BioMark: Biomarker Analysis Tool

Mehmet Ali Balikci, Cyrille Mesue Njume, Ali Cakmak*
Istanbul Technical University, Maslak 34867, Istanbul, Turkiye
Corresponding author: Ali Cakmak, ali.cakmak@itu.edu.tr

***Abstract*—Biomarkers play a pivotal role in disease diagnosis and prognosis by offering molecular insights into biological states. The rapid growth of high-throughput omics technologies has enabled the generation of large-scale biomarker datasets, yet analyzing these complex, high-dimensional data remains a major challenge—particularly for researchers lacking advanced computational expertise. While numerous tools exist for omics data analysis, many fall short in providing an integrated, user-friendly environment tailored specifically for biomarker discovery and interpretation.**

To address this gap, we present BioMark, a web-based platform designed to streamline biomarker analysis across diverse omics types. BioMark integrates robust statistical methods with widely used machine learning algorithms to support key workflows, including statistical analysis, dimensionality reduction, classification, and subsequent model explanation. The platform emphasizes accessibility, offering intuitive visualizations and automated reporting to facilitate the interpretation and dissemination of results. Notably, BioMark also offers a feature-ranking strategy that consolidates outputs from multiple analytical methods, enhancing the robustness of biomarker identification. By lowering the barrier to advanced biomarker analytics, BioMark empowers a broader range of researchers to uncover clinically relevant molecular signatures and accelerate translational research. Biomark is available online at https://bioinf.itu.edu.tr/biomark.

***Index Terms*—disease diagnosis, biomarker discovery, artificial intelligence, multivariate analysis**

I. Introduction

BIOMARKERS are useful to diagnose diseases and monitor the prognosis of patients. These molecular indicators provide critical insights into biological states, aiding in the identification of disease presence and the prediction of disease trajectory and outcome [1]. Biomarkers encompass a wide range of molecular entities, such as genomic alterations, gene expression levels (transcriptomics), protein abundance (proteomics), and metabolite concentrations (metabolomics), and they are fundamental to understanding the molecular underpinnings of health and disease. Their accurate identification and interpretation across diverse biological contexts are paramount for improving diagnostic capabilities, monitoring disease progression, predicting treatment responses, and ultimately personalizing treatment strategies in various medical conditions.

The advent of high-throughput technologies has led to an exponential increase in the volume and complexity of biomarker data. These technologies, including next-generation sequencing, microarray analysis, and advanced mass spectrometry, generate vast datasets that often characterize thousands or even millions of molecular features simultaneously [2], [3], [4], [5], [6], [7], [8], [9]. Analyzing these large, heterogeneous, and intricate datasets presents significant computational challenges. Issues such as data noise, missing values, batch effects, high dimensionality (where the number of features far exceeds the number of samples), and the need for integrating data from multiple omics layers require sophisticated computational, statistical, and machine learning methods for effective analysis [6], [10], [11]. For many researchers and clinicians, who may not possess extensive programming skills or have access to dedicated bioinformatics support and powerful computational resources, applying these advanced analytical methods remains a considerable barrier [11], [12].

To address these challenges, a diverse landscape of computational tools has been developed. Many platforms offer comprehensive suites for general omics analysis, including data processing, statistical tests, and visualization, often delivered via accessible web interfaces to support researchers without extensive programming backgrounds [13], [7], [8], [2], [5]. Examples range from broad platforms like **OmicsAnalyst** [14] and **PROMO** [15] to more specialized resources for metabolomics like **Metabolomics Workbench** [16] and **iMAP** [17]. Other tools focus on specific functionalities such as multi-omics integration [6], cancer prognosis analysis [1], [18], or pathway informatics using databases like KEGG [19] and Reactome [20]. Despite this rich ecosystem, a critical gap remains. Many existing tools have limitations in their ease of use for non-experts, the breadth of integrated analytical methods, or the intuitive visualization of results, forcing researchers to use multiple, disconnected tools for a complete analysis [11], [12]. There is a persistent need for a platform that seamlessly integrates the essential biomarker analysis workflows—identifying significant differences, discovering patient subgroups, and building predictive models—into a single, cohesive environment.

In response to this need, we have developed BioMark, a novel web-based tool designed to provide a user-friendly and comprehensive platform for biomarker data analysis. BioMark enables researchers to upload their high-dimensional biomarker datasets and perform crucial analyses including dimensionality reduction and visualization to discover hidden patient subgroups, and classification analysis to build predictive models for diagnosis and prognosis. The tool employs a combination of robust statistical methods and widely-used machine learning techniques, and provides intuitive visualizations to aid in the interpretation of complex results. Furthermore, BioMark offers a unique capability to combine and rank biomarker lists obtained from its distinct Statistical Analysis and Model Explanation modules, providing a more robust identification of important features by consolidating results from various perspectives. BioMark aims to make advanced

biomarker analysis accessible to a wider audience, facilitating the identification of biomarkers useful for disease diagnosis and prognosis and accelerating research in these areas, with the capability to generate comprehensive reports summarizing the analysis findings and visualizations for easy dissemination of results.

To demonstrate its practical utility, we apply BioMark to real-world transcriptomics datasets from prostate and breast cancer studies. These case studies showcase the platform's effectiveness in identifying significant biomarkers, revealing patient subgroups through dimensionality reduction and visualization, and constructing accurate diagnostic models. The remainder of this paper is organized as follows: Section II reviews related work, Section III details the architecture and methodologies of BioMark, Section IV presents the performance evaluation and case study results, and Section V provides concluding remarks.

II. RELATED WORK

The increasing availability of high-dimensional omics and biomarker data has led to the development of numerous computational tools and platforms designed to assist researchers in data analysis and interpretation. These tools can be broadly categorized under three umbrella groups: (i) comprehensive analysis platforms, (ii) niche specialized applications, and (iii) resources for downstream biological interpretation. Positioning BioMark within this landscape requires an examination of their respective strengths and limitations.

Comprehensive web-based platforms represent the most common category, offering end-to-end analysis workflows (see Table I for a qualitative comparison in different dimensions). For instance, **MetaboAnalyst** [2], **MetaboAnalystR** [3] and **OmicsAnalyst** [14] provide a powerful suite of tools for statistical analysis, functional enrichment, and visualization, but are primarily tailored for metabolomics and multi-omics data integration, respectively. Platforms like **Omics Playground** [13] and **GenoCraft** [8] excel in providing interactive and user-friendly interfaces for exploratory data analysis, including dimensionality reduction and clustering. However, a common limitation of these general-purpose tools is that their biomarker discovery workflows often rely on individual statistical methods. While powerful, they typically lack an integrated mechanism to consolidate feature importance rankings from multiple, methodologically distinct approaches (e.g., combining statistical significance from ANOVA with model-based importance from SHAP). This consolidation is a core feature of BioMark, designed to enhance the robustness and reliability of identified biomarker signatures.

A second category consists of specialized tools designed for specific research questions or data types. Numerous web servers, such as **OSacc** [18], **OSuvm** [21], and others focusing on specific cancers [22], [23], [24], [25], are dedicated to survival and prognosis analysis, often using pre-existing public datasets like The Cancer Genome Atlas (TCGA). While invaluable for prognostic studies, these tools are not designed for users to upload and analyze their own datasets for differential expression or classification model building, which is a primary function of BioMark. Other specialized resources focus on advanced visualization (e.g., **MicrobioSee** [26]), complex multi-omics integration pipelines (e.g., **MOMIC** [6]), or specific methods for outcome prediction (e.g., **X-Tile** [27]).

Finally, foundational bioinformatics resources like **KEGG** [19], **Reactome** [20], and **Pathway Tools** [28] provide the essential pathway and genome informatics necessary for the biological interpretation of results. While indispensable for downstream analysis, these are databases and toolkits rather than platforms for upstream statistical discovery from user-supplied data matrices.

While this diverse landscape offers powerful capabilities, a clear gap remains. Researchers often face a trade-off: either use a general-purpose platform that may lack dedicated, robust biomarker discovery workflows, or piece together multiple specialized tools, which can be inefficient and requires significant bioinformatics expertise. BioMark addresses this gap by providing a single, cohesive web platform specifically architected for biomarker discovery. It uniquely integrates three critical, yet often separate, workflows: (i) robust differential analysis featuring an integrated multi-method feature ranking system to increase confidence in results, (ii) intuitive visualization for sample stratification, and (iii) comprehensive classification model building and evaluation. By focusing on this core biomarker analysis pipeline and prioritizing an intuitive user experience, BioMark empowers researchers without extensive computational backgrounds to perform end-to-end analyses, from raw data to validated biomarker lists and high-performance predictive models.

III. METHODS

The BioMark tool is designed as a web-based platform to facilitate the analysis and visualization of biomarker data, primarily focusing on high-dimensional datasets such as gene expression and protein levels. The tool's architecture and functionalities are structured to enable researchers, including those without extensive programming backgrounds, to easily perform advanced analyses and gain meaningful biological insights

A. Architecture

The BioMark web tool operates through a client-server architecture, as illustrated in Fig. 1. Users interact with the platform through a web browser, which serves as the client interface.

The back-end employs a **hybrid architecture** to combine a responsive web interface with powerful scientific computation. A persistent **Node.js** server handles all client-side HTTP requests and session management. When an analysis is requested, the Node.js server acts as an orchestrator, launching the corresponding **Python** analysis script (from the `services` module) as an asynchronous **child process** (`spawn`). This Python process handles all intensive data processing, statistical analyses, and machine learning modeling.

Upon completion, the Python script communicates its results (such as file paths to generated graphs and tables, or JSON-formatted metadata) back to the Node.js server by writing to

TABLE I: Qualitative comparison of BioMark features against established omics analysis platforms. BioMark distinguishes itself through its focus on multi-method rank aggregation and deep model explainability (SHAP/LIME).

Feature / Capability	BioMark	MetaboAnalyst	OmicsAnalyst	Omics Playground
Primary Application	Biomarker Discovery	Metabolomics Profiling	Multi-omics Integration	Exploratory Data Analysis
Feature Ranking	**Multi-Method Aggregation** (RRF, Rank Prod.)	Single Method	Single Method	Single Method
AI Explainability	**Deep** (SHAP, LIME, Waterfall Plots)	Basic (Variable Importance)	Basic (Loadings/ Importance)	Moderate (SHAP Support)
ML Algorithms	Broad (XGBoost, CatBoost, RF, SVM, etc.)	Standard (PLS-DA, RF, SVM)	Standard (RF, PLS-DA)	Broad (AutoML Features)
Output Format	Interactiv Plots + **PDF Report**	Interactive	Interactive	Interactive
Pathway Enrichment	Planned	**Extensive**	Extensive	Extensive

its standard output stream (`stdout`). The Node.js server efficiently captures this output, aggregates the results, and sends the final response back to the user's web browser for display. This architecture ensures that computationally intensive tasks are handled asynchronously by the Python runtime—without blocking the main Node.js event loop—providing a robust, scalable, and responsive user experience even during long-running analyses.

B. Data Uploading

The BioMark tool is designed to work with omics datasets, typically in a tabular format where rows represent samples (e.g., patients or experiments) and columns represent biological entities (e.g., genes, proteins, metabolites) or sample metadata (e.g., disease status, sample ID). The primary data format format is CSV (Comma Separated Values). The platform also robustly handles other common tabular formats such as **TSV, TXT, and Excel (.xlsx)**, as well as compressed files **(.gz, .zip)**. Users can upload their own datasets through the web interface. For users new to the platform or wishing to explore its functionalities, a pre-loaded demo dataset is also available and can be downloaded. The tool is optimized to handle high-dimensional data, often characterized by a large number of biomarkers (columns) relative to the number of samples (rows). While primarily focused on numerical data (such as expression levels or concentrations), the tool can also handle categorical columns, particularly for the target variable.

C. Data Preprocessing and Transformation

Upon data upload, BioMark automatically identifies numerical and categorical (object-type) columns. To ensure the scientific validity of downstream analyses, the tool implements a robust preprocessing pipeline.

A critical step in this pipeline is the handling of categorical data. Unlike methods that apply ordinal encoding (e.g., assigning 1, 2, 3 to groups), BioMark utilizes **One-Hot Encoding** (`pd.get_dummies` or `OneHotEncoder`). This transformation converts a single categorical feature into multiple binary (1/0) columns, one for each unique category.

This **model-agnostic "healthy" approach** is essential, as it prevents the introduction of artificial ordinal relationships (e.g., "Group C" > "Group B") that would invalidate the results of linear models (like Logistic Regression, SVC) and distance-based algorithms (like PCA and t-SNE). Numerical features are processed using median imputation for missing values and optional `StandardScaler` for normalization.

D. Selecting Patient Groups

The analysis often involves comparing or grouping samples based on specific criteria. BioMark allows for the selection of different classes within a target variable, such as disease status (e.g., healthy vs. diseased). The platform can automatically detect different classes present in the uploaded target variable column, and users can then select the specific classes they wish to include in their analysis. This functionality is particularly important for differential analysis and classification tasks where distinct groups of samples need to be defined.

E. Determining Analysis Groups

Following the selection of patient groups based on class labels, BioMark allows for determining the specific groups to be compared in the analysis. This step is integral to the Statistical Analysis module, where the tool aims to find statistically significant differences in biomarker levels between predefined groups (e.g., comparing biomarker profiles of a 'patient' group against a 'healthy' group). The tool is designed to facilitate the comparison between different combinations of selected classes, enabling researchers to explore various group comparisons relevant to their study.

F. Selecting an Analysis Type and Customization

BioMark offers multiple analysis options, which are grouped into four main modules: Statistical Analysis, Model Explanation, Dimensionality Reduction & Visualization, and Classification Analysis.

- **Statistical Analysis:** This module uses traditional statistical tests such as the T-test and ANOVA to identify biomarkers with statistically significant differences between predefined groups.
- **Dimensionality Reduction & Visualization:** Used to visualize the underlying structure of the data and explore natural groupings among samples. This is achieved

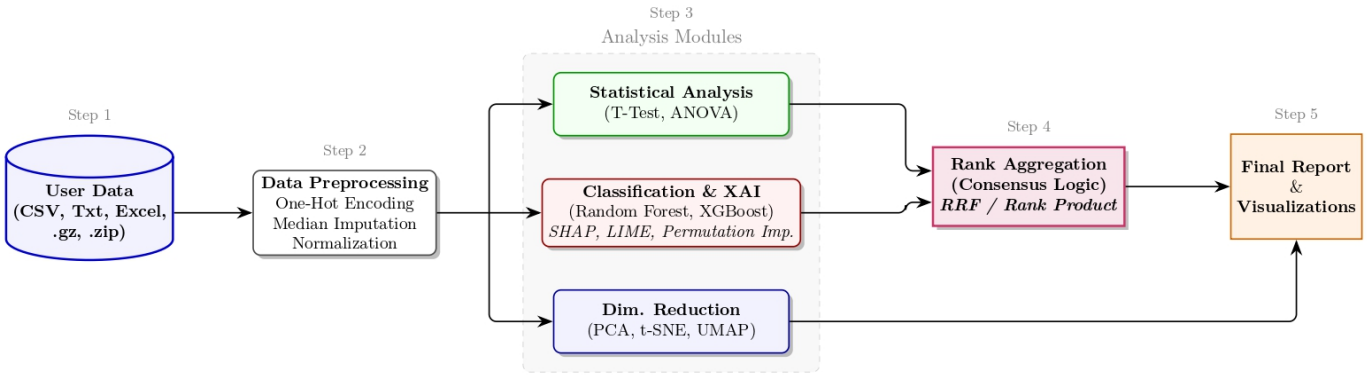


Fig. 1: The BioMark Analysis Workflow. The pipeline begins with data upload (supporting various formats, including compressed files) and automated preprocessing. The data is then processed in parallel by statistical and machine learning modules (including explainability via SHAP, LIME, and Permutation Importance). Feature importance rankings from these distinct methods are consolidated via the Rank Aggregation module to identify robust biomarkers, which are presented in the final report alongside dimensionality reduction visualizations.

by projecting high-dimensional biomarker data into a lower-dimensional space (2D or 3D) using techniques such as PCA (Principal Component Analysis), t-SNE (t-Distributed Stochastic Neighbor Embedding), and UMAP (Uniform Manifold Approximation and Projection). The resulting visualizations can reveal inherent patterns and potential subgroups within the samples based on their biomarker profiles.

- **Classification Analysis:** This module focuses on building and evaluating predictive models. Users can select from a wide range of machine learning algorithms (e.g., Logistic Regression, XGBClassifier, SVC, etc.) to train a model that assigns samples to specific categories. This step produces performance metrics (e.g., Accuracy, F1-Score) and, for tree-based models, their native feature importance scores (e.g., 'feature_importances_').
- **Model Explanation:** This module is applied *after* a model has been trained in the Classification Analysis module. It allows users to interpret the behavior of the specific model they just trained. It includes model-agnostic techniques like SHAP [29], [30], LIME [31], and Permutation Feature Importance to rank biomarkers based on their contribution to the model's predictions.

BioMark provides options for users to either utilize default parameters for these analyses or to customize detailed parameter settings for more advanced control (such as test dataset size, cross-validation folds, and model tuning options).

1) Feature Selection and Workflow Stages: BioMark incorporates a robust feature selection mechanism crucial for enhancing the interpretability and performance of downstream analyses, particularly for high-dimensional omics data. This process allows researchers to focus on the most informative biomarkers by reducing noise and computational complexity. The platform distinguishes between two primary analytical stages based on feature selection: "Without Feature Selection" and "After Feature Selection."

Feature Selection Mechanism: The core feature selection process is initiated by using the methods within the **Statistical Analysis** and **Model Explanation** modules. These modules compute and consolidate feature importance scores from the suite of methods it contains. These methods can be broadly categorized as follows:

- **Statistical Tests:** Standalone tests like ANOVA and T-test that assess each feature independently.
- **Model-Based Importance:** A set of methods that rely on internally trained machine learning models. Importance scores are derived via:
  - Model-specific attributes (XGBoost Feature Importance, Random Forest Feature Importance).
  - Explainability frameworks like SHAP and LIME are applied **in a model-agnostic manner** to the user-selected, internally trained model (e.g., XGBoost, Random Forest, etc.) to provide detailed feature attributions.

These individual method-specific rankings are then consolidated using a flexible rank aggregation methodology to generate a final consensus list. This multi-method aggregation strategy enhances the confidence in the identified top features, as they are consistently deemed important across different analytical perspectives. The detailed methodology for this consolidation, including the platform's support for multiple aggregation strategies like Reciprocal Rank Fusion (RRF) and Rank Summation, is described in Section III-H.

"Without Feature Selection" Stage: In this initial stage, all analyses, including preliminary visualizations and classification model training, are performed using the entire high-dimensional dataset with all available features. Results from this stage represent a baseline performance reference before any dimensionality reduction based on feature importance is applied, and corresponding figures are often explicitly labeled "Without Feature Selection."

"After Feature Selection" Stage: Following the feature selection process, a refined dataset is created by retaining only the top N selected biomarkers (as determined by the aggregated ranking) along with the target class column. All subsequent analyses, including advanced dimensionality re-

duction visualizations (e.g., PCA, t-SNE, UMAP plots) and re-training of classification models, are then performed exclusively on this reduced, most informative feature set. Results and figures generated in this stage are explicitly labeled with "After Feature Selection," showcasing the enhanced insights and improved model performance achieved by focusing on the most relevant molecular signatures. This two-stage approach allows for a direct comparison of analytical outcomes, emphasizing the value of feature selection in biomarker discovery.

BioMark enforces fixed random seeds (default random_state=42 in Python/Scikit-learn) for all stochastic algorithms (t-SNE, UMAP, Random Forest, XGBoost) to ensure that biomarker rankings and visualizations are fully reproducible across independent runs.

G. Explainability: Visualizing Biomarkers

Visualizations are a key component of the BioMark tool, enabling users to intuitively explore and interpret their biomarker data and analysis results. To enhance accessibility for users without a computational background, info-buttons are placed throughout the interface to describe each analysis task. Furthermore, the results page features a "Plot Guide" button, which provides detailed explanations on how to interpret each generated plot. The platform generates various types of plots tailored to the specific analyses performed:

- **Heatmap:** Displays the levels of important biomarkers across different samples using color intensity, revealing patterns and groupings. Heatmaps can also show comparative feature rankings from different statistical methods in summary analysis.
- **PCA Plot:** Visualizes how samples are distributed and grouped in a selected lower-dimensional space (2D or 3D) after PCA dimension reduction.
- **t-SNE Plot:** Similar to PCA plots, this visualizes sample distribution and grouping in a selected lower-dimensional space (2D or 3D) after t-SNE dimension reduction.
- **SHAP Summary Plot:** From SHAP analysis, this plot shows the overall impact of each feature on the model output and the direction of the effect (positive/negative).
- **SHAP Dependence Plot:** Illustrates how the model output changes as the value of a specific feature changes.
- **LIME Explanation Plots:** Provide local feature importances that explain the prediction for a single sample based on LIME analysis.
- **Statistical Method Summary Plot:** Creates a plot summarizing the results of different statistical methods used in differential analysis, showing comparative feature rankings. Separate summaries can be generated for different pairs of classes.

H. Combining and Ranking Multi-Analysis Results

A significant challenge in biomarker discovery is that different analytical methods—each with its own statistical assumptions and biases—often yield distinct lists of candidate biomarkers. Relying on a single method can lead to findings that are not robust. To address this, BioMark incorporates a powerful and flexible feature that consolidates and ranks biomarker lists generated from multiple analytical approaches, providing a more reliable, consensus-based result.

The core of this consolidation is a flexible rank aggregation methodology. BioMark supports several established strategies to generate this unified ranking, allowing users to select the method best suited for their data. The supported methods include:

- **Reciprocal Rank Fusion (RRF):** The **default and recommended method**. RRF calculates a score (Score = $\sum \frac{1}{k+\text{rank}}$) that is highly robust to outlier rankings and strongly rewards features consistently ranked at the top [32].
- **Rank Product:** A non-parametric method that identifies features with a rank product (geometric mean of ranks) that is significantly better than expected by chance.
- **Weighted Borda:** An extension of the Borda count where users can assign custom weights to different analytical methods (e.g., giving SHAP a higher weight than ANOVA).
- **Rank Summation (Total Rank Score):** The simplest baseline method, included for its interpretability, where a "Total Rank Score" is calculated by summing its individual ranks.

This methodological flexibility allows researchers to explore the robustness of their findings under different aggregation assumptions.

To illustrate the fundamental *concept* of rank aggregation using its most straightforward example, consider a hypothetical analysis using the Rank Summation method. Imagine five biomarkers (Gene A to Gene E) analyzed using three different methods, yielding the ranked lists shown in Table II.

TABLE II: Hypothetical ranked outputs from three different analysis methods.

ANOVA Results	SHAP Results	PF Importance Results
1. Gene C	1. Gene A	1. Gene A
2. Gene A	2. Gene C	2. Gene E
3. Gene E	3. Gene B	3. Gene C
4. Gene B	4. Gene D	4. Gene B
5. Gene D	5. Gene E	5. Gene D

To consolidate these rankings using the **Rank Summation** method, BioMark extracts the rank of each gene from each list and aggregates them. The process is detailed in Table III.

TABLE III: A Hypothetical Example of the Rank Summation Methodology. The table shows the individual ranks of five biomarkers from three different methods and the calculated Total Rank Score used for the final consolidated ranking.

Biomarker	ANOVA	SHAP	PF Rank	Total Score
Gene A	2	1	1	4
Gene C	1	2	3	6
Gene E	3	5	2	10
Gene B	4	3	4	11
Gene D	5	4	5	14

After calculating the Total Rank Score for each biomarker, they are sorted from the lowest to the highest score to produce the final consolidated list for this specific method:

1) **Gene A** (Total Score: 4)
2) **Gene C** (Total Score: 6)
3) **Gene E** (Total Score: 10)
4) **Gene B** (Total Score: 11)
5) **Gene D** (Total Score: 14)

In this *example*, Gene A is identified as the most robust biomarker under the Rank Summation method. This demonstrates the core principle of rank aggregation: filtering out biomarkers that appear significant due to the artifacts of one method. This concept of consensus ranking—whether through the simple Rank Sum example shown here or the platform's more advanced default RRF method—increases confidence in the final selected candidates for downstream validation and research.

I. Analysis Report Generation

Upon completion of the analyses, BioMark enables users to generate and download reports that contain all the constructed visualizations and analysis results in an organized format. These reports are typically provided in PDF format. Furthermore, to facilitate further offline analysis and integration with other tools, key results such as feature rankings and statistical test outputs can be downloaded directly as CSV files from the results page. The platform also displays the processing time taken for analyses, which is particularly useful for longer computations.

J. Asynchronous Analysis and Notifications

To enhance the user experience for computationally intensive tasks, which can take several hours (as shown in Table V), BioMark implements an asynchronous analysis and notification system.

When a user initiates an analysis, a heuristic assesses the "duration risk." For high-risk tasks (e.g., XGBoost Feature Importance or optional hyperparameter tuning), a modal window appears, warning the user of the potential runtime. This modal provides an **optional field** for the user to enter their email address.

If an email is provided, the backend assigns a unique job ID and runs the analysis as a background process. This **asynchronous architecture** (detailed in Section III-A) allows the user to safely close their browser window without interrupting the task. Upon completion, the system sends an email notification. This email contains a unique, secure link to a "Results Viewer" page, where the user can access their completed plots and analysis data. This entire notification flow is optional and does not affect the standard in-browser analysis for users who choose not to provide an email.

IV. Results

This section presents the performance evaluation of the BioMark web tool through load tests and analysis time measurements, followed by the analytical findings from its application to two distinct cancer gene expression datasets from literature [33]: prostate cancer and breast cancer. The aim is to demonstrate BioMark's efficiency, scalability, and capabilities in handling high-dimensional biomarker data and extracting meaningful biological insights.

A. Load Test

To comprehensively evaluate the BioMark web tool's performance and scalability, a series of load tests were conducted focusing on the impact of varying dataset sizes under different user loads. The BioMark tool's backend was deployed on a server running **Ubuntu 24.04.1 LTS (kernel 6.8.0-62-generic)** with an **Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz** (112 cores, 2 sockets) and **251 GB of RAM**. The load testing itself was performed using **Locust 2.25.0**, a Python-based load testing tool, executed from a separate client machine to ensure an isolated testing environment.

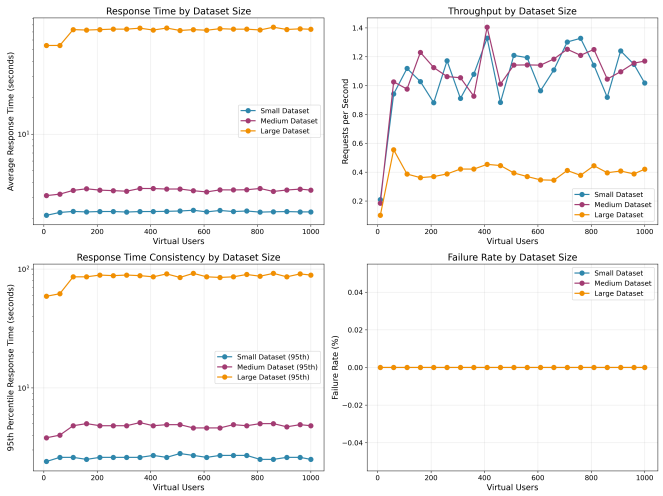
To evaluate the performance and scalability of the data ingestion component of BioMark, which represents the primary user entry point and a potential performance bottleneck, our load test focused on a specific, high-traffic scenario: the concurrent upload of user datasets. Each virtual user simulated this core action to measure the system's response under varying data loads. This scenario was executed using three distinct dataset sizes, which were subsampled from a single, real-world public dataset to ensure a consistent data structure across the tests. The source data was obtained from the NCBI Gene Expression Omnibus (GEO) repository under accession number **GSE120584** [34], corresponding to a human miRNA expression study for dementia risk prediction. The three derived datasets were as follows:

- **Small Dataset:** 20 samples, 10 molecules (file size: 2.011 kB).
- **Medium Dataset:** 200 samples, 100 molecules (file size: 234.366 kB).
- **Moderately-sized dataset:** 1000 samples, 1000 molecules (file size: 11.905 MB).

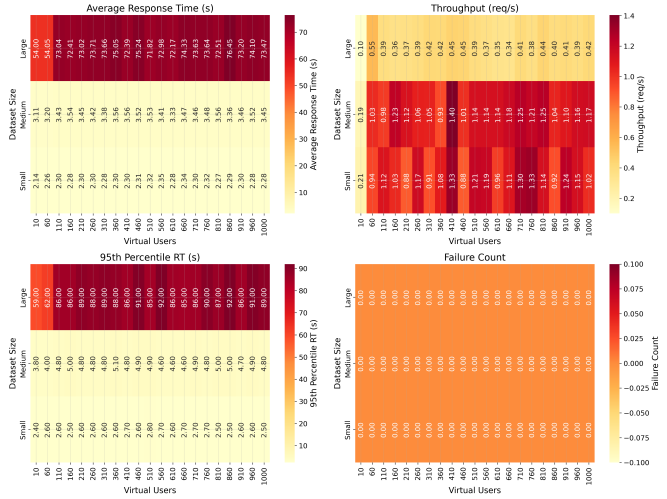
Each test was executed for a duration of 30 seconds, while varying the number of concurrent Virtual Users (VUs) from 10 to 1000. This approach allowed for a detailed assessment of how dataset size influences key performance metrics across a wide range of user loads.

The results, comprehensively summarized in Figure 2 and Table IV, highlight the critical impact of dataset size on BioMark's performance. Across all dataset sizes and virtual user loads, the system consistently demonstrated robust stability with a **0% failure rate**. This indicates BioMark's reliability in handling file uploads without errors, even under stress.

Response Time Performance: As depicted in Figure 2a (top-left panel), average response times varied significantly with dataset size. The *Small Dataset* consistently yielded the fastest response times, remaining below 5 seconds even at 1000 VUs (e.g., 2.28s at 1000 VUs, as shown in Table IV). The *Medium Dataset* showed a moderate increase, with average response times typically between 20-30 seconds (e.g., 3.5s at 1000 VUs). The *Moderately-sized dataset* presented the highest response times, averaging around 70-80 seconds (e.g., 73.5s at 1000 VUs). This direct correlation is further exemplified in Table IV, which summarizes performance at low and high user loads. At 1000 VUs, the average response time for the Moderately-sized dataset (73.47 s) is drastically higher than for the Medium (3.47 s) and Small (2.28 s) datasets, clearly demonstrating the performance impact of data size.



(a) Performance Metrics (Response Time and Throughput)



(b) Heatmap Distribution of Metrics

Fig. 2: Performance evaluation of BioMark. (a) Detailed comparison of average response times, throughput, 95th percentile response times, and failure rates. (b) Heatmap visualizations illustrating the distribution of metrics across different dataset sizes and virtual user loads.

The 95th percentile response times (Figure 2a, bottom-left panel, and Figure 2b, bottom-left panel) also show a similar trend, indicating that the impact of dataset size affects the vast majority of user experiences.

Throughput Capacity: Figure 2a (top-right panel) illustrates the throughput performance. The *Small Dataset* and *Medium Dataset* achieved similar peak throughput values, ranging from 0.8 to 1.4 requests/second, and demonstrated robust scaling behavior as user loads increased. In contrast, the *moderately-sized dataset* exhibited significantly lower throughput, stabilizing around 0.4-0.5 requests/second. This suggests that while BioMark can process larger files, the per-second processing capacity is substantially reduced due to the increased computational overhead associated with larger data volumes.

TABLE IV: Summary of Load Test Performance Metrics at Low, Medium, and High Virtual User (VU) Loads for Each Dataset Size.

Dataset Size	Virtual Users	Avg. Response Time (s)	95th Percentile RT (s)	Throughput (req/s)
Small	60	2.26	2.6	0.94
	510	2.30	2.6	1.14
	1000	2.28	2.5	1.17
Medium	60	3.20	4.0	1.02
	510	3.53	4.9	1.14
	1000	3.47	4.8	1.17
Moderately-sized	60	54.05	59.0	0.55
	510	71.82	85.0	0.39
	1000	73.47	91.0	0.38

The load test results underscore that BioMark effectively handles varying data sizes, maintaining 0% failure rates across all tested conditions. However, performance metrics like response time and throughput are heavily influenced by the size of the uploaded dataset. Small and medium datasets are processed efficiently, making BioMark highly responsive for typical biomarker panels. For large, high-dimensional omics datasets, while the system remains reliable, processing times naturally increase, indicating a need for careful consideration of computational resources and potential optimizations for extremely large files in future developments. These findings provide crucial insights for capacity planning and user expectations when analyzing diverse omics data with BioMark.

B. Analysis Time Evaluation

This subsection details the execution times for various computationally intensive analytical tasks performed by BioMark, highlighting the tool's efficiency in processing complex biomarker data. The analysis times were recorded on the same server environment used for load testing.

Analysis time was evaluated for critical operations such as differential analysis, dimensionality reduction tasks, and classification across two distinct datasets: the prostate cancer dataset (137 samples, 27,828 features) and the breast cancer dataset (108 samples, 16,382 features). The detailed execution times for each analysis type are presented in Table V.

Table V demonstrates that BioMark efficiently processes large and complex omics data within practical timeframes for many critical operations. For instance, most dimensionality reduction tasks (e.g., PCA) and simpler classification models (e.g., Logistic Regression, Decision Tree, SVC) complete within seconds, providing rapid insights. Differential analysis methods like ANOVA and T-test also execute quickly.

However, certain computationally intensive machine learning methods, particularly in Model Explanation (e.g., XGBoost Feature Importance, Permutation Feature Importance) and classification (e.g., CatBoosting Classifier), require significantly extended durations. For the prostate cancer dataset, XGBoost Feature Importance took approximately 9.4 hours

TABLE V: Analysis Execution Times for Prostate and Breast Cancer Datasets

Analysis Type	Prostate	Breast
Statistical Analysis		
ANOVA	7 s	7 s
T-test	20 s	15 s
Dimensionality Reduction and Visualization		
PCA	10 s	10 s
t-SNE	12 s	11 s
UMAP	13 s	11 s
Classification Analysis (Training & Native Importance)		
Logistic Regression	11 s	9 s
Random Forest	8m 30s	4m 58s
RF Native Feat. Imp.	*(incl. in 8m 30s)*	*(incl. in 4m 58s)*
XGBClassifier	9h 23m 07s	2h 57m
XGB Native Feat. Imp.	*(incl. in 9h 23m 07s)*	*(incl. in 2h 57m)*
Decision Tree	11 s	9 s
Gradient Boosting	3m 16s	1m 29s
CatBoost Classifier	1h 19m	23m
AdaBoost Classifier	55 s	29 s
MLPClassifier	17 s	15 s
SVC	11 s	10 s
Model-Agnostic Explanation (using trained models)		
SHAP (with XGBClassifier)	33 s	18 s
LIME (with Random Forest)	73 s	42 s
Permutation Feat. Imp. (with Random Forest)	2h 28m 18s	51m 47s

(33787 s), and Permutation Feature Importance took around 2.5 hours (8898 s). Similarly, CatBoosting Classifier required about 1.3 hours (4741 s). While these durations are substantial, they are typical for high-dimensional feature importance calculations and complex model training on large datasets. The breast cancer dataset, being smaller, generally resulted in faster execution times for these computationally demanding tasks (e.g., XGBoost Feature Importance: 10620 s $\approx$ 2.9 hours; CatBoosting Classifier: 1381 s $\approx$ 23 minutes).

These measurements highlight BioMark's capability as a versatile tool for biomarker discovery and validation. While common analytical tasks are highly responsive, users should anticipate longer processing times for advanced, computationally intensive algorithms on high-dimensional omics data, especially for larger datasets.

C. Case Study on Prostate Cancer

This section details the application of the BioMark web tool to a prostate cancer transcriptomics dataset (137 samples, 27,828 gene transcripts, where 'Factors' column includes cancer/healthy status).

1) Analysis Results:

a) Biomarker Identification and Analysis: To identify statistically significant biomarkers capable of distinguishing between prostate cancer and healthy samples, we employed a comprehensive approach using BioMark's analytical modules. This process leverages both traditional **Statistical Analysis** and advanced **Model Explanation** techniques, providing a multi-faceted view of the molecular signatures driving the disease.

The analysis began with classical statistical tests like ANOVA to identify biomarkers with significant mean expression differences between the cancer and healthy cohorts. As shown in the ANOVA results (Fig. 4a), biomarkers such as *LOC105374013*, *DANCR*, and *PCA3* demonstrated high F-values, indicating strong statistical power in differentiating the two groups based on their expression levels alone.

To gain deeper, model-driven insights, we employed model-based explainability techniques. While BioMark's architecture is **model-agnostic**—allowing SHAP and LIME to be applied to any classifier trained by the platform (e.g., XGBoost, SVC, etc.)—for the specific analyses in this case study, we applied **SHAP** to interpret the predictions of an internally trained **XGBoost classifier** and **LIME** to explain the predictions of a **Random Forest model**. Unlike ANOVA, which assesses each feature in isolation, these methods evaluate a biomarker's contribution within the context of a predictive model that learns complex interactions. The SHAP Summary Plot (Fig. 4b) offers a global perspective on feature importance. For instance, it reveals that high expression of biomarkers like *NEK5* and *LOC105374013* (indicated by red dots) consistently yields high positive SHAP values, pushing the model's prediction towards cancer. Conversely, low expression of these same markers (blue dots) contributes to a "healthy" prediction. This visualization powerfully illustrates not only the importance of each biomarker but also the direction and magnitude of its effect across all samples.

This global view is complemented by instance-level explanations. The SHAP Heatmap (Fig. 4b) visualizes the collective behavior of SHAP values across the entire dataset, clearly delineating the cancer and healthy cohorts through distinct biomarker contribution patterns. For a more granular view, SHAP Waterfall plots (Fig. 4d) deconstruct the prediction for individual samples. They illustrate precisely how each biomarker contributes to shifting the model's output from the baseline to the final prediction for both a representative cancer sample and a healthy one. For example, in the cancer sample, high expression of *RUSC1* and *NEK5* are shown as primary drivers of the cancer classification.

Similarly, the LIME plot (Fig. 4e) provides a local explanation for a single cancer sample, highlighting that high expression of *STAT6* and *MCM8* were key contributors to its specific classification. The slight differences in top biomarkers between global methods (SHAP Summary, ANOVA) and local methods (LIME, SHAP Waterfall) underscore the biological heterogeneity within the cancer group and demonstrate BioMark's capacity to capture both population-level trends and patient-specific molecular events. This dual approach, combining robust statistical validation with nuanced model-based interpretation, facilitates a more comprehensive identification

of clinically relevant biomarkers.

To further justify the utility of the rank aggregation strategy, we analyzed the concordance between different analytical modules. Figure 3 illustrates the overlap between the top 20 biomarkers identified by ANOVA (statistical), SHAP (model-based), and the final aggregated list. While individual methods yielded distinct subsets of candidates due to their differing mathematical assumptions, the aggregated list successfully captured the consensus features—those lying at the intersection of statistical significance and predictive power—thereby filtering out potential method-specific artifacts.

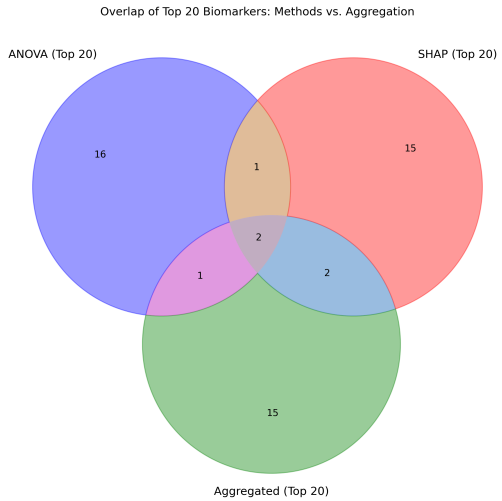


Fig. 3: Venn diagram visualizing the overlap between the top 20 features identified by ANOVA, SHAP, and the final Aggregated Rank (Rank Summation). The diagram demonstrates how the aggregation strategy captures consensus features shared across methods (intersections) while filtering out biomarkers that are isolated artifacts of a single statistical approach.

2) Literature-Based Validation of Identified Biomarkers: A key measure of BioMark's efficacy is its ability to identify biomarkers that are corroborated by existing scientific literature. Our analysis successfully pinpointed a range of molecules with well-documented roles in prostate cancer, thus validating our platform's analytical pipeline. For instance, top-ranked features identified by BioMark through its biomarker combination feature (Fig. 5) such as the lncRNAs **DANCR** and **PCAT7** are strongly supported by previous studies, which confirm their roles in promoting invasion and metastasis [35], [36]. Similarly, our tool highlighted **NEK5**, whose elevated expression is known to support tumor growth [37]; **KLF8**, a co-activator for the androgen receptor linked to poor survival [38]; **CRISP3**, a marker for a particularly aggressive molecular subtype of prostate cancer [39]; and the tumor suppressor **SERPINB5** (Maspin), whose loss is associated with cancer progression [40]. Alongside these established markers, BioMark also prioritized candidates like *SNORD12B* and *SNORD105B*, for which direct links to prostate cancer are not yet extensively documented. This demonstrates BioMark's dual capability: robustly confirming known, clinically relevant biomarkers while simultaneously generating novel hypotheses for future experimental investigation.

Beyond their individual validations, the top-ranked biomarkers identified by BioMark show a functional convergence on pathways driving tumor aggressiveness and metastasis. Specifically, the selection of long non-coding RNAs DANCR and PCAT7, along with the transcription factor KLF8, points to a deregulation of the Epithelial-to-Mesenchymal Transition (EMT) machinery, a critical process for cancer cell invasion. This signature is further reinforced by the inclusion of NEK5, a kinase essential for cell cycle progression. Collectively, these findings suggest that the BioMark-derived signature does not merely represent random distinct genes, but rather captures a cohesive biological phenotype characterized by enhanced proliferative capacity and the acquisition of invasive metastatic potential

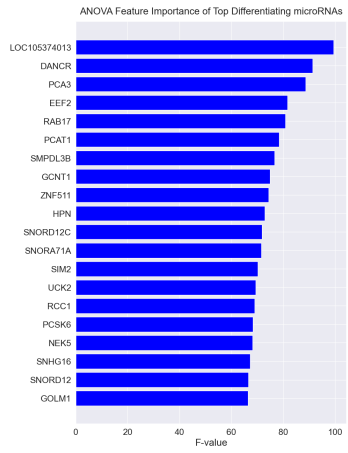
a) Dimensionality Reduction and Visualization: Dimensionality reduction and visualization were performed to investigate the inherent structure of the prostate cancer dataset and to explore natural groupings among samples based on their biomarker profiles. This unsupervised approach utilizes dimensionality reduction techniques to visualize high-dimensional data in a lower-dimensional space, revealing patterns without prior knowledge of the sample labels (cancer or healthy). The analyses were conducted "After Feature Selection," meaning they were applied to a refined subset of the most informative biomarkers. This crucial preliminary step, as detailed in Section III-F1, reduces data noise and computational complexity, allowing the algorithms to focus on the strongest biological signals and thus produce clearer, more meaningful visualizations.

To achieve this, we employed three powerful dimensionality reduction methods: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP), all offered in BioMark. The results, visualized in 3D space in Figure 6, consistently demonstrate a striking separation between the cancer and healthy samples.

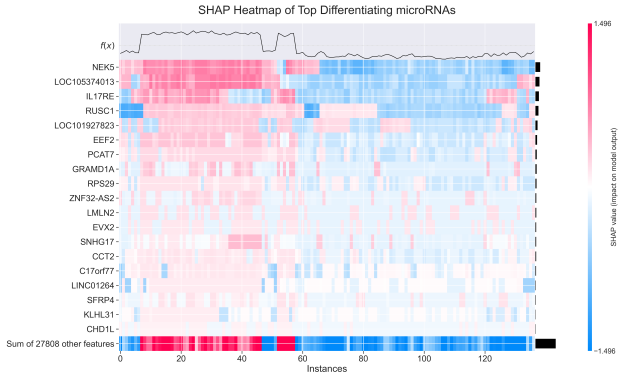
As a linear method, PCA identifies the axes of maximum variance in the data. The resulting PCA plot (Fig. 6a) shows a clear and effective separation, with the healthy samples forming a tight, distinct cluster and the cancer samples forming another, more dispersed group. This indicates that a significant linear component of the variation in the biomarker data is sufficient to distinguish the two groups.

To capture more complex, non-linear relationships, we applied t-SNE and UMAP. The t-SNE method, which excels at preserving local neighborhood structures, also revealed a very strong separation between the two classes (Fig. 6b), further confirming that the groups are distinct even when considering non-linear patterns. Similarly, UMAP, which effectively balances the preservation of both local and global data structure, provided the most striking visualization (Fig. 6c). It produced exceptionally compact and well-defined clusters with a pronounced margin between them, reinforcing the robustness of the identified biomarker signature.

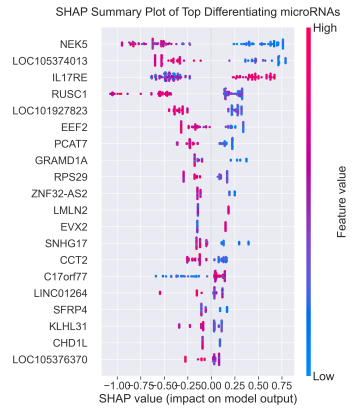
The convergence of these three distinct methods—linear (PCA) and non-linear (t-SNE, UMAP)—on the same conclu-



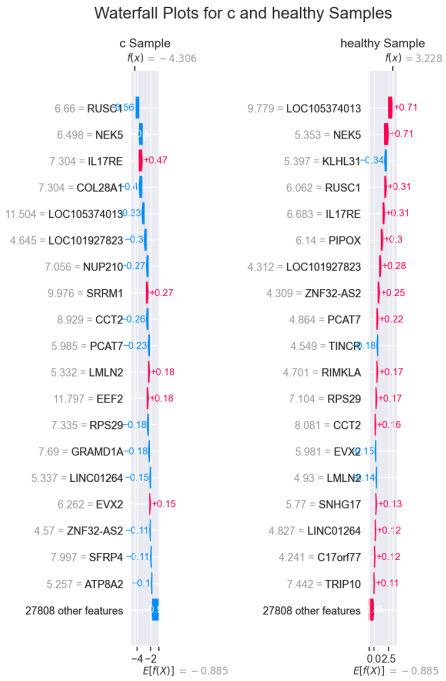
(a) ANOVA Feature Importance



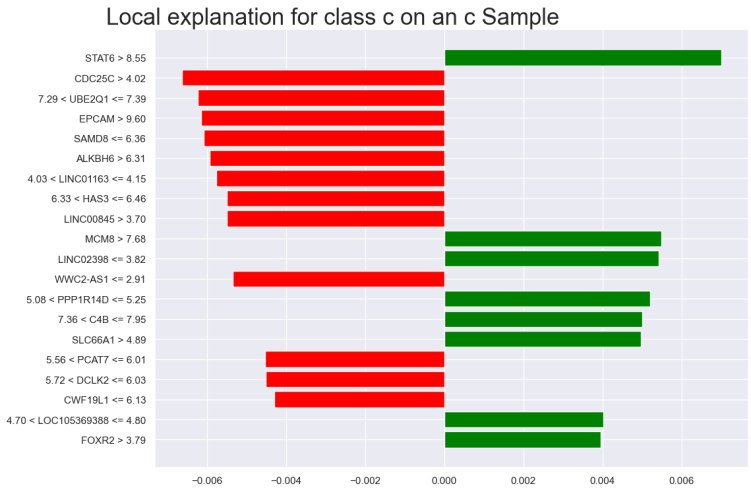
(b) Heatmap of Top Differentiating Biomarkers



(c) SHAP Summary Plot



(d) SHAP Waterfall Plot



(e) LIME Local Explanation

Fig. 4: Comprehensive biomarker explainability analysis for Prostate Cancer. (a) Top discriminative biomarkers detected by ANOVA. (b) Global SHAP summary illustrating feature impact. (c)Heatmap of Top Differentiating Biomarkers in Prostate Cancer: Displaying expression patterns of top 19 biomarkers across cancer and healthy samples. (d) SHAP Waterfall plots contrasting prediction logic. (e) LIME plot showing local feature contributions.

sion provides powerful evidence for the validity of the selected biomarkers. The ability of these unsupervised techniques to so clearly partition the samples based on their biological status underscores the fundamental transcriptomic differences between cancerous and healthy prostate tissue.

b) Classification Analysis and Predictive Model Performance: To assess the diagnostic potential of the identified biomarker signature, a comprehensive classification analysis was performed using the Biomark tool. Various machine learning models were trained to distinguish between cancer and healthy samples using the transcriptomic profiles. The performance of these models was systematically evaluated with stratified k-fold cross validation and is presented in Table VI. As detailed in Section III-F1, the analysis was conducted in two stages: first using the "Without Feature Selection" approach on the entire high-dimensional feature set, and subsequently using the "After Feature Selection" approach on a refined subset of the most informative biomarkers. The results clearly demonstrate the strength of the identified biomarker set in building accurate predictive models from omics data. When models were trained on the full dataset, their performance was moderate, reflecting the challenge of learning from noisy, high-dimensional data. However, applying feature selection to focus on identified key biomarkers led to

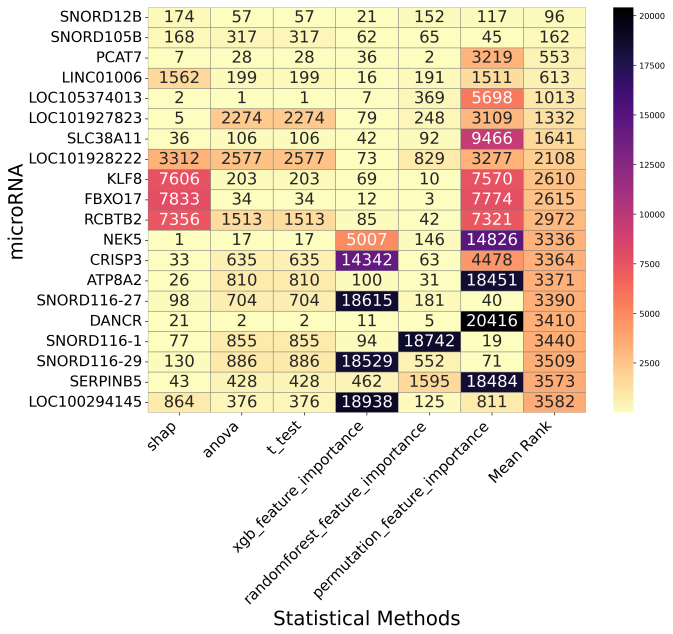


Fig. 5: Consolidated ranking (via Rank Summation) of the top 20 differentiating biomarkers in Prostate Cancer. The plot summarizes feature importance scores from multiple methods (SHAP, LIME, ANOVA, etc.) to provide a robust consensus list of molecules distinguishing cancer samples from healthy controls.

a dramatic and universal improvement in all metrics evaluated for almost every model. The Support Vector Classifier (SVC) emerged as the top-performing model after biomarker selection, achieving an impressive accuracy of 0.94 and a robust F1-score of 0.88. Other models, such as Logistic Regression and the MLP Classifier, also showed excellent performance, with accuracies of 0.92 and 0.90, respectively. This significant boost in predictive power strongly validates the diagnostic relevance of the selected biomarker panel and showcases BioMark's effectiveness in facilitating an end-to-end workflow, from biomarker discovery to the construction of high-performance diagnostic tools.

D. Case Study on Breast Cancer

This section details the application of the BioMark web tool to a breast cancer biomarker dataset (108 samples, 16,382 features, 'Factors' column for cancer/healthy status), demonstrating the tool's adaptability across varying data characteristics.

1) Analysis Results:

a) Biomarker Identification and Analysis: To demonstrate BioMark's adaptability, the same comprehensive analytical workflow was applied to the breast cancer dataset. The process again combined traditional **Statistical Analysis** with **Model Explanation** techniques to uncover the key molecular drivers distinguishing cancer from healthy breast tissue samples.

Statistical analysis with ANOVA (Fig. S5 in Sppl. Material) identified biomarkers with the most significant differences in mean expression between the groups. Features such as

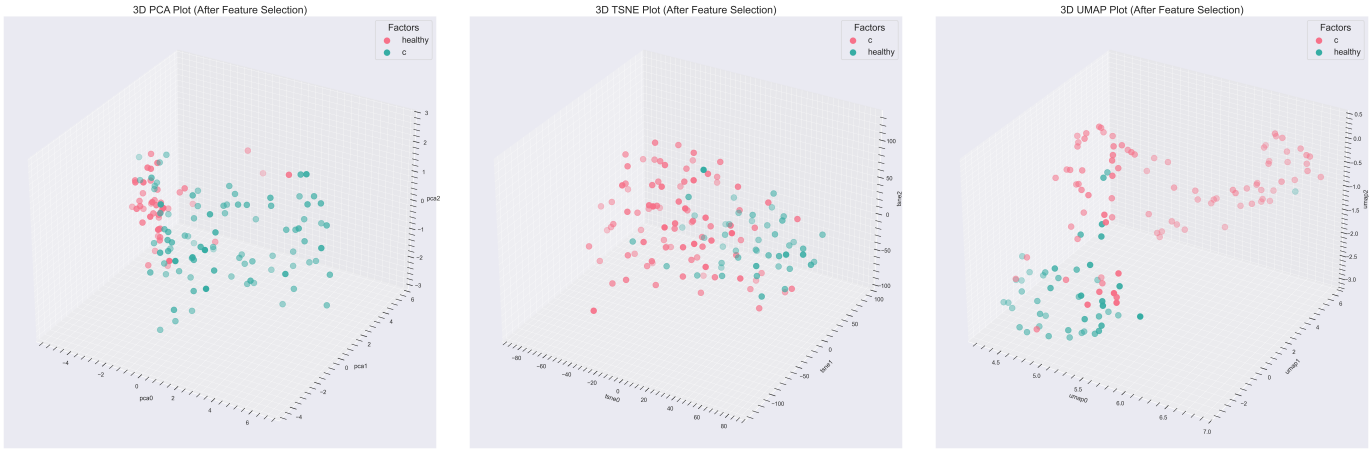
TABLE VI: Classification Model Performance Metrics for Prostate Cancer Diagnosis

Without Feature Selection				
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.78	0.59	0.97	0.72
Random Forest	0.80	0.71	0.48	0.56
XGBClassifier	0.82	0.72	0.61	0.66
Decision Tree	0.75	0.55	0.67	0.59
Gradient Boosting	0.82	0.68	0.74	0.69
CatBoosting Classifier	0.81	0.83	0.41	0.53
AdaBoost Classifier	0.81	0.72	0.60	0.62
MLPClassifier	0.69	0.49	0.93	0.64
SVC	0.77	0.62	0.48	0.53
After Feature Selection				
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.92	0.86	0.87	0.84
Random Forest	0.88	0.81	0.77	0.78
XGBClassifier	0.89	0.82	0.80	0.79
Decision Tree	0.80	0.69	0.58	0.62
Gradient Boosting	0.80	0.63	0.70	0.66
CatBoosting Classifier	0.90	0.79	0.87	0.82
AdaBoost Classifier	0.88	0.78	0.80	0.79
MLPClassifier	0.90	0.80	0.90	0.83
SVC	0.94	0.88	0.90	0.88

VIT, *SEMA3G*, *PDE2A*, and *KL* ranked highest based on their F-values, establishing their strong statistical relevance in separating the cohorts.

For more nuanced, model-driven insights, we employed model-based explainability techniques. While BioMark's architecture is **model-agnostic**—allowing SHAP and LIME to be applied to any classifier trained by the platform (e.g., XGBoost, SVC, etc.)—for the specific analyses in this case study, we applied **SHAP** to interpret the predictions of an internally trained **XGBoost classifier** and **LIME** to explain the predictions of a **Random Forest model**. The SHAP Summary Plot (Fig. S6 in Sppl. Material) provides a global overview of the features most impactful to the model. Notably, for top biomarkers like *KL* and *ZHX3*, low expression levels (blue dots) are associated with high positive SHAP values, indicating that their downregulation is a strong predictor of breast cancer. This contrasts with the patterns observed in the prostate cancer dataset and highlights the disease-specific nature of biomarker signatures. The SHAP Heatmap (Fig. S1 in the Suppl. Material) visualizes this phenomenon across all samples, revealing a clear separation where the cancer cohort (right side) is characterized by patterns of biomarker contributions that are distinct from the healthy cohort (left side).

Instance-level explanations further clarify these relationships. The SHAP Waterfall plots (Fig. S7 in Suppl. Material) elegantly contrast the model's reasoning for a healthy versus a cancer sample. For the healthy sample, high expression of *KL* and *ZHX3* strongly contributes to a "healthy" prediction. Conversely, for the cancer sample, low expression of these same biomarkers are the primary drivers of the "cancer" classification. The LIME plot (Fig. S8 in Suppl. Material) offers a similar local validation for a single healthy sample, showing that high expression of *PCYOX1* and appropriate



(a) PCA (Linear Separation) (b) t-SNE (Local Structure) (c) UMAP (Global & Local Structure)

Fig. 6: Dimensionality reduction visualizations for Prostate Cancer "After Feature Selection". (a) 3D PCA plot showing linear separation axes. (b) 3D t-SNE plot revealing local neighborhood clusters. (c) 3D UMAP plot demonstrating the most distinct separation between cancer and healthy groups.

levels of other markers like *ABCA6* and *DENND2D* solidified its correct classification. By integrating these multi-faceted insights, BioMark successfully identified significant biomarkers like *DENND2D*, *PYROXD2*, and *DPT*, and elucidated their complex predictive roles, underscoring the platform's power in extracting context-rich biological information from diverse datasets.

2) Literature-Based Validation of Identified Biomarkers: To further validate our tool's performance, the biomarkers identified in the breast cancer dataset were cross-referenced with published findings, revealing a strong overlap with established research across diverse biological functions. BioMark successfully flagged multiple genes with confirmed roles in breast cancer pathology. For example, **DPT** (Dermatopontin) is a known tumor suppressor whose downregulation is linked to malignancy [41]. **IKBKE** and **PARVA** were also identified, and both are documented to promote invasion and metastasis [42], [43]. Additionally, our tool prioritized **CHEK1**, a crucial marker in triple-negative breast cancer [44]; **MPPED2**, a tumor suppressor silenced by hypermethylation [45]; and **CRY2**, a circadian clock gene whose low expression correlates with aggressiveness and poor survival [46]. The congruence between BioMark's automated analysis and this broad spectrum of manually curated literature findings powerfully underscores the tool's reliability and precision. The identification of less-studied features like *PYROXD2* further highlights its capacity to pinpoint novel candidates worthy of future research.

Biologically, the top markers prioritized in the breast cancer analysis reflect a distinct disruption of tumor suppressive mechanisms and cell adhesion dynamics. The strong predictive value of downregulated KL (Klotho) and DPT (Dermatopontin) highlights the loss of critical tumor suppressor functions associated with aging and extracellular matrix integrity. Conversely, the identification of IKBKE and PARVA implicates the activation of inflammatory signaling and cytoskeletal reorganization. This functional grouping suggests that BioMark successfully isolated a molecular profile defined by the dis-

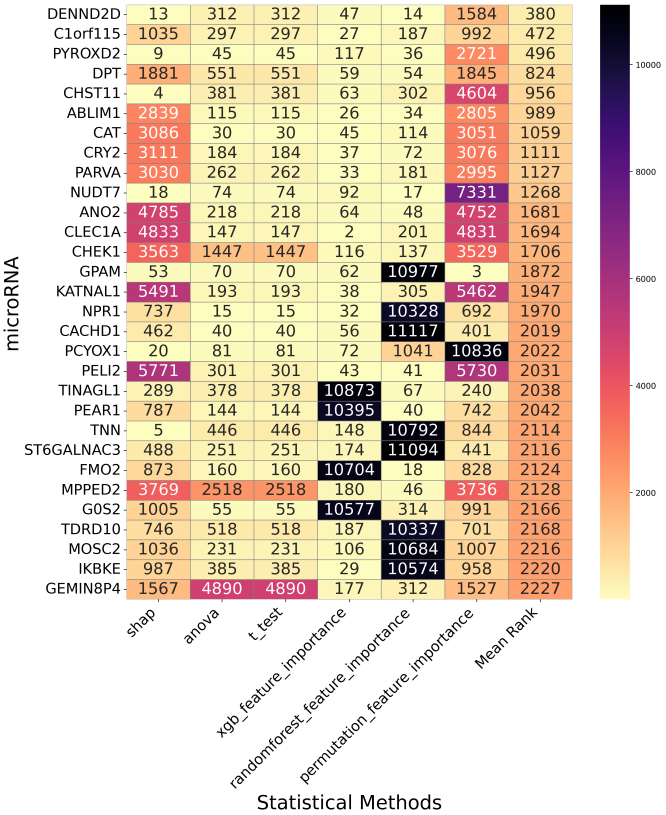


Fig. 7: Consolidated (via Rank Summation) ranking of the top 30 differentiating biomarkers in Breast Cancer. The plot summarizes feature importance scores from multiple methods (SHAP, LIME, ANOVA, etc.) to provide a robust consensus list of molecules distinguishing cancer samples from healthy controls.

mantling of protective cellular barriers and the activation of pathways prerequisite for migration and malignancy.

a) Dimensionality Reduction and Visualization: The dimensionality reduction and visualization pipeline was subsequently applied to the breast cancer dataset to evaluate its inherent structure and the separability of the cancer and healthy groups. Following the protocol detailed in Section III-F1, PCA, t-SNE, and UMAP visualizations were generated after feature selection. In striking contrast to the prostate cancer results, the breast cancer samples exhibited a greater degree of inter-group similarity, making their separation more challenging.

The resulting visualizations are presented in the **Supplementary Material**. The PCA plot (**Fig. S2**) revealed only a partial separation between the cancer and healthy samples. While a general trend is visible, there is significant overlap between the two groups. Similarly, the t-SNE analysis (**Fig. S3**), which focuses on local data structure, also resulted in a largely mixed distribution of samples, failing to resolve them into distinct clusters.

However, the UMAP analysis proved to be the most effective method for this more complex dataset (**Fig. S4**). Despite the challenges, UMAP successfully partitioned the majority of the samples into two discernible clusters. This demonstrates UMAP's strength in preserving a more meaningful balance of local and global data structure. The ability of BioMark to still uncover underlying data structures, even in a dataset with less distinct separation, underscores its utility for patient stratification

b) Classification Analysis and Predictive Model Performance: Predictive models for breast cancer diagnosis were trained using the same array of machine learning algorithms as for prostate cancer, with their performance evaluated through cross-validation. As with the prostate cancer analysis, model performance was compared between the "Without Feature Selection" approach (using the full feature set) and the "After Feature Selection" approach (using the refined biomarker subset), summarized in Table VII. In contrast to the prostate cancer case, a couple of models (i.e., LR and SVC) for breast cancer performed strongly even without feature selection, suggesting a robust initial biomarker signature. However, the application of feature selection was still crucial for optimizing performance for most models. While some models showed little change, others improved significantly. Notably, the AdaBoost Classifier's performance surged after feature selection, reaching the highest accuracy of 0.90 and the best F1-score of 0.88. Other models like Random Forest and CatBoosting Classifier also saw their F1-scores improve to 0.85 after feature selection. These results underscore that even for a strong initial set of biomarkers, the feature selection workflow within BioMark is invaluable for identifying the optimal model and achieving the highest possible diagnostic accuracy.

The classification models, particularly the CatBoosting Classifier, achieved high predictive performance in predicting breast cancer status, underscoring BioMark's robustness and broad applicability for biomarker-driven classification tasks across different cancer datasets.

TABLE VII: Classification Model Performance Metrics for Breast Cancer Diagnosis

Without Feature Selection				
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.86	0.81	0.90	0.85
Random Forest	0.84	0.87	0.76	0.80
XGBClassifier	0.80	0.80	0.76	0.77
Decision Tree	0.78	0.73	0.84	0.77
Gradient Boosting	0.79	0.75	0.84	0.78
CatBoosting Classifier	0.84	0.84	0.79	0.81
AdaBoost Classifier	0.78	0.79	0.73	0.74
MLPClassifier	0.85	0.83	0.84	0.83
SVC	0.84	0.80	0.84	0.81
After Feature Selection				
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.85	0.84	0.82	0.82
Random Forest	0.87	0.85	0.87	0.85
XGBClassifier	0.85	0.82	0.84	0.83
Decision Tree	0.80	0.77	0.78	0.77
Gradient Boosting	0.83	0.80	0.78	0.79
CatBoosting Classifier	0.87	0.85	0.87	0.85
AdaBoost Classifier	0.90	0.87	0.89	0.88
MLPClassifier	0.86	0.84	0.84	0.84
SVC	0.80	0.78	0.79	0.78

V. Limitations and Future Directions

While BioMark lowers the barrier to advanced biomarker discovery, we acknowledge several limitations in the current implementation. First, regarding data preprocessing, the platform is designed primarily for downstream analysis of processed data. Consequently, it does not currently include upstream quality control (QC) modules for sample-level outlier detection or automated batch-effect correction. Users are advised to perform normalization and batch correction prior to uploading data to ensure the validity of downstream statistical inferences.

Second, the statistical analysis module currently relies on parametric tests (T-test and ANOVA). While robust for many omics datasets, these methods may not be optimal for data that strictly violates normality assumptions. Future updates will incorporate non-parametric alternatives, such as the Wilcoxon rank-sum test and Kruskal-Wallis test, to broaden the tool's applicability.

Third, the current scope of the platform is limited to binary classification tasks (e.g., Case vs. Control). Datasets requiring multi-class outcome prediction are not supported in the present version. Additionally, the pipeline does not explicitly apply automated resampling techniques (such as SMOTE or ADASYN) to handle severe class imbalance. Users with heavily skewed class distributions should therefore interpret classification metrics with caution.

Fourth, regarding biological interpretation, the current version focuses on identifying and ranking molecular signatures but does not yet feature built-in pathway or Gene Ontology (GO) enrichment analysis. Integrating these databases (e.g., KEGG, Reactome) is a priority for the next major release to provide automated biological context.

Finally, as indicated in our load testing (Section IV-A), scalability remains a consideration. While the asynchronous archi-

tecture handles standard datasets efficiently, execution times for computationally intensive tasks—specifically permutation-based feature importance—can be substantial for very large, high-dimensional datasets.

Future development work will focus on addressing these limitations, specifically by extending support to multi-class problems, integrating survival analysis modules, and adding the aforementioned non-parametric and enrichment capabilities.

VI. Conclusion

In this paper, we introduced BioMark, a web-based platform designed to bridge the gap between complex high-throughput omics data and researchers seeking meaningful biological insights. By integrating robust statistical methods, machine learning algorithms, and intuitive visualizations into a single, user-friendly environment, BioMark successfully lowers the barrier to advanced biomarker analysis.

Our case studies on prostate and breast cancer datasets demonstrated the platform's practical utility. BioMark not only identified numerous biomarkers validated by existing literature but also constructed highly predictive classification models with accuracies reaching up to 94%. This highlights its dual capability as both a reliable discovery tool and a powerful diagnostic model builder. Furthermore, the platform's ability to pinpoint under-investigated yet statistically significant molecules showcases its potential for generating novel, data-driven hypotheses for translational research. By continuing to enhance its capabilities—particularly by addressing the limitations outlined in Section V—BioMark can serve as a valuable resource for the scientific community, empowering a broader range of researchers to accelerate the pace of biomarker-driven medicine.

VII. Declarations

A. Ethics approval and consent to participate

Not applicable.

B. Consent for publication

Not applicable.

C. Availability of data and materials

The source codes are available at the following repository: https://github.com/itu-bioinformatics-database-lab/biomark.

For case studies and performance evaluations, the employed datasets are publicly available at the following links:

- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120584
- https://doi.org/10.5281/zenodo.7874228

A demo video is available on the "User Guide" page of the Biomark website.

D. Availability and Requirements

Project name: BioMark
Project home page: https://bioinf.itu.edu.tr/biomark
Operating system(s): Platform independent
Programming language: Node.js (JavaScript) and Python
Other requirements: Python 3.8 or higher
License: GNU GPL
Any restrictions to use by non-academics: License needed.

E. Competing interests

None.

F. Funding

This work was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) – EU Joint Programme - Neurodegenerative Disease Research (JPND), under Grant 124N069, by the Scientific Research Projects Unit of Istanbul Technical University (ITU BAP) [Grant No. TGA-2025-46998], and by the National Center for High-Performance Computing (UHEM) [Grant Number: 1009742021].

G. Authors' contributions

MAB implemented the frontend, CNM implemented the backend, AC conceived the study and acquired the funding. All authors contributed to the manuscript's writing.

References

[1] H. Zheng, G. Zhang, L. Zhang, Q. Wang, H. Li, Y. Han, L. Xie, Z. Yan, Y. Li, Y. An, H. Dong, W. Zhu, and X. Guo, "Comprehensive review of web servers and bioinformatics tools for cancer prognosis analysis," *Frontiers in Oncology*, vol. 10, p. 68, 2020.

[2] J. Xia, N. Psychogios, N. Young, and D. S. Wishart, "MetaboAnalyst: a web server for metabolomic data analysis and interpretation," *Nucleic Acids Research*, vol. 37, no. suppl_2, pp. W652–W660, 2009.

[3] J. Chong and J. Xia, "MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data," *Bioinformatics*, vol. 34, no. 24, pp. 4313–4314, 2018.

[4] A. Mendes, J. F. Havelund, J. Lemvig, V. Schwämmle, and N. J. Færgeman, "MetaboLink: a web application for streamlined processing and analysis of large-scale untargeted metabolomics data," *Bioinformatics*, vol. 40, no. 7, p. btae459, 2024.

[5] S. Cardoso, T. Afonso, M. Maraschin, and M. Rocha, "WebSpecmine: A website for metabolomics data analysis and mining," *Metabolites*, vol. 9, no. 10, p. 237, 2019.

[6] L. Madrid-Márquez, C. Rubio-Escudero, B. Pontes, A. González-Pérez, J. C. Riquelme, and M. E. Sáez, "MOMIC: A multi-omics pipeline for data analysis, integration and interpretation," *Applied Sciences*, vol. 12, no. 8, p. 3987, 2022.

[7] A. Gruca, J. Henzel, I. Kostorz, T. Stęclik, Ł. Wróbel, and M. Sikora, "MAINE: a web tool for multi-omics feature selection and rule-based data exploration," *Bioinformatics*, vol. 38, no. 6, pp. 1773–1775, 2022.

[8] Y. Lu, M. Shen, L. Yue, C. Li, L. Chen, X. Wang, D. Herrington, Y. Wang, Y. Zhao, T. Fu, and C. Van Rechem, "GenoCraft: A comprehensive, user-friendly web-based platform for high-throughput omics data analysis and visualization," *arXiv*, 2024.

[9] J. Zoppi, J.-F. Guillaume, M. Neunlist, and S. Chaffron, "MiBiOmics: an interactive web application for multi-omics data exploration and integration," *BMC Bioinformatics*, vol. 22, no. 1, p. 6, 2021.

[10] G. Camele, S. Menazzi, H. Chanfreau, A. Marraco, W. Hasperué, M. D. Butti, and M. C. Abba, "Multiomix: a cloud-based platform to infer cancer genomic and epigenomic events associated with gene expression modulation," *Bioinformatics*, vol. 38, no. 3, pp. 866–868, 2022.

[11] M. Lawrence, E.-K. Lee, D. Cook, H. Hofmann, and E. Wurtele, "exploRase: Exploratory data analysis of systems biology data," in *Proceedings of the Fourth International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV'06)*. IEEE, 2006, pp. 40–79.

[12] R. Tautenhahn, G. J. Patti, D. Rinehart, and G. Siuzdak, "XCMS Online: A web-based platform to process untargeted metabolomic data," *Analytical Chemistry*, vol. 84, no. 11, pp. 5035–5039, 2012.

[13] M. Akhmedov, A. Martinelli, R. Geiger, and I. Kwee, "Omics Playground: a comprehensive self-service platform for visualization, analytics and exploration of Big Omics Data," *NAR Genomics and Bioinformatics*, vol. 2, no. 1, p. lqz019, 2019.

[14] G. Zhou, J. D. Ewald, and J. Xia, "Omicsanalyst: a comprehensive web-based platform for visual analytics of multi-omics data," *Nucleic Acids Research*, vol. 49, no. W1, pp. W476–W482, 2021.

[15] D. Netanely, N. Stern, I. Laufer, and R. Shamir, "PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets," *BMC Bioinformatics*, vol. 20, no. 1, p. 732, 2019.

[16] M. Sud, E. Fahy, D. Cotter, K. Azam, I. Vadivelu, C. Burant, A. Edison, O. Fiehn, R. Higashi, K. S. Nair, V. Warty, M. Washabaugh, D. Wolan, and I. Zaslavsky, "Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools," *Nucleic Acids Research*, vol. 44, no. D1, pp. D463–D470, 2016.

[17] D. Zhou, W. Zhu, T. Sun, Y. Wang, Y. Chi, T. Chen, and J. Lin, "iMAP: A web server for metabolomics data integrative analysis," *Frontiers in Chemistry*, vol. 9, p. 659656, 2021.

[18] L. Xie, Q. Wang, F. Nan, L. Ge, Y. Dang, X. Sun, N. Li, H. Dong, Y. Han, G. Zhang, W. Zhu, and X. Guo, "OSacc: Gene expression-based survival analysis web tool for adrenocortical carcinoma," *Cancer Management and Research*, vol. 11, pp. 9145–9152, 2019.

[19] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Research*, vol. 45, no. D1, pp. D353–D361, 2017.

[20] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio, "The Reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 48, no. D1, pp. D498–D503, 2020.

[21] F. Wang, Q. Wang, N. Li, L. Ge, M. Yang, Y. An, G. Zhang, H. Dong, S. Ji, W. Zhu, and X. Guo, "OSuvm: An interactive online consensus survival tool for uveal melanoma prognosis analysis," *Molecular Carcinogenesis*, vol. 59, no. 1, pp. 56–61, 2020.

[22] G. Zhang, Q. Wang, M. Yang, Q. Yuan, Y. Dang, X. Sun, Y. An, H. Dong, L. Xie, W. Zhu, Y. Wang, and X. Guo, "OSblca: A web server for investigating prognostic biomarkers of bladder cancer patients," *Frontiers in Oncology*, vol. 9, p. 466, 2019.

[23] G. Zhang, Q. Wang, M. Yang, X. Yao, X. Qi, Y. An, H. Dong, L. Zhang, W. Zhu, Y. Li, and X. Guo, "OSpaad: An online tool to perform survival analysis by integrating gene expression profiling and long-term follow-up data of 1319 pancreatic carcinoma patients," *Molecular Carcinogenesis*, vol. 59, no. 3, pp. 304–310, 2020.

[24] Y. An, Q. Wang, L. Zhang, F. Sun, G. Zhang, H. Dong, Y. Li, Y. Peng, H. Li, W. Zhu, S. Ji, Y. Wang, and X. Guo, "OSlgg: An online prognostic biomarker analysis tool for low-grade glioma," *Frontiers in Oncology*, vol. 10, p. 1097, 2020.

[25] Y. An, Q. Wang, G. Zhang, F. Sun, L. Zhang, H. Li, Y. Li, Y. Peng, W. Zhu, S. Ji, and X. Guo, "OSlihc: An online prognostic biomarker analysis tool for hepatocellular carcinoma," *Frontiers in Pharmacology*, vol. 11, p. 875, 2020.

[26] J. Li, Y. Sang, S. Zeng, S. Mo, Z. Zhang, S. He, X. Li, G. Su, J. Liao, and C. Jiang, "MicrobioSee: A web-based visualization toolkit for multi-omics of microbiology," *Frontiers in Genetics*, vol. 13, p. 853612, 2022.

[27] R. L. Camp, M. Dolled-Filhart, and D. L. Rimm, "X-Tile: A new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization," *Clinical Cancer Research*, vol. 10, no. 21, pp. 7252–7259, 2004.

[28] P. D. Karp, S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I. M. Keseler, and R. Caspi, "Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology," *Briefings in Bioinformatics*, vol. 11, no. 1, pp. 40–79, 2009.

[29] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[30] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[31] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

[32] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2009, pp. 758–759. [Online]. Available: https://doi.org/10.1145/1571941.1572114

[33] E. Benedetti, M. d. C. P. de Lima, A. C. de A. Garcia, A. Fidelis, T. F. de Almeida, B. P. F. de Souza, C. A. O. de Melo, D. C. de O. e Silva, F. R. S. de Souza, I. T. dos Santos, L. L. da C. e Silva, L. R. R. de Oliveira, M. C. T. de Oliveira, T. G. de Araújo, V. E. de C. da Silva, and A. G. de S. e Silva, "A multimodal atlas of tumor metabolism reveals the architecture of gene-metabolite covariation," Oct. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7874228

[34] National Center for Biotechnology Information, "Serum metabolomics of TRAMP mice reveals novel biomarkers of prostate cancer - GSE120584," 2019, accessed: August 9, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120584

[35] J. Jia, F. Li, X.-S. Tang, S. Xu, Y. Gao, Q. Shi, W. Guo, X. Wang, D. He, and P. Guo, "Long noncoding RNA DANCR promotes invasion of prostate cancer through epigenetically silencing expression of TIMP2/3," *Oncotarget*, vol. 7, no. 25, pp. 17 449–17 461, 2016.

[36] C. Lang, Y. Dai, Z. Wu, Q. Yang, S. He, X. Zhang, W. Guo, Y. Lai, H. Du, H. Wang, D. Ren, and X. Peng, "SMAD3/SP1 complex-mediated constitutive active loop between lncRNA PCAT7 and TGF-β signaling promotes prostate cancer bone metastasis," *Molecular Oncology*, vol. 14, no. 4, pp. 808–828, 2020.

[37] Y. Huang, H. Zhu, Z. Liang, W. Wei, H. Yang, Q. Wang, H. Huang, H. He, R. Mo, J. Ye, Q. Dai, W. Zhong, and Y. Liang, "Development and validation of a kinase-related gene signature as a novel diagnostic and prognostic model for prostate cancer," *Biochimica et Biophysica Acta - Molecular Basis of Disease*, vol. 1871, no. 4, p. 167722, 2025.

[38] H. jiang He, X. feng Gu, W. hai Xu, D. jun Yang, X. min Wang, and Y. Su, "Krüppel-like factor 8 is a novel androgen receptor co-activator in human prostate cancer," *Acta Pharmacologica Sinol*, vol. 34, no. 2, pp. 282–288, 2013.

[39] S. A. Bashir, M. Alshalalfa, S. A. Hegazy, M. Dolph, B. Donnelly, and T. A. Bismar, "Cysteine-rich secretory protein 3 (CRISP3), ERG and PTEN define a molecular subtype of prostate cancer with implication to patients' prognosis," *Journal of Hematology & Oncology*, vol. 7, no. 21, p. 21, 2014.

[40] S. S. Y. Teoh, J. C. Whisstock, and P. I. Bird, "Maspin (SERPINB5) is an obligate intracellular serpin," *Journal of Biological Chemistry*, vol. 285, no. 14, pp. 10 862–10 869, 2010.

[41] D. Ye, Y. Wang, X. Deng, X. Zhou, D. Liu, B. Zhou, W. Zheng, X. Wang, and L. Fang, "DNMT3a-dermatopontin axis suppresses breast cancer malignancy via inactivating YAP," *Cell Death & Disease*, vol. 14, no. 2, p. 106, 2023.

[42] W. Xie, Q. Jiang, X. Wu, and L. Wang, "IKBKE phosphorylates and stabilizes Snail to promote breast cancer invasion and metastasis," *Cell Death and Differentiation*, vol. 29, no. 8, pp. 1528–1540, 2022.

[43] Y. Sun, Y. Ding, C. Guo, C. Liu, P. Ma, S. Ma, Z. Wang, J. Liu, T. Qian, L. Ma, Y. Deng, and C. Wu, "α-parvin promotes breast cancer progression and metastasis through interaction with G3BP2 and regulation of TWIST1 signaling," *Oncogene*, vol. 38, no. 24, pp. 4856–4874, 2019.

[44] H.-J. Kim, B.-G. Seo, E.-C. Seo, K.-M. Lee, and C. Hwangbo, "Checkpoint Kinase 1 (CHK1) functions as both a diagnostic marker and a regulator of Epithelial-to-Mesenchymal Transition (EMT) in Triple-Negative Breast Cancer," *Current Issues in Molecular Biology*, vol. 44, no. 12, pp. 5848–5865, 2022.

[45] S. Pellecchia, R. Sepe, A. Federico, M. Cuomo, S. C. Credendino, P. Pisapia, C. Bellevicine, P. Nicolau-Neto, M. S. Ramundo, E. Crescenzi, G. D. Vita, L. M. Terracciano, L. Chiariotti, A. Fusco, and P. Pallante, "The Metallophosphoesterase-Domain-Containing Protein 2 (MPPED2) gene acts as tumor suppressor in breast cancer," *Cancers*, vol. 11, no. 6, p. 797, 2019.

[46] Y. Mao, A. Fu, A. E. Hoffman, D. I. Jacobs, M. Jin, K. Chen, and Y. Zhu, "The circadian gene CRY2 is associated with breast cancer aggressiveness possibly via epigenomic modifications," *Tumor Biology*, vol. 36, no. 5, pp. 3533–3539, 2015.