8:81-89.

doi: 10.1016/j.coisb.2017.12.005

12. Vickers AJ. Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Med Res Methodol.* 2005;5:35.

doi: 10.1186/1471-2288-5-35

13. Hopkins S, Dettori JR, Chapman JR. Parametric and nonparametric tests in spine research: Why do they matter? *Global Spine J*. 2018;8(6):652-654.

doi: 10.1177/2192568218782679

14. Ritchie ME, Phipson B, Wu D, *et al*. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.

doi: 10.1093/nar/gkv007

15. Sinsomboonthong S. Performance comparison of new adjusted min-max with decimal scaling and statistical column normalization methods for artificial neural network classification. *Int J Math Math Sci*. 2022;2022:3584406.

doi: 10.1155/2022/3584406

16. Henderi H, Wahyuningsih T, Rahwanto E. Comparison of min-max normalization and Z-score normalization in the K-nearest neighbor (KNN) algorithm to test the accuracy of types of breast cancer. *Int J Inform Informat Syst.* 2021;4(1):13-20.

17. Välikangas T, Suomi T, Elo LL. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform*. 2018;19(1):1-11.

doi: 10.1093/bib/bbw095

18. Li B, Tang J, Yang Q, *et al*. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. *Sci Rep.* 2016;6:38881.

doi: 10.1038/srep38881

19. Uh HW, Klaric L, Ugrina I, Lauc G, Smilde AK, Houwing-Duistermaat JJ. Choosing proper normalization is essential for discovery of sparse glycan biomarkers. *Mol Omics*. 2020;16(3):231-242.

doi: 10.1039/c9mo00174c

20. Kwak SG, Park SH. Normality test in clinical research. *J Rheum Dis*. 2019;26(1):5-11.

doi: 10.4078/jrd.2019.26.1.5

21. Khatun N. Applications of normality test in statistical analysis. *Open J Stat*. 2021;11(1):113-122.

doi: 10.4236/ojs.2021.111006

22. Das KR. A brief review of tests for normality. *Am J Theor Appl Stat*. 2016;5(1):5.

doi: 10.11648/j.ajtas.20160501.12

23. Thadewald T, Büning H. *Jarque-Bera Test and its Competitors for Testing Normality: A Power Comparison. Diskussionsbeiträge.* Freie Universität Berlin, Fachbereich Wirtschaftswissenschaft, Berlin; 2004. Available from: https://hdl.handle.net/10419/49919 [Last accessed on 2025 Apr 19].

24. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-darling tests. *J Stat Model Anal*. 2011;2:21-33.

25. Thadewald T, Büning H. *Jarque-Bera Test and its Competitors for Testing Normality: A Power Comparison. Diskussionsbeiträge*. Freie Universität Berlin, Fachbereich Wirtschaftswissenschaft, Berlin; 2004.

26. Livingston EH. The mean and standard deviation: What does it all mean? *J Surg Res*. 2004;119(2):117-123.

doi: 10.1016/j.jss.2004.02.008

27. Ugoni A, Walker BF. The chi square test: An introduction. Aust Chiropr Osteopathy. 1995;4(3):85-91.

28. McHugh ML. The Chi-square test of independence. *Biochem Med (Zagreb)*. 2013;23(2):143-149.

doi: 10.11613/BM.2013.018

29. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*. 2009;25(6):765-771.

doi: 10.1093/bioinformatics/btp053

30. Thanavathi C. *Advanced Educational Research and Statistics*; 2017. Available from: https://www.researchgate.net/publication/337991541 [Last accessed on 2025 Apr 19].

31. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis

of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*. 2019;20:40.

doi: 10.1186/s12859-019-2599-6

32. Boareto M, Caticha N. t-Test at the probe level: An alternative method to identify statistically significant genes for microarray data. *Microarrays*. 2014;3(4):340-351.

doi: 10.3390/microarrays3040340

33. Zhang L, Zhu T, Zhang JT. Two-sample Behrens-Fisher problems for high-dimensional data: A normal reference scale-invariant test. *J Appl Stat*. 2023;50(3):456-476.

doi: 10.1080/02664763.2020.1834516

34. Hong S, Coelho CA, Park J. An exact and near-exact distribution approach to the Behrens-fisher problem. *Mathematics*. 2022;10(16):2953.

doi: 10.3390/math10162953

35. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol.* 2014;14:135.

doi: 10.1186/1471-2288-14-135

36. Dao PB. On Wilcoxon rank sum test for condition monitoring and fault detection of wind turbines. *Appl Energy*. 2022;318:119209.

doi: 10.1016/j.apenergy.2022.119209

37. Botlagunta M, Khatri K, Devi BM, Doneti R, Pasha A, Pawar SC. Differential expression of DDX3 and microRNAs in response to hormone and cisplatin against cervical cancer. EJMO. 2022;6(4):307-316.

doi: 10.14744/ejmo.2023.96531

38. Larsson O, Wahlestedt C, Timmons JA. Considerations when using the significance analysis of microarrays (SAM) algorithm. *BMC Bioinformatics*. 2005;6:129.

doi: 10.1186/1471-2105-6-129

39. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001;98(9):5116-5121.

doi: 10.1073/pnas.091062498

40. Bewick V, Cheek L, Ball J. Statistics review 10: Further nonparametric methods. *Crit Care*. 2004;8(3):196-199.

doi: 10.1186/cc2857

41. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc.* 1952;47(260):583-621.

doi: 10.2307/2280779

42. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*. 2002;18(11):1454-1461.

doi: 10.1093/bioinformatics/18.11.1454

43. Massey FJ Jr. The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc*. 1951;46(253):68-78.

doi: 10.2307/2280095

44. Steinskog DJ, Tjøtheim DB, Kvamstø NG. A cautionary note on the use of the Kolmogorov-Smirnov test for normality. *Mon Weather Rev*. 2007;135(3):1151-1157.

doi: 10.1175/MWR3326.1

45. Pushap AC, Sudershan S, Sudershan A. Type of error in statistics: A review. *Haya Saudi J Life Sci*. 2023;8(03):39-43.

doi: 10.36348/sjls.2023.v08i03.001

46. Kaur P, Stoltzfus J. Type I, II, and III statistical errors: A brief overview. *Int J Acad Med.* 2017;3(2):268-270.

doi: 10.4103/IJAM.IJAM_92_17

47. Shaffer JP. *Multiple Hypothesis Testing: A Review. Technical Report No. 23.* Research Triangle Park, NC: National Institute of Statistical Sciences; 1994. Available from: https://www.niss.org [Last accessed on 2025 Apr 19].

48. El-Gohary TM. Hypothesis testing, type I and type II errors: Expert discussion with didactic clinical scenarios. *Int J Health Rehabil Sci*. 2019;8(3):132.

doi: 10.5455/ijhrs.0000000180

49. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1):6.

doi: 10.1186/s12864-019-6413-7

50. Lise S, Archambeau C, Pontil M, Jones DT. Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. *BMC Bioinformatics*. 2009;10:365.

doi: 10.1186/1471-2105-10-365

51. Gohary T. Hypothesis testing, type I and type II errors: Expert discussion with didactic clinical scenarios. *Int J Health Rehabil Sci*. 2019;8(3):132.

doi: 10.5455/ijhrs.0000000180

52. Jafari M, Ansari-Pour N. Why, when and how to adjust your P values? *Cell J*. 2019;20(4):604-607.

doi: 10.22074/cellj.2019.5992

53. Greenland S, Senn SJ, Rothman KJ, *et al.* Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337-350.

doi: 10.1007/s10654-016-0149-3

54. Silicon Genetics. *Multiple Testing Corrections.* Redwood City, CA: Silicon Genetics; 2003.

55. Vasilopoulos T, Morey TE, Dhatariya K, Rice MJ. Limitations of significance testing in clinical research:

A review of multiple comparison corrections and effect size calculations with correlated measures. *Anesth Analg*. 2016;122(3):825-830.

doi: 10.1213/ANE.0000000000001107

56. Sedgwick P. Multiple significance tests: The Bonferroni correction. *BMJ*. 2012;344:e509.

doi: 10.1136/bmj.e509

57. Vickerstaff V, Omar RZ, Ambler G. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. *BMC Med Res Methodol*. 2019;19(1):129.

doi: 10.1186/s12874-019-0754-4

58. Blakesley RE, Mazumdar S, Dew MA, *et al*. Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology*. 2009;23(2):255-264.

doi: 10.1037/a0012850

59. Kang G, Ye K, Liu N, Allison DB, Gao G. Weighted multiple hypothesis testing procedures. *Stat Appl Genet Mol Biol*. 2009;8(1):23.

doi: 10.2202/1544-6115.1437

60. Cox DD, Lee JS. Pointwise testing with functional data using the Westfall-Young randomization method. *Biometrika*. 2008;95(3):621-634.

doi: 10.1093/biomet/asn021

61. Westfall PH, Young SS. p Value adjustments for multiple tests in multivariate binomial models. *J Am Stat Assoc*. 1989;84(407):780-786.

doi: 10.1080/01621459.1989.10478837

62. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*. 2005;21(13):3017-3024.

doi: 10.1093/bioinformatics/bti448

63. Acharya A. *A Complete Review of Controlling the False Discovery Rate in a multiple Comparison Problem Framework: The Benjamini-Hochberg Algorithm*. *arXiv:1406.7117v1 [stat.ME]*; 2014.

doi: 10.48550/arXiv.1406.7117

64. Benjamini Y. Discovering the false discovery rate. *J R Stat Soc Series B Stat Methodol*. 2010;72(4):405-416.

doi: 10.1111/j.1467-9868.2010.00746.x

65. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29(4):1165-1188.

doi: 10.1214/aos/1013699998

66. Chakraborty A, Jiang G, Boustani M, Liu Y, Skaar T, Li L. Simultaneous inferences based on empirical Bayes methods and false discovery rates in eQTL data analysis. *BMC Genomics*. 2013;14(Suppl 8):S8.

doi: 10.1186/1471-2164-14-S8-S8

67. Efron B. Microarrays, empirical Bayes and the two-groups model. *Stat Sci*. 2008;23(1):1-22.

doi: 10.1214/07-STS236

68. Gu T, Zhao X, Barbazuk WB, Lee JH. miTAR: A hybrid deep learning-based approach for predicting miRNA targets. *BMC Bioinformatics*. 2021;22(1):96.

doi: 10.1186/s12859-021-04026-6

69. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.

doi: 10.1186/s13059-014-0550-8

70. Available from: https://gatk.broadinstitute.org/hc [Last accessed 2025 Jul 03].

71. Sekhon A, Singh R, Qi Y. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. *Bioinformatics*. 2018;34(17):i891-i900.

doi: 10.1093/bioinformatics/bty612

72. Gomez CG, Rosa-Calatrava M, Fouret J. Optimizing *in silico* drug discovery: Simulation of connected differential expression signatures and applications to benchmarking. *Brief Bioinform*. 2024;25(4):bbae299.

doi: 10.1093/bib/bbae299

73. Peng H, Wang H, Kong W, *et al*. Optimizing differential expression analysis for proteomics data via high-performing rules and ensemble inference. *Nat Commun*. 2024;15:3922.

doi: 10.1038/s41467-024-47899-w

74. Aurelio AMM, Fabián CAF, Iván CCC, Felipe GL. Optimized method for differential gene expression analysis in non-model species: Case of *Cedrela odorata* L. *MethodsX*. 2023;11:102449.

doi: 10.1016/j.mex.2023.102449

75. Miao Z, Deng K, Wang X, Zhang X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*. 2018;34:3223-3224.

doi: 10.1093/bioinformatics/bty332

76. Available from: https://github.com/kharchenkolab/pagoda2 [Last accessed on 2025 Jul 15].

77. Hao Y, Stuart T, Kowalski MH, *et al*. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol*. 2024;42(2):293-304.

doi: 10.1038/s41587-023-01767-y

78. Senabouth A, Lukowski SW, Hernandez JA, *et al*. ascend: R package for analysis of single-cell RNA-seq data. *Gigascience*. 2019;8(8):giz087.

doi: 10.1093/gigascience/giz087

79. Hussain SI, Toscano E. Optimized deep learning for mammography: Augmentation and tailored architectures. *Information*. 2025;16(5):359.

doi: 10.3390/info16050359

80. Xu Z, Zhong S, Gao Y, *et al.* Optimizing breast lesions diagnosis and decision-making with a deep learning fusion model integrating ultrasound and mammography: A dual-center retrospective study. *Breast Cancer Res*. 2025;27:80.

doi: 10.1186/s13058-025-02033-6

81. Shetty B, Fernandes R, Rodrigues AP, *et al.* Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Sci Rep.* 2022;12:18134.

doi: 10.1038/s41598-022-22644-9

82. Hussain SI, Toscano E. An extensive investigation into the use of machine learning tools and deep neural networks for the recognition of skin cancer: Challenges, future directions, and a comprehensive review. *Symmetry*. 2024;16(3):366.

doi: 10.3390/sym16030366

83. Hussain SI, Toscano E. Enhancing recognition and categorization of skin lesions with tailored deep convolutional networks and robust data augmentation techniques. *Mathematics*. 2025;13(9):1480.

doi: 10.3390/math13091480

84. Available from: https://davidbioinformatics.nih.gov/home.jsp [Last accessed on 2025 Jul 02].