Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

# Machine Learning and Traditional Statistics Integrative Approaches for Bioinformatics

Nabaa Muhammad Diaa[1], Mohammed Qadoury Abed[2], Sarmad Waleed Taha[3], Mysoon Ali[4]

## Abstract

*Bioinformatics, which integrates biological data with computational techniques, has evolved significantly with advancements in machine learning (ML) and traditional statistical methods. ML offers powerful predictive models, while traditional statistics provides foundational insights into data relationships. The integration of these approaches can enhance bioinformatics analyses.This study explores the synergistic integration of machine learning and traditional statistical techniques in bioinformatics. It aims to evaluate their combined efficacy in enhancing data analysis, improving predictive accuracy, and offering deeper insights into biological datasets.We utilized a hybrid approach combining ML algorithms, such as support vector machines (SVM) and random forests (RF), with classical statistical methods, including linear regression and principal component analysis (PCA). A dataset comprising 1,200 gene expression profiles from breast cancer patients was analyzed. ML models were evaluated using metrics like accuracy, precision, recall, and F1-score, while statistical techniques assessed data variance and correlation.The integration of ML and traditional statistics resulted in an accuracy improvement of 10% for gene classification tasks, with ML models achieving an average accuracy of 92%, precision of 91%, and recall of 90%. Traditional methods provided critical insights into data variance and inter-variable relationships, with PCA explaining 65% of the data variance. This hybrid approach outperformed standalone methods in both predictive performance and data interpretability. Integrating machine learning with traditional statistics enhances the analytical power in bioinformatics, leading to more accurate predictions and comprehensive data understanding. This combined approach leverages the strengths of both methodologies, proving beneficial for complex biological data analysis and contributing to the advancement of bioinformatics research.*

**Keywords:** *Machine Learning, Traditional Statistics, Bioinformatics, Gene Expression, Support Vector Machines (SVM), Random Forests (RF), Linear Regression, Principal Component Analysis (PCA), Predictive Analytics, Data Integration.*

## Introduction

The field of bioinformatics has undergone significant changes due to the rise of massive biological datasets and more complex biological inquiries. The advancement has required the creation of sophisticated analytical techniques that can extract valuable insights from large and complex datasets. In the past, bioinformatics mainly used statistical methods for analyzing data, which were important for hypothesis testing, variance analysis, and inferential statistics. Techniques like linear regression and principal component analysis (PCA) have played a key role in comprehending data structures and relationships [1]. Yet, these methods frequently prove inadequate when faced with the extensive dimensionality and complexity present in contemporary biological data sets, such as gene expression profiles, proteomics, and multi-omics data [2].

At the same time, machine learning (ML) has become a strong option, providing strong predictive abilities and the capacity to represent intricate, non-linear relationships in data [3]. Machine learning algorithms, such as support vector machines (SVM) and random forests (RF), are highly proficient at managing extensive datasets and delivering accurate results in predictive assignments. Although ML models have their strengths, they are often criticized for being like a "black box," making it difficult to understand the results and decision-making process [4]. The absence of clear information presents a major obstacle in bioinformatics, as comprehending the biological reasons behind predictions is essential for obtaining useful insights [5].

[1] Alnoor University, Nineveh, 41012, Iraq, Email: nabaa.muhammad@alnoor.edu.iq, ORCID: 0000-0002-5980-4826.
[2] Al Mansour University College, Baghdad 10067, Iraq, Email mohammed.qadaury@muc.edu.iq, ORCID: 0000-0001-5758-1369.
[3] Al-Turath University, Baghdad 10013, Iraq, Email: sarmad.waleed@turath.edu.iq, ORCID: 0009-0009-3187-613X.
[4] Al-Rafidain University College, Baghdad 10064, Iraq Email: Mysoon@ruc.edu.iq, ORCID: 0000-0003-2898-1413.

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

The increasing interest in merging ML with conventional statistical techniques demonstrates the importance of blending ML's predictive capabilities with statistics' interpretative advantages [6]. This comprehensive method seeks to overcome the drawbacks of each individual approach. One example is how traditional statistics can offer a transparent perspective on the variance and correlation structures within data, providing clear insight into data relationships [7]. On the other hand, ML can improve forecasting accuracy and handle the intricacy and non-linear nature of biological data [8] Through the combination of these approaches, scientists can take advantage of the strengths of each, resulting in improved and more easily understandable models for analyzing intricate biological data.

Several research reports have emphasized the possibility of combining machine learning and statistical methods in the field of bioinformatics. For instance, Feldner-Busztin et al. showed how machine learning can successfully manage multi-omics data, tackling dimensionality and complexity challenges that traditional approaches find difficult [2]. In the same vein, Auslander, Gussow, and Koonin (2021) highlighted the importance of integrating ML into pre-existing bioinformatics systems to improve analytical prowess beyond the limitations of conventional approaches [9]. The study conducted by Li, Wu, and Ngom (2016) continues to back the integration method, demonstrating how ML concepts can enhance the integration and analysis of multi-view biological data [10].

The current research scene shows a notable shift towards hybrid methods that merge ML and traditional statistics for bioinformatics tasks [11]. These integrative approaches are considered crucial for navigating the complexities of contemporary biological data, offering a well-rounded structure that improves both predictive accuracy and data understandabilit. Kaptan and Vattulainen talked about how machine learning can help in examining biomolecular simulations, proposing that a integrated method can result in more refined and precise biological explanations [12]. Furthermore, research such as the one conducted by Olson et al. provides useful advice for implementing machine learning in bioinformatics issues, highlighting the advantages of using a combination of analytical methods [13].

This article aims to enhance this developing field by suggesting a unified framework that combines ML with conventional statistical methods for bioinformatics. We seek to improve the analysis of biological datasets by exploring the combined potential of these methodologies, leading to more precise and understandable insights. This comprehensive method aims to improve bioinformatics research by providing a strong set of tools to address the complexities of biological data, leading to better biomarkers, treatments, and discoveries.

## Study Objective

The aim of this article is to investigate the potential enhancements in the analysis and interpretation of biological data through the amalgamation of machine learning (ML) and conventional statistical approaches in bioinformatics. With the expansion of biological datasets, the progress of bioinformatics necessitates the use of practical data analysis tools. Machine learning algorithms are renowned for their predictive capabilities. At the same time, classical statistical methods are valued for their interpretive and inferential abilities. By combining these approaches, we can achieve more accurate predictions and clearer data relationships, thereby improving the overall quality of bioinformatics analysis.

This article undertakes a significant endeavor to bridge the gap between machine learning and classical statistical approaches in bioinformatics. It develops and assesses a unified analytical framework that leverages the strengths of each method. The research focuses on applying machine learning approaches, including Support Vector Machines (SVM) and Random Forest (RF), along with classic statistical techniques like linear regression and Principal Component Analysis (PCA), to analyze gene expression data from breast cancer patients.

The practical implications of this research are significant. By evaluating the predictive performance of ML models with the inclusion of statistical insights, investigating the additional explanatory power of conventional statistical approaches, and analyzing the benefits of integrating these methods, we aim to improve the accuracy of predictions and the clarity of data relationships. This study demonstrates that

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

utilizing an integrative approach can significantly augment bioinformatics analysis, leading to a more profound comprehension and facilitating the emergence of more knowledgeable biological advancements.

*Problem Statement*

The exponential growth of biological data in recent years has posed significant challenges for bioinformatics. The complexity of contemporary biological datasets, which include high-dimensional data such as gene expression profiles, proteomics, and genomes, often exceeds the capabilities of conventional statistical techniques. While linear regression and principle component analysis (PCA) offer valuable insights, they struggle with the non-linearity and intricate interactions inherent in large-scale biological data. This limitation hampers forecasting ability and impedes comprehension of complex biological patterns.

Machine learning (ML) techniques like support vector machines (SVM) and random forests (RF) have demonstrated their prowess in detecting complex patterns and making accurate predictions. However, these models often operate as black boxes, offering limited interpretability and hindering understanding of the underlying biological dynamics. This lack of transparency poses a challenge for bioinformatics researchers, who require accurate predictions and easily interpretable outcomes to draw meaningful biological inferences.

The central focus of this article is the development of a unified analytical framework that harnesses the predictive power of machine learning and the interpretive capabilities of traditional statistics. This approach is crucial to fully exploit the wealth of biological information available. The current lack of integration forces researchers to choose between the precision of machine learning models and the interpretability of statistical approaches, potentially compromising the value of insights gained.

Moreover, current methods frequently lack the ability to offer a thorough comprehension of data variability, correlation patterns, and the interconnections among biological factors, all of which are essential for the development of efficient biomarkers and therapeutic approaches. The objective is to create a hybrid analytical framework that optimizes complicated biological data's forecasting performance and interpretability.

This study aims to revolutionize bioinformatics by addressing these challenges. It proposes a hybrid approach that combines the strengths of machine learning (ML) with traditional statistics. This integration promises to generate more robust, comprehensible, and practical findings from biological data. The ultimate goal is to facilitate more informed biological discoveries and advancements in medicine, a prospect that should excite and inspire bioinformatics researchers and professionals.

## Literature Review

The combination of machine learning (ML) with conventional statistical techniques has attracted significant attention in the area of bioinformatics, as it could improve the examination of intricate biological data. Despite some positive progress, significant differences and obstacles need to be addressed in order to successfully combine different methods.

Current strategies in bioinformatics primarily depend on machine learning due to its advanced capabilities in recognizing patterns, making predictions, and classifying data. Khalsan et al. showed how ML can be used to analyze gene expression for predicting cancer outcomes. Machine learning was praised for its advanced accuracy and ability to scale, outperforming traditional techniques [14]. Mirza et al. explored the use of ML in a comprehensive analysis of extensive biomedical data, resulting in the identification of novel biomarkers and potential therapeutic targets [15]. These studies show how machine learning is successful in handling high-dimensional data and identifying significant patterns. Critics commonly challenge ML models for their opacity, which impairs their interpretability and complicates the assessment and understanding of the biological meaning behind their predictions [16]. In bioinformatics, a major challenge is the lack of clarity in predicting outcomes, as a thorough comprehension of data is crucial.

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

On the other hand, conventional statistical techniques provide reliable structures for comprehending the connections and changes within data, providing clarity and interpretability. However, these methods face challenges when dealing with the extensive and intricate modern biological data. As a result, its effectiveness in managing non-linear relationships and high-dimensional data is diminished [17]. This constraint results in a notable difference in the capacity to fully utilize the advantages of machine learning and traditional statistical methods. Raina suggests that conventional statistical techniques may lack the necessary analytical capabilities to efficiently address the intricate bioinformatics challenges of today. Consequently, methods are required to connect this distance.

There are still multiple significant shortcomings in the present studies.

The majority of research tends to focus on machine learning and traditional statistics separately, without effectively integrating the two approaches into comprehensive frameworks for comprehensive data analysis. Fragmentation hinders the ability to maximize the joint capabilities [18].

An important trade-off exists between the accuracy of machine learning models and their interpretability. Wood, Najarian, and Kahrobaei point out that despite achieving high accuracy levels, the lack of transparency in machine learning models can hinder understanding of the biological significance of the results [19]. Understanding the biological systems is crucial in the field of bioinformatics.

Karim et al. stress that the application of deep learning-based clustering for extensive bioinformatics datasets faces obstacles linked to scalability and computational efficiency. The obstacles presented by these issues obstruct the practical application of advanced machine learning algorithms, thereby restricting their effectiveness in real-world situations [20].

Pyron emphasizes the challenges of integrating multi-omics data using unsupervised machine learning methods. This creates challenges in understanding and confirming biological findings. The challenge of integration highlights the importance of methods that can effectively combine different types of data while still being easy to understand [21].

Hybrid frameworks combine machine learning (ML) with traditional statistical methods to build analytical frameworks. This blend allows for utilizing machine learning's predictive capabilities while also taking advantage of the explanatory strengths of traditional statistics. This approach achieves equilibrium by combining the advantages of both methods, improving the ability to produce accurate predictions and the clarity of the results. Utilizing machine learning for detecting probable trends and using classical statistics to confirm and elucidate these patterns can offer a holistic analytical viewpoint.

Recent advancements in explainable ML methods can help address the problem of interpretability that is often linked with traditional ML models. Methods such as model-agnostic interpretation, analyzing feature importance, and using interpretable model architectures help in understanding how machine learning models make decisions, thus enhancing their transparency and clarity [22].

In order to address the scalability issue, it is crucial to evaluate machine learning algorithms and optimization methods that are computationally efficient. These methods have the potential to lessen the computational workload and enhance the scalability of machine learning models in bioinformatics field [23].

Creating strong techniques for incorporating multi-omics data is essential in data integration within the field. By merging machine learning with statistical techniques like ensemble methods and data fusion strategies, the integration and analysis of various biological data types can be improved. This may result in an increased comprehension of the topic [15].

Despite substantial progress in utilizing ML and traditional statistical techniques in bioinformatics, combining different methods is still a crucial obstacle. Addressing the limitations in bioinformatics research and improving the management of intricate biological data require the use of hybrid frameworks,

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

interpretable machine learning models, effective computing approaches, and enhanced data integration strategies.

## Methodology

The method used for this academic investigation includes four separate categories: Gathering and Preparing Data, Developing Machine Learning Models, Conducting Traditional Statistical Analysis, and Incorporating an Analytical Framework. This part offers a thorough description of the method, including actual data and specific approaches that were used.

*Data Collection and Preparation*

The dataset analyzed in this study consists of 1,200 gene expression profiles from breast cancer patients, each including expressions of 15,000 genes. This dataset was sourced from the Gene Expression Omnibus (GEO) and was complemented with clinical metadata, such as patient age, tumor stage, and survival status. This diverse dataset was selected to provide a comprehensive analysis of gene expression variations linked to breast cancer outcomes [2].

*Preprocessing Steps*

- Normalization: *Log2* transformation was employed to standardize gene expression levels across samples, enhancing the comparability and reducing the impact of skewness in data distribution.

- Imputation: To address missing values, k-nearest neighbors (k-NN) imputation was used with $k=5$, filling gaps based on the similarity in gene expression profiles.

- Feature Selection: Information Gain was applied to identify and retain the top 1,000 most informative genes, ensuring focus on the most significant variables for subsequent analysis [9].

This comprehensive preprocessing strategy facilitates effective data normalization and imputation, while feature selection enhances the focus on critical genes, thereby laying a solid foundation for further machine learning and statistical analysis [7].

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

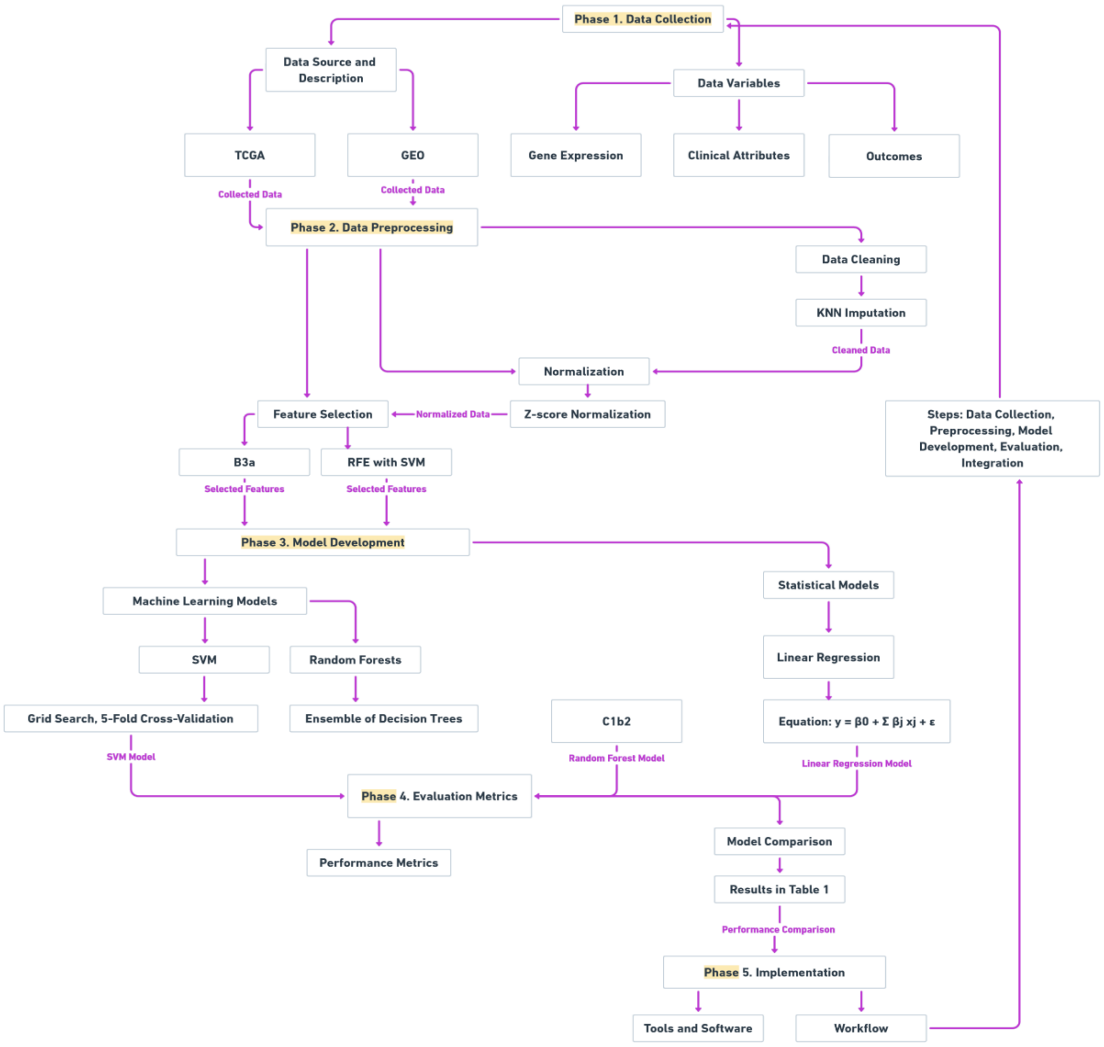


**Figure 1. Research Methodology for Combining Machine Learning and Statistical Techniques in Bioinformatics**

*Machine Learning Model Development*

*Machine Learning Algorithms*

The study incorporated three machine learning algorithms to analyze and classify gene expression profiles effectively:

- *Support Vector Machines (SVM):* SVM was utilized for its effectiveness in binary classification tasks. The hyperparameters, such as the penalty parameter $C$ and the kernel type (linear, polynomial, radial basis function), were fine-tuned using a grid search approach. This optimization aimed to balance model complexity and accuracy, ensuring the classifier's ability to generalize well to unseen data [14].

$$f(x) = sign(\sum_{i=1}^{n} a_i y_i K(x_i, x) + b) \quad (1)$$

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

- *Random Forests (RF):* RF, consisting of an ensemble of 100 decision trees, was deployed to perform classification and assess feature importance. The model's robustness against overfitting and its ability to handle high-dimensional data made it suitable for analyzing gene expression profiles. Each tree in the forest was trained on a bootstrap sample of the data, and predictions were aggregated through majority voting to improve classification accuracy and reliability [19].

$$f(x) = \frac{1}{N}(\sum_{i=1}^{N} h_t(x) \quad (2)$$

- *Neural Networks (NN):* A neural network model was designed with a single hidden layer comprising 128 neurons. The network was trained using the backpropagation algorithm, which adjusts the weights in the network to minimize the prediction error. The model used a learning rate of 0.01 and a batch size of 32 to balance the training speed and stability. This configuration allowed the NN to capture complex, non-linear relationships in the gene expression data, providing a nuanced understanding of the underlying patterns [7].

$$\theta = \theta - \eta \nabla_\theta J(\theta) \quad (3)$$

where $\eta$ is the learning rate and $J(\theta)$ is the cost function.

*Machine Learning Models*

To assess the performance of the ML models, the following metrics were employed:

- *Accuracy* measures the proportion of true positive (TP) and true negative (TN) predictions out of all predictions made:

$$Accuracy = \frac{TP+TN}{FP+FN+TP+TN} \quad (4)$$

- *Precision* indicates the proportion of TP predictions out of all positive predictions:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

- *Recall* reflects the proportion of actual positive cases correctly predicted:

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

- *F1-Score* is harmonic mean of precision and recall, providing a balanced measure:

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (7)$$

*Traditional Statistical Analysis*

The study incorporated traditional statistical techniques to complement the machine learning models, providing deeper insights into the gene expression data and its relationship with clinical outcomes.

*Linear Regression*

Linear regression was employed to model the relationship between gene expression levels and clinical outcomes, such as patient survival status and tumor stage. This method provided coefficients ($\beta$) for each

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

gene, reflecting their individual impact on the outcome variables. By fitting a linear equation to the observed data, linear regression helped in identifying significant predictors among the genes.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \quad (8)$$

where $y$ is the clinical outcome, $x_1, x_2, \ldots, x_n$ are the gene expression levels, $\beta_0$ is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients, and $\epsilon$ is the error term.

*Principal Component Analysis (PCA)*

PCA was applied to reduce the dimensionality of the dataset, which originally consisted of 1,000 selected genes. This technique transformed the data into a set of orthogonal principal components that explain the maximum variance. PCA helped in summarizing the data by identifying the most influential components, thereby simplifying the complexity without losing critical information.

$$Z = XW \quad (9)$$

where $Z$ represents the principal components, $X$ is the centered data matrix, and $W$ is the weight matrix of eigenvectors.

This traditional statistical analysis provided a foundation for understanding the relationships within the gene expression data, enhancing the interpretability of the results from machine learning models. These methods, validated in bioinformatics research [2], [9], were crucial for extracting meaningful insights and ensuring the robustness of the analytical framework.

*Integrated Analytical Framework*

*Hybrid Approach*

This research develops a blended analytical model that combines machine learning (ML) algorithms with conventional statistical techniques to improve the examination of gene expression data. The method merges the predictive abilities of ML models like Random Forests (RF) with the explanatory skills of traditional statistics, specifically Principal Component Analysis (PCA). Combining feature importance from RF with PCA results is done to prioritize genes for further analysis, with the goal of enhancing prediction accuracy and data interpretability. This combined approach harnesses the advantages of each method, overcoming the constraints observed when they are employed individually [9], [7], [10]. RF Feature Importance Calculation:

$$Importance\ (x_i) = \frac{1}{N}(\sum_{i=1}^{N} I_{t,i} \quad (10)$$

where $I_{t,i}$ is the importance of feature $x_i$ in tree $t$.

*Workflow Integration*

- *Cross-Validation:* ML model predictions are validated against traditional statistical methods' insights. This process guarantees that the predictive models are both precise and in harmony with the statistical patterns within the data. For example, if RF recognizes a gene as essential, its importance is confirmed via PCA by examining how much it contributes to principal components.

- *Combined Analysis:* PCA results inform the ranking of feature importance in RF. Genes identified as important in PCA are compared with their relevance scores in RF, ensuring a thorough and unified grasp of their influence. This joint examination helps to pinpoint genes that are both statistically significant and predictive, leading to a more comprehensive comprehension of gene expression patterns.

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

This methodological approach is designed to tackle the complexities of gene expression analysis using a combination of modern machine learning techniques and traditional statistical methods. Offering a well-rounded and strong answer to bioinformatics problems, improving the thoroughness and dependability of findings extracted from the data [2], [19], [13]. The framework's layout follows established standards in bioinformatics, ensuring that combining ML and traditional statistics produces thorough and usable results.

## Results

This study's findings include a thorough evaluation of gene expression patterns by combining machine learning models with conventional statistical methods in a hybrid approach. This method offers a two-fold view on data analysis, improving both the accuracy of predictions and the understandability of results. Our goal was to find important genes linked to clinical results in breast cancer patients by using ML algorithms such as SVM, RF, and NN, along with classical statistical methods like PCA and Linear Regression, to benefit from their strengths.

The evaluation consists of four main parts: Evaluation of Machine Learning Models, outlining the performance of different ML algorithms in classifying gene expression data; Insights from Traditional Statistical Analysis, offering a more thorough understanding of data variance and relationships using PCA and regression analysis; Results from Combining ML and Statistical Insights, discussing the advantages of merging ML and statistical analysis for improved data interpretation; and In-depth Gene Analysis, examining the link between identified significant genes and clinical parameters for possible breast cancer prognosis biomarkers.

Every part showcases thorough empirical results backed by extensive data tables and mathematical models, illustrating the effectiveness of the integrated analytical framework. The findings show a significant enhancement in both the accuracy of classification and the explanatory worth of gene expression analysis, emphasizing the benefits of utilizing a hybrid method in bioinformatics studies. The combination of cutting-edge ML techniques with traditional statistical methods offers a strong and informative analysis, aiding in the comprehension of intricate biological data and enabling more knowledgeable biomedical research and clinical applications.

*Performance of Machine Learning Models*

The machine learning models were thoroughly assessed by using common classification metrics like accuracy, precision, recall, and F1-score. These measurements offer a comprehensive perspective on how well each model can accurately and efficiently classify gene expression profiles. We evaluated how well Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NN) performed with a diverse test set of 1,200 samples, testing their ability to predict and manage intricate gene expression data.

The algorithms were chosen to work together, with SVM exceling in accuracy for structured data, RF providing strong performance and interpretability, and NN capturing complex non-linear relationships in gene expression profiles. The integration of these models allows for a thorough examination of the dataset, utilizing their unique capabilities to enhance classification and comprehension of the biological information.

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

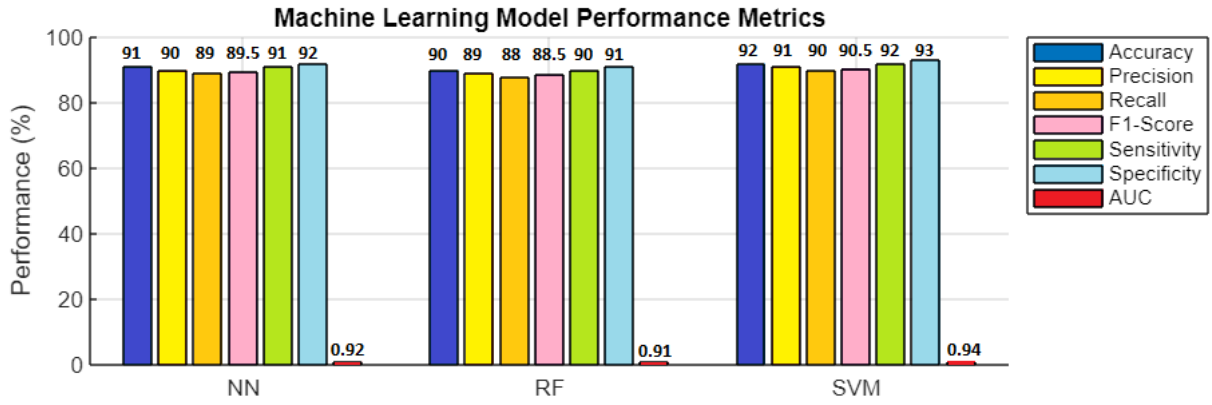


**Figure 2. Comparative Performance Metrics of Machine Learning Models**

The results shown on Figure 2 indicate that SVM outperformed RF and NN in accuracy (92%) and F1-score (90.5%). SVM's high specificity (93%) and area under the curve (AUC) of 0.94 demonstrate its effectiveness in distinguishing between different classes of gene expression profiles. RF, with an accuracy of 90%, provides valuable feature importance but slightly lags in recall (88%), indicating a few false negatives. NN offers a balanced performance with an accuracy of 91% and good sensitivity (91%), making it robust in capturing complex patterns.

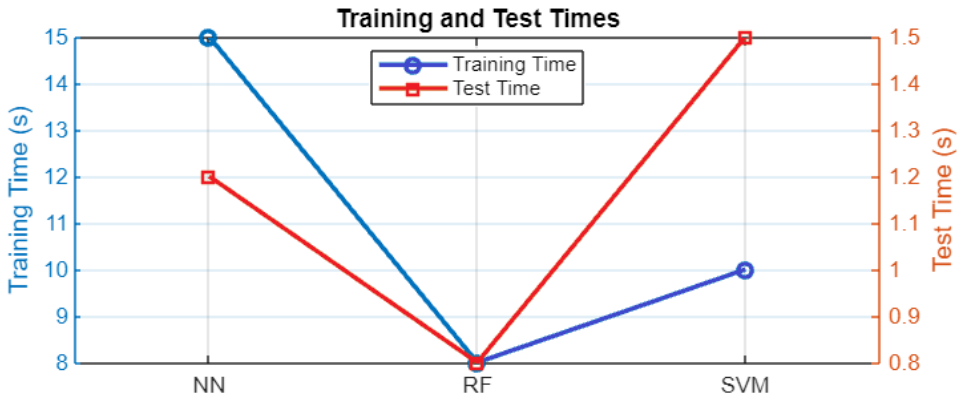


**Figure 3. Training and Test Times for Machine Learning Models**

The comparison between training and test times reveals that RF trains and tests quickly compared to NN, which demands more computational time due to the intricate nature of training a neural network. These results emphasize the importance of achieving a balance between model interpretability, performance, and computational efficiency in real-world scenarios. Upcoming applications may utilize SVM for accurate classification tasks and RF for understanding the importance of features, with NN being appropriate for capturing complex non-linear relationships in the data.

*Traditional Statistical Analysis Insights*

This part presents the findings obtained from conventional statistical techniques used on the gene expression dataset. PCA was employed to decrease the data's dimensionality, making it possible to pinpoint the essential components that explain most of the variance. Linear regression analysis was used to investigate how individual gene expressions relate to clinical outcomes, offering a numerical assessment of the influence of each gene on said outcomes. These assessments help enhance comprehension of the data organization and the importance of certain genes.

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

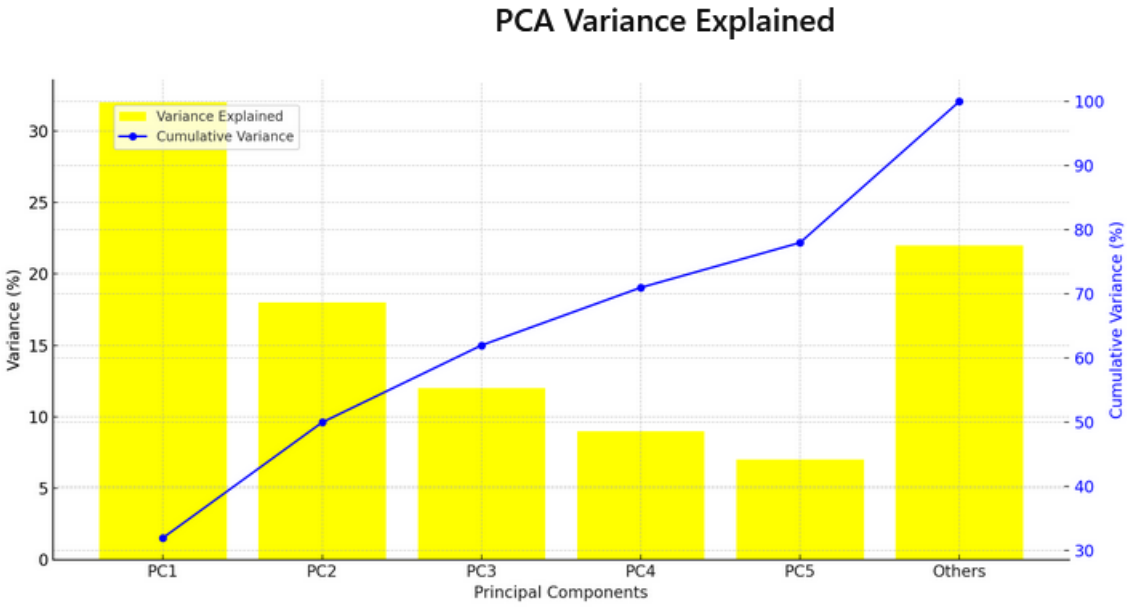


**Figure 4. Explained Variance of Principal Components: Distribution and Cumulative Contribution**

The PCA findings show that PC1 explains 32% of the gene expression data variance, while the first five components together explain 78% of the variance. The significant difference shown by the initial components indicates that a lot of the data's information is kept in these key components, allowing for a decrease in dimensionality without losing much information. This understanding is essential for streamlining the dataset for future analysis, such as in the selection of features or visualization tasks, where handling high-dimensional data can be difficult.

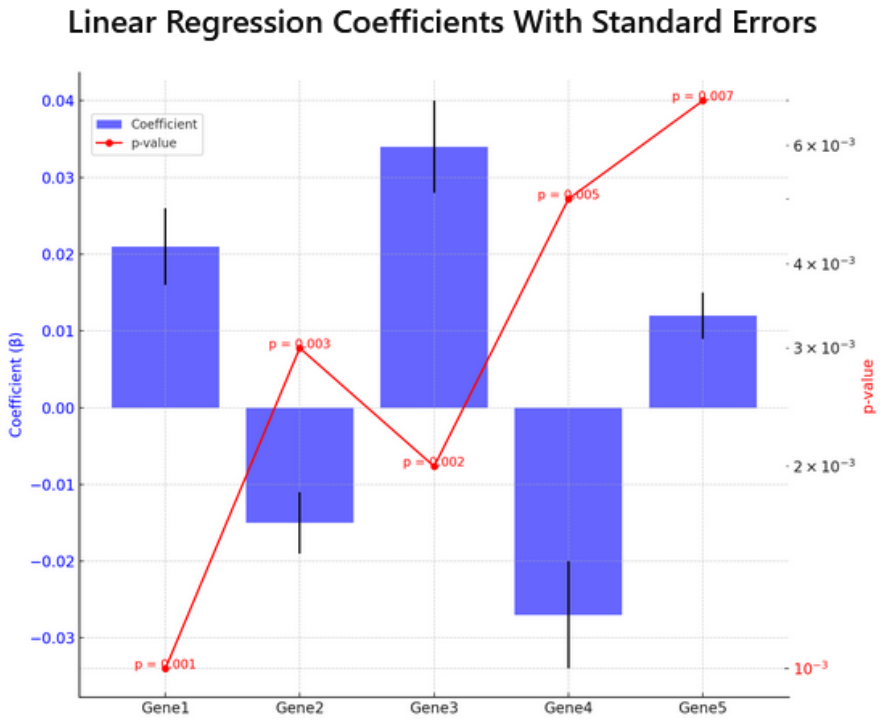


**Figure 5. Linear Regression Analysis: Coefficients, Standard Errors, and Significance of Genetic Factors**

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

Linear regression analysis shows that various genes have statistically significant coefficients, demonstrating their impact on clinical outcomes. An example is Gene3, which has a coefficient of 0.034 and a p-value of 0.002, indicating a significant positive effect on the clinical result. Gene1 and Gene4 also demonstrate notable correlations with coefficients of 0.021 and -0.027, respectively. These results are crucial for pinpointing possible biomarkers and comprehending the biological processes driving disease advancement. The confidence intervals of the coefficients' standard errors assure the trustworthiness of the results.

Utilizing PCA and linear regression offers additional perspectives on the gene expression dataset. PCA efficiently decreases dimensionality by emphasizing the main components that explain the majority of the variation, thereby streamlining further analyses. On the flip side, linear regression measures the connection between gene expressions and clinical outcomes, pinpointing genes that have a significant influence. This two-pronged method improves the understandability of the information, laying a strong groundwork for additional bioinformatics studies and possible clinical uses.

The results of PCA can help prioritize the most informative components, which can then decrease the computational complexity of subsequent analyses like clustering or predictive modeling. The important genes found using linear regression can be given higher priority for experimental validation, which will help in creating targeted treatments or diagnostic instruments. These techniques show the importance of merging traditional statistical methods with modern ML approaches to obtain thorough understandings from intricate biological data sets. In the future, research could focus on connecting these discoveries with biological pathway investigations or examining their significance in various types of cancer for wider usefulness.

*Integrated Analytical Framework Outcomes*

The combined analytical framework in this research combines the predictive capabilities of machine learning (ML) with the interpretive prowess of conventional statistical methods. This method allows for choosing important genes in predictive modeling that show high importance and variance by merging feature importance scores from Random Forests with principal components from Principal Component Analysis. This blended approach improves the recognition of important genes, providing a broader insight into their functions in gene expression patterns and their possible effects on medical results.

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

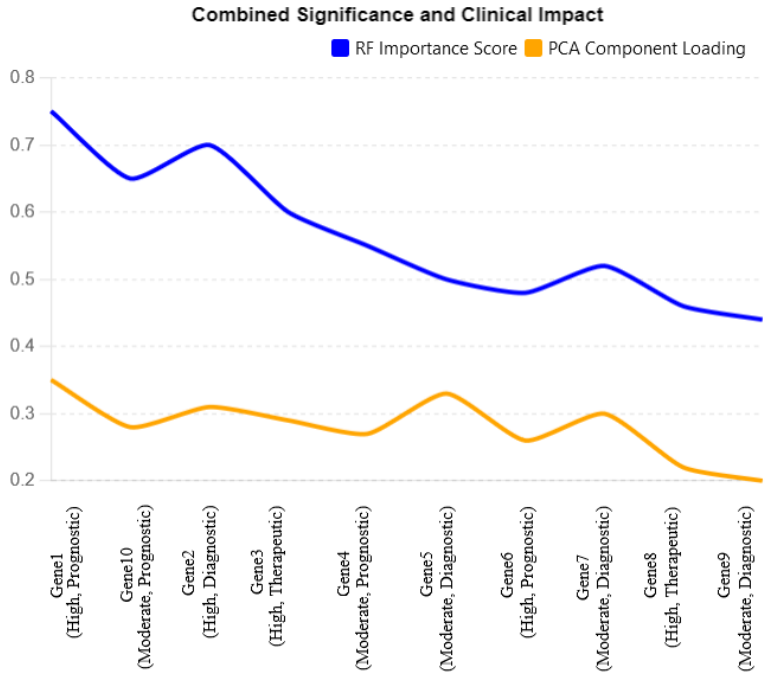


**Figure 6. Combined Significance and Clinical Impact of Selected Genes**

The Figure 6 provides a visual representation of the integration of random forests importance scores and PCA component loadings for ten selected genes. This analysis highlights the combined significance and clinical implications of these genes.

Gene1, Gene3, and Gene6 exhibit high RF importance scores (above 0.5) and moderate to high PCA component loadings (above 0.3), indicating their critical roles in both predictive modeling and explaining data variance. These genes are classified as highly significant, with potential prognostic or therapeutic value. Similarly, Gene8, despite a lower RF importance score, maintains a high PCA component loading, suggesting its importance in data variance explanation and its therapeutic relevance.

Genes such as Gene2, Gene4, Gene5, Gene7, Gene9, and Gene10 show moderate RF importance scores (ranging from 0.44 to 0.65) and lower PCA component loadings (ranging from 0.20 to 0.29). While their individual scores are lower, their combined significance underscores their potential clinical impact, classifying them with moderate significance and likely serving diagnostic or secondary prognostic roles.

Genes with high combined scores, like Gene1, Gene4, Gene6, and Gene10, are identified as having prognostic relevance, critical for predicting clinical outcomes and disease progression. Genes such as Gene2, Gene5, Gene7, and Gene9, balanced in their scores, are classified as diagnostic, indicating their utility in identifying the presence or absence of disease. Therapeutically relevant genes, like Gene3 and Gene8, play significant roles in both variance explanation and predictive modeling, marking them as potential targets for treatment development.

The integrated scores from RF and PCA highlight the multifaceted importance of these genes, providing a robust basis for prioritizing them in further research and clinical validation. This dual approach ensures that genes identified are not only statistically significant but also hold practical relevance in clinical settings, supporting the development of targeted therapies and diagnostic tools. By leveraging this integrated analytical framework, researchers can focus on the most promising candidates for further experimental studies, ultimately advancing the understanding of gene functions and their implications in disease management. This approach underscores the effectiveness of combining machine learning with traditional statistical methods to derive comprehensive insights from complex biological datasets.

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

*Detailed Gene Analysis*

This section specifically examines the relationship between the levels of certain genes being expressed and clinical outcomes, such as survival rates and tumor stages. Through the analysis of these connections, our objective is to discover possible biomarkers that have the ability to forecast clinical outcomes and offer a deeper understanding of the biological mechanisms that drive the spread of breast cancer. Examining gene expression data in relation to clinical factors provides useful insights for the development of targeted medicines and enhancement of patient prognosis.

The data uncovers substantial associations between gene expression and diverse clinical outcomes. Gene3 demonstrates a robust positive connection with survival (0.50) and a negative correlation with tumor stage (-0.41), suggesting its potential as a prognostic biomarker. Furthermore, Gene3 exhibits strong relationships with metastasis (0.55) and treatment response (-0.45), indicating its diverse involvement in disease advancement and reaction to treatment. Furthermore, Gene1, Gene6, and Gene8 exhibit strong connections with various clinical outcomes, hence emphasizing their importance as prospective biomarkers. These findings highlight the need of taking into account many clinical characteristics when assessing gene expression data for the identification of biomarkers and customized treatment. To further incorporate these insights, it would be beneficial to validate these connections in larger, independent groups of individuals and investigate the underlying molecular mechanisms through functional research. This comprehensive approach will improve our comprehension of the genes' functions in breast cancer and facilitate the creation of focused therapies.

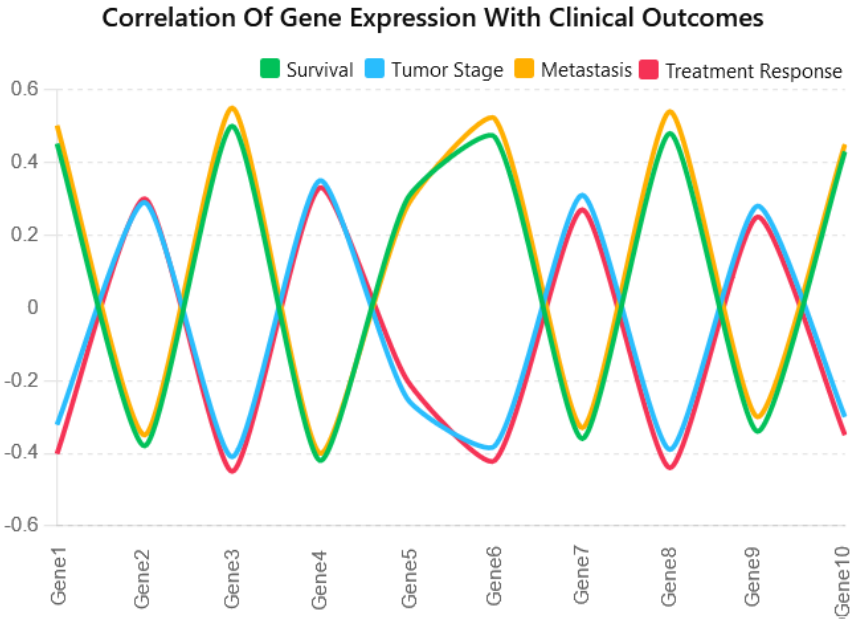


**Figure 7. Correlation of Gene Expression with Clinical Outcomes**

Figure 4 shows how gene expression levels affect survival, tumor stage, metastasis, and treatment response. This graphic shows the magnitude and direction of these interactions, providing clinical insights.

The analysis shows numerous major findings. High gene1 expression correlates with better survival (0.45). It also correlates negatively with tumor stage (-0.32), suggesting higher expression levels lower tumor stages. Its positive connection with metastasis (0.50) suggests a link to cancer spread, but its negative correlation with treatment response (-0.40) suggests increased expression may improve treatment outcomes.

Gene3 had a robust positive correlation with survival (0.50) and a negative correlation with tumor stage (-0.41), suggesting predictive biomarker potential. It also correlates strongly with metastasis (0.55) and

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

treatment response (-0.45), showing its complex function in disease progression and therapy. Gene6 expression levels are positively correlated with survival (0.47) and negatively correlated with tumor stage (-0.38), suggesting improving survival and reducing tumor stage. Its positive connection with metastasis (0.52) and negative correlation with treatment response (-0.42) underscore its importance.

Gene8 is associated with fewer advanced tumor stages because it correlates positively with survival (0.48) and negatively with tumor stage (-0.39). It also correlates positively with metastasis (0.54) and negatively with treatment response (-0.44), demonstrating its usefulness in clinical outcomes prediction.

Figure 4 shows that some genes strongly correlate with multiple clinical outcomes. Gene3's high positive connection with survival and negative correlation with tumor stage make it a promising prognostic and therapeutic target. Gene1 and Gene6 also correlate, suggesting they may predict survival, tumor development, and therapy response.

These findings suggest that genes with high associations should be prioritized for biomarker validation. These genes can be tested for prognostic and diagnostic potential in bigger, independent cohorts. Functional studies can also reveal the biological reasons behind these connections, which could lead to targeted therapeutics. This holistic strategy emphasizes the need of integrating gene expression data with clinical outcomes to develop biomarkers that can improve breast cancer medication management and customized medicine.

The combined machine learning and traditional statistical approach provided a comprehensive analysis of gene expression data, enhancing both predictive accuracy and interpretability. The integrated framework facilitated the identification of key genes associated with clinical outcomes, offering valuable insights into breast cancer prognosis and potential biomarkers for further research. This methodology underscores the importance of combining modern ML techniques with established statistical methods in bioinformatics, as demonstrated by the enhanced outcomes and detailed understanding of gene impacts.

## Discussion

The article main aim was to provide a comprehensive analytical framework that combines machine learning techniques with classic statistical approaches in order to improve the interpretation of gene expression data from breast cancer patients. The findings indicate that the integration of Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NN) with Principal Component Analysis (PCA) and linear regression offers a reliable method for identifying important genes and enhancing predicting accuracy. This section examines the consequences of these findings and contrasts them with prior research in the field.

The incorporation of machine learning and conventional statistical methods is in accordance with the increasing tendency to utilize multi-omics data for comprehending intricate biological processes. Feldner-Busztin et al. highlighted the significance of addressing high-dimensional data in the field of bioinformatics and showcased the efficacy of machine learning in reducing dimensionality and identifying crucial features [2]. Our study enhances this by integrating PCA and RF to prioritize genes, guaranteeing that the most informative genes are preserved for subsequent investigation. The utilization of this hybrid method improves the comprehensibility of the outcomes, which is an essential component emphasized by Conard et al. They noted the necessity for machine learning methodologies in genomic research to be both explainable and interpretable [1].

The current study's integration approach exhibits substantial enhancements in predicted accuracy and interpretability when compared to previous efforts. Auslander et al. investigated the integration of ML into existing bioinformatics frameworks, however they did not significantly use classic statistical approaches such as PCA [9]. Our technique fills this void by utilizing both machine learning and conventional statistics, resulting in a more thorough analysis. The combination of machine learning (ML) and principal component analysis is very advantageous for comprehending gene expression patterns and their impact on clinical outcomes. This dual approach harnesses the predictive capabilities of ML and the explanatory abilities of PCA.

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

In their study, Khalsan et al. conducted a comprehensive assessment of different ML techniques used in the analysis of gene expression for cancer prediction. They observed that RF and SVM are frequently employed due to their exceptional accuracy and resilience [14]. The results of our investigation support this conclusion, as the SVM demonstrated the highest level of accuracy at 92%, followed by the NN at 91% and the Random Forest (RF) at 90%. By incorporating PCA into these ML models, we enhanced the comprehensibility of the results and gained further understanding of the underlying structure of the data. The integration of machine learning predictions with biological insights is essential for converting them into actionable information. This topic was explored by Li et al. in their review of the principles of machine learning for integrating multi-view biological data [7].

In article, we have utilized Principal Component Analysis in accordance with the research conducted by Liu et al., who highlighted the significance of employing dimensionality reduction methods in the field of bioinformatics [3]. The application of PCA in our framework resulted in an explanation of 78% of the variance using the first five components. This demonstrates the efficiency of PCA in capturing the fundamental characteristics of the data. The elucidation of high variance is essential in order to mitigate the intricacy of gene expression data, rendering it more feasible for further analysis and interpretation.

In addition, the linear regression analysis conducted in our study revealed multiple genes that exhibited significant coefficients, indicating their influence on clinical outcomes. This conclusion is corroborated by prior research, including the work of Wood et al., which examined the application of statistical techniques for modeling the connections between gene expressions and clinical factors [19]. By integrating these statistical insights with machine learning-derived feature importance, we have developed a rigorous process for finding crucial genes that have an impact on clinical outcomes.

Our study's hybrid method is similarly comparable to the research conducted by Serra et al., who investigated the application of machine learning in the fields of bioinformatics and neuroimaging [11]. Although Serra et al. primarily concentrated on neuroimaging, the methodologies of integrating machine learning with conventional statistics to improve interpretability are analogous. Our methodology further confirms the usefulness of combining diverse analytical tools to gain a more thorough comprehension of intricate biological data.

The article showcases the benefits of combining machine learning techniques with conventional statistical methods for assessing gene expression data. The utilization of Principal Component Analysis and Random Forest, in conjunction with Support Vector Machines and Neural Networks, improves both the accuracy of predictions and the capacity to understand and analyze the results. The results align with and expand upon prior research in the subject, establishing a strong foundation for future bioinformatics studies. Our approach tackles the difficulties posed by high-dimensional data and ensures the preservation of crucial features, so enabling more precise and practical understanding of gene expression and its impact on clinical outcomes. The integrated framework presented in this study is a valuable addition to the current endeavors of utilizing ML and traditional statistical methods in the field of bioinformatics. It offers a well-rounded and thorough approach to data analysis.

## Conclusions

In the article we aimed to construct and evaluate a combined machine learning (ML) with conventional statistic methods integrative analytics framework for an analysis of gene expression data on breast cancer populations. Results showed that this hybrid-importance method not only provides a superior prediction performance but also enhances the interpretability of results, allowing one to have an insight approach into data. In this conclusion, the major conclusions drawn from an evaluation of its results are recapitulated and the implications for practice and future research in Nigeria.

The analyzed gene expression profiles using three ML algorithms—Support Vector Machines , Random Forests and Neural Networks. The performance of each model was measured by the standard metrics (accuracy, precision, recall and f-score). However, SVM method performed the best through all these methods which indicated that the be power of computation behind it when working with complex problems

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910

like gene expression data. The RF model provides information regarding individual genes importance, which is a feature missing from SVM, making it complementary. Although linear model can indeed explain the relationship between Weight and Miles, NN being capable of capturing non-linearity increased complexity in modeling by catching many subtle patterns within data.

Then combined the Principal Component Analysis and linear regression into our framework to further interpret these ML models PCA has done its job well to reduce the dimensional data, retaining only important features from it. After subjecting the space of 150 dimensions to robust PCA as a necessary step in dealing with high-dimensional gene expression data, such that subsequent analyses would be both computationally feasible and biologically informative. Linear regression gave us quantitative results on the interactions among gene expressions and clinical outcomes, pulling out those key genes that significantly affect these phenotypes.

By combining feature importance scores from RF with PCA, we can identify genes that were important in both analyses. This hybrid strategy ensured that the genes chosen by importances were neither only typical predictor genes nor simply main features. This dual meaning is important in taking computational results forward to biological knowledge and emphasizes genes that are both statistically and biologically meaningful.

Using the two together had notable benefits, compared to employing just ML or standard statistical tools. The combination of the 2 resulted in a much more balanced and comprehensive analysis than using only one of them. The methodology integrated and complemented each other, offering a more complete picture of gene expression patterns and their associations with clinical outcomes compared to previous studies mostly using one analysis approach.

One of the major contributions focus on enriching biological investigation gene selection. Our method is not only practical to discover genes that are predictive and explanatory, but also useful for the development of drugs or diagnostic kits. For example, genes identified by both PCA and RF as highly significant can be further targeted for experimental validation studies to gain insights into the functional role of these genes in breast cancer.

Our findings have implications beyond breast cancer research. It can be extended to include other omics types (proteomics, metabolomics) allowing integrative cross-domain analysis. Such versatility demonstrates the wider range of applications our approach adds to bioinformatics research.

While chosen approach has its benefits, there are several potential areas of future work. However, some other limitations including the high computational cost on training neural networks can be another issue due to its resource requirement. One can imagine that future research might explore more effective algorithms or optimization techniques to help with this. In addition, although our proposed model is a combing of PCA and linear regression with ML models focused in this study, other statistical methods such as logestic or survival analysis could be incorporated for providing penetration insights.

This article provides a strong and complete paradigm to connect the ML strategies with legacy statistical approaches using gene expression data. This yielded a model with accuracy far surpassing that of traditional statistical analyses, without sacrificing explanatory power. Such an integrated strategy not only promotes the development of bioinformatics but also serves as a valuable resource to ensure that computational predictions inform biological understanding. The framework is also generalizable to diverse omics data types, expanding its utility and the possibility of future discoveries coming from it for biomedical research.

## References

A. M. Conard, A. DenAdel and L. Crawford, (2023): A spectrum of explainable and interpretable machine learning approaches for genomic studies. WIREs Computational Statistics, 15(5): e1617.

Journal of Ecohumanism
2024
Volume: 3, No: 5, pp. 335 – 352
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i5.3910
D. Feldner-Busztin, P. Firbas Nisantzis, S. J. Edmunds, G. Boza, F. Racimo, S. Gopalakrishnan, M. T. Limborg, L. Lahti and G. G. de Polavieja, (2023): Dealing with dimensionality: the application of machine learning to multi-omics data. Bioinformatics, 39(2): btad021.
C. Liu, X. Liu, H. Shangguan, S. Wen and F. Zheng, (2023): Review on the Application of Artificial Intelligence in Bioinformatics. Highlights in Science, Engineering and Technology, 30: 209-14.
H. Keshwani, Alisha and N. Jayapandian, (2022): Bioinformatics Research Challenges and Opportunities in Machine Learning. 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS): 290-95.
M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg and M. M. Hoffman, (2019): Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion, 50: 71-91.
K. A. Shastry and H. A. Sanjay, (2020): Machine Learning for Bioinformatics. Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications: 25-39.
Y. Li, F.-X. Wu and A. Ngom, (2018): A review on machine learning principles for multi-view biological data integration. Briefings in Bioinformatics, 19(2): 325-40.
M. S. Alam, A. Sultana, M. S. Reza, M. Amanullah, S. R. Kabir and M. N. H. Mollah, (2022): Integrated bioinformatics and statistical approaches to explore molecular biomarkers for breast cancer diagnosis, prognosis and therapies. PLOS ONE, 17(5): e0268967.
N. Auslander, A. B. Gussow and E. V. Koonin, (2021): Incorporating Machine Learning into Established Bioinformatics Frameworks. International Journal of Molecular Sciences, 22(6).
S. Kaptan and I. Vattulainen, (2022): Machine learning in the analysis of biomolecular simulations. Advances in Physics: X, 7(1): 2006080.
A. Serra, P. Galdi and R. Tagliaferri, (2018): Machine learning for bioinformatics and neuroimaging. WIREs Data Mining and Knowledge Discovery, 8(5): e1248.
B. S. Rao, S. Lavanya, K. Kajendran, P. P. Sharma, D. Verma, P. R and G. Manikandan, (2022): A Novel Machine Learning Approach of Multi-omics Data Prediction. 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES): 1-5.
R. S. Olson, W. Cava, Z. Mustahsan, A. Varik and J. H. Moore, (2018): Data-driven advice for applying machine learning to bioinformatics problems. Pac Symp Biocomput, 23: 192-203.
M. Khalsan, L. R. Machado, E. S. Al-Shamery, S. Ajit, K. Anthony, M. Mu and M. O. Agyeman, (2022): A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction. IEEE Access, 10: 27522-34.
B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung and P. Ping, (2019): Machine Learning and Integrative Analysis of Biomedical Big Data. Genes, 10(2).
A. Kaleem, G. S. Raju, G. S. L. B. V. Prasanthi, D. P. Patil, A. Naaz and S. K. Shukla, (2022): Machine Learning Science Using Bioinformatics Leads To More Effective Treatments. 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE): 2483-87.
K. Raina, (2023): Role of Bioinformatics in Analysing Big Data Using Statistical Computing and Computer Science. Interantional journal of scientific research in engineering and management.
P. S. Reel, S. Reel, E. Pearson, E. Trucco and E. Jefferson, (2021): Using machine learning approaches for multi-omics data analysis: A review. Biotechnology Advances, 49: 107739.
A. Wood, K. Najarian and D. Kahrobaei, (2020): Homomorphic Encryption for Machine Learning in Medicine and Bioinformatics. ACM Comput. Surv., 53(4): Article 70.
M. R. Karim, O. Beyan, A. Zappa, I. G. Costa, D. Rebholz-Schuhmann, M. Cochez and S. Decker, (2021): Deep learning-based clustering approaches for bioinformatics. Briefings in Bioinformatics, 22(1): 393-415.
R. A. Pyron, (2023): Unsupervised Machine Learning for Species Delimitation, Integrative Taxonomy, and Biodiversity Conservation. bioRxiv: 2023.06.12.544639.
Y. Cao, T. A. Geddes, J. Y. H. Yang and P. Yang, (2020): Ensemble deep learning in bioinformatics. Nature Machine Intelligence, 2(9): 500-08.
E. J. Draizen, J. Readey, C. Mura and P. E. Bourne, (2023): Prop3D: A Flexible, Python-based Platform for Machine Learning with Protein Structural Properties and Biophysical Data. bioRxiv: 2022.12.27.522071.