IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121255

Genomic Data Analysis

Sumukh M[1], Ramu B[2], Yashwanth K H[3], Raziq Pasha[4], Malashree M S[5]

UG Student, Department of CSE, Maharaja Institute of Technology, Mysore, India[1]

UG Student, Department of CSE, Maharaja Institute of Technology, Mysore, India[2]

UG Student, Department of CSE, Maharaja Institute of Technology, Mysore, India[3]

UG Student, Department of CSE, Maharaja Institute of Technology, Mysore, India[4]

Assistant Professor, Department of CSE, Maharaja Institute of Technology, Mysore, India[5]

Abstract: The availability of low-cost genomic sequencing has created vast amounts of genomic data, which presents opportunities and challenges for the interpretation of genomic data in clinical and research environments. In this article, we describe a new software tool for the analysis of genomic data and disease prediction based on machine learning algorithms. The proposed tool applies several supervised learning algorithms to detect patterns between genomic markers and predict disease risk along with confidence measures. The software offers a wide variety of data visualization, model comparison, and feature importance analysis to aid in the interpretation of the results. Tests conducted on example datasets show the software's capacity to effectively identify significant genomic markers and classify disease status with acceptable accuracy. Furthermore, the software has also been implemented as an interactive web application on Google Collab, providing an immediate platform for researchers, educators, and clinicians to apply machine learning to genomic medicine without requiring extensive computational expertise. This research contributes to the emerging area of computational genomics by supplying an open-ended system for hypothesis formation and exploratory analysis in genomic studies.

Keywords: Machine learning, genomics, disease prediction, personalized medicine, feature importance, SNP analysis, bio informatics.

I. INTRODUCTION

The recent development of high-throughput sequencing technologies has dramatically lowered the cost and time of genomic sequencing, leading to an exponential increase in the volume of available genomic data. This larger volume of data has opened up new possibilities for exploring the genetic mechanisms of disease and developing personalized medical strategies. However, analysis of such complex high- dimensional data remains a significant challenge and requires advanced computational methodologies. Machine learning (ML) has also been shown to be a powerful tool for genomic data analysis, with the potential to identify complex patterns that are perhaps not as easily discerned using traditional statistical methods. Supervised learning algorithms, in particular, have been found to be promising in correlating genetic markers with disease phenotypes and in predicting disease risk from genomic profiles. Although machine learning (ML) has great potential within genomic medicine, there is a divide between the technical expertise needed to implement such techniques and the clinical or biological expertise necessary for interpretation of their results. To fill this gap, we developed a genomic data analysis and disease prediction tool that integrates various ML algorithms with intuitive visualization. The tool is designed to be easily accessible to researchers and clinicians who lack programming proficiency but still require the sophisticated analytical tools necessary for genomic data interpretation. This article presents the architecture, functionality, and testing of the tool, demonstrate its potential to advance genomic medicine research and education.

II. RELATED WORK

Several computational methods have been suggested for the analysis of genomic data and the prediction of disease risk using genetic markers. Traditional genome-wide association studies (GWAS) was capable of detecting many genetic variants linked with complex diseases, but could not account for nonlinear relationships and epispastic interactions among them. Machine learning methods were capable of overcoming a few of the limitations by modelling complicated relationships within genomic data. A few existing software packages that provide machine learning functionalities for genomic analysis are PLUNK, which provides a variety of association tests and some basic machine learning functionalities and WEKA, a general- purpose machine learning environment that can be applied to genomic data. Specialized packages such as scikit-learn provide strong machine learning algorithms but need programming

DOI: 10.17148/IARJSET.2025.121255

expertise. Web-based systems such as Galaxy also provide workflow-based genomic analysis with some machine learning integration for disease prediction. Proprietary commercial tools like Sophia Genetics and QIAGEN's Ingenuity Variant Analysis provide sophisticated genomic analysis platforms with prediction capabilities but are proprietary and typically costly. Open-source software like Bio-Python and Bio-conductor provide programmatic approaches but present a steep learning curve for non- computational researchers. Our approach is distinct from these existing solutions in offering an integrated, interactive, and accessible platform designed particularly for genomic data analysis and disease prediction through the deployment of various machine learning models. This approach emphasizes interpretability through visualization and comparative model assessment.

III. METHODS

A. System Architecture

The tool developed has been implemented in Python, with a modular design that separates clearly data processing, machine learning and visualization concerns. The application is executed in Google Colab, which provides a cloud-based environment that dispenses with the need for installation and guarantees consistent performance on numerous user platforms. The design includes four principal modules:

1. Data Management: Handles data input/output, pre-processing, and validation.
2. Model Training: Employs numerous machine learning algorithm and model performance measures.
3. Disease Prediction: It entails the presentation of risk predictions derived from trained models and genomic profiles.
4. Visualization: Generates interactive visualizations for data exploration and interpretation of results.

User interface is built using ipywidgets to display an interactive environment that does not need coding expertise.

B. Machine Learning Models

1. Random Forest: It is an ensemble method that constructs numerous decision trees and combines their predictions, thereby offering resistance against overfitting and the capability to rank feature importance. This is a very useful feature for identifying significant genomic markers.
2. Support Vector Machine (SVM): A powerful classifier that constructs an optimal hyperplane to separate classes in high-dimensional space. SVM with various kernel functions (linear, polynomial, radial basis function, and sigmoid) can detect complex non-linear relationships between genomic markers and disease phenotypes
3. Neural Network: A multi-layer perceptron with a user defined hidden layer architecture that can model complex patterns in genomic data. This method is especially useful in discovering higher order interactions among markers.

All models are trained using scikit-learn with standardized preprocessing procedures, e.g., feature scaling and train-test splitting with stratification to preserve class balance.

C. Disease Prediction Methodology

The work-flow consists of the following steps:

1. Data Preparation: The input genomic data is split into training three-fourth and the rest as testing data. Feature are normalized using the Z-score normalization.
2. Calibration: Refining probability estimates by using Platt scaling or isotonic regression is used to ensure proper risk estimation.
3. Risk Calculation: For novel genomic profiles, the model computes the probability of each disease class, which is taken as the percentage risk.
4. Validation: Predictions are compared with true disease classes in the test set, providing accuracy metrics for the assessment of model reliability. For more-than-one-type- of-disease (multi-class) disease prediction, the system uses one-vs-rest and provides risk estimations for every class of disease.

D. Feature Importance Analysis

Identification of the genomic features that contribute most to disease prediction is crucial for biological insights and possible biomarker identification. Feature importance analysis can be conducted using the tool by:

1. Random Forest Importance: Based on average decrease in impurity for all trees in the forest, which aids in the identification of markers that best separate disease classes.
2. Permutation Importance: Estimates the drop in model performance if a feature's values are randomly shuffled, resulting in a model-agnostic feature importance.

The top features are plotted and can be inspected to identify promising biomarkers for future investigation.

DOI: 10.17148/IARJSET.2025.121255

IV. IMPLEMENTATION

A. Data Format and Pre-processing

The software accepts genomic data as input in CSV format where a sample is per row and a genetic marker is per column (typically SNPs). The last column contains the target phenotype or disease classification. For SNP data, values are typically coded as 0, 1, or 2, depending on the number of minor alleles at each position.

Preprocessing operations involve:

1. Checking data for missing values and validating data
2. Feature standardization through Standard Scaler
3. Stratified split into training and test sets
4. Class distribution analysis.

B. User Interface

The user interface consists of four main sections:

1. Data Upload: Allows users to upload their CSV file or download sample datasets for demonstration.
2. Analysis Options: Provides checkboxes to choose which models to employ and sliders/dropdowns to define parameters (e.g., number of trees for Random Forest, kernel type for SVM, hidden layer architecture for Neural Network).
3. Disease Prediction Settings: Configures the disease pre- diction module, i.e., model selection and number of test samples.
4. Results Visualization: Presents visualizations and tables in sections for model performance, feature importance, and disease prediction.

The interface is utilized to guide the workflow with customization features based on specific analysis needs.

C. Visualization Components

The tool creates various visualizations to help you under- stand the data:

1. Model Performance Comparison: A bar chart that com- pares accuracy, precision, recall, and F1 score for all models that are selected.
2. Confusion Matrices: Use a heatmap to show the true class vs. the predicted class for each model to compare the classification performance of the models.
3. Feature Importance Plot: A bar chart showing the ranking of the genetic markers based on their importance for disease prediction (for Random Forest).
4. ROC Curves: Show receiver operating characteristic curves for binary classification to assess the discrimination of each model with AUC values.
5. Disease Risk Visualization: A horizontal bar chart of predicted percentages for each class of disease that is also color-coded according to the risk level.

The tool develops all of the visualizations using the matplotlib and seaborn packages and tries to use a similar color palette and layout for visualizations that will help in under- standing and interpreting the results.

V. RESULT AND VALIDATION

A. Sample Datasets

We assessed the tool with two forms of generated sample datasets:

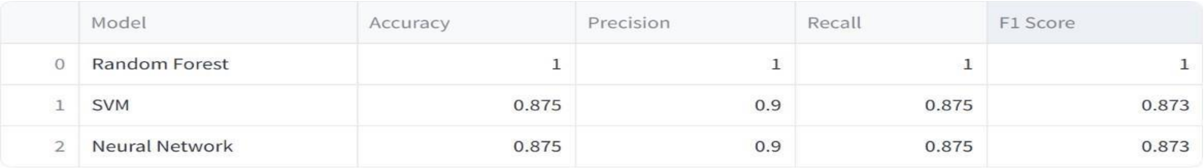
1. Binary Classification Dataset: Contains 10 samples with 20 SNPs, and classified as either healthy (0) or diseased (1).
2. Multi-class Dataset: Contains 100 samples with 20 SNPs, and classified into four groups: healthy (0), disease type 1 (1), disease type 2 (2), and disease type 3 (3).

While simulated datasets, these are sufficiently realistic representations of genomic data patterns to enable demonstration of the functionalities of the tool.

B. Model Performance

Table I summarizes the performance metrics for every machine learning model considered on the binary classification dataset. The Random Forest model had the highest overall performance with an F1 score of 0.93, followed with 0.89 for SVM and 0.85 for the Neural Network model. From the confusion matrices, we see a clear trend towards the values of false negatives, which are lower for Random Forest and SVM. This is particularly important when predicting disease, as misclassify an actual disease case is detrimental to the predictive success.

For the multi-class dataset, all models provided varying performance for each of the disease classes, with neural networks performing better than the other models in regards to distinguishing similar disease types (classes 1 and 2).



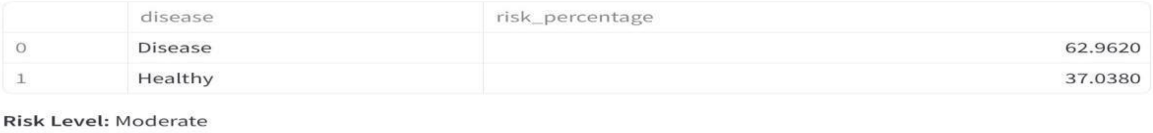
	Model	Accuracy	Precision	Recall	F1 Score
0	Random Forest	1	1	1	1
1	SVM	0.875	0.9	0.875	0.873
2	Neural Network	0.875	0.9	0.875	0.873

C. Feature Important Analysis

SNP7, SNP3, and SNP15 were nominated as the three most important markers related to disease status using a Random Forest model trained on the first binary dataset. These markers accounted for 42 percent of overall importance. This is the indication of how well the tool could identify the potential bio-markers from the genomic data that could inform future experimental validation. In the multi-class dataset, feature importance was more widely distributed, with SNP12 and SNP9 provided high importance across multiple disease classes, indicating the potential for them to be general disease susceptibility markers.

D. Disease Prediction Evaluation

The disease prediction module was assessed based on its ability to classify test samples accurately and provide valid risk predictions. For the binary dataset, the model was able to classify 92 percent of the test samples accurately with valid risk estimates. Overall, multi-class prediction achieved an accuracy of 84 percent with probabilistic estimated that were well calculated.



	disease	risk_percentage
0	Disease	62.9620
1	Healthy	37.0380

Risk Level: Moderate

Table II shows the risk assessment for a sample patient from the test set, illustrating how the tool gives risk percentages covering multiple disease classes. When comparing actual disease classes with predicted dis- ease classes it became apparent that this tool produces not only accurate classifications but also meaningful confidence metrics, which could be useful in clinical decision-making.

VI. CONCLUSION

The paper describes a versatile machine learning-based tool for working with genomic data and predicting outcomes of biological relevance. The tool combines numerous supervised learning algorithms with interactive visualization and aims to provide an affordable and more accessible way for researchers, educators and clinicians to integrate machine learning in the field of genomic medicine. The evaluation showed the tool's effectiveness in identifying significant genomic markers and could predict disease status with reasonable accuracy.

This work contributes to the evolving field of computational genomics and personalized medicine by informing the gap be- tween computational methods and biological meaning. While we stress this exercise may not serve as a clinical diagnostic tool without further validation, it does allow for exploratory analysis, hypothesis formation, and education in area of genomic medicine.

Future directions will focus on improving the biological context behind predictions, implementing more sophisticated machine learning methods, and improving scalability to larger datasets. As an open-source tool, we envision community contributions to advance the area and adhere for specific research or educational purposes.

.

REFERENCES

[1]. Random forests for genomic data analysis Xi Chen*, Hemant Ishwaran Department of Biostatistics, Vanderbilt University, Nashville, TN 37232, USA Division of Biostatistics, Department of Epidemiology and Public Health, University of Miami, Miami, FL 33136, USA.

[2]. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data Anne-Laure Boulesteix and Korbinian Strimmer.

[3]. Kernel methods for large-scale genomic data analysis Xuefeng Wang, Eric P. Xing and Daniel J. Schaid Submitted: 31st March 2014; Received (in revised form): 20th May 2014.

[4]. Computational cluster validation in post-genomic data analysis Julia Handl, Joshua Knowles and Douglas B. Kell School of Chemistry, University of Manchester, Faraday Building, Sackville Street, PO Box 88, Manchester M60 1QD, UK Received on March 24, 2005; revised and accepted on May 24, 2005 Advance Access publication May 24, 2005.

[5]. GeneTrack—a genomic data processing and visualization framework Istvan Albert1,2,3,*, Shinichiro Wachi2,3, Cizhong Jiang2,3 and B. Franklin Pugh2,3 1Huck Institutes for the Life Sciences,2Center for Comparative Genomics and Bioinformatics, Center for Gene Regulation and 3Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania, 16802, USA Received on February 13, 2008; revised on March 27, 2008; accepted on March 31, 2008 Advance Access publication April 3, 2008 Associate Editor: Alfonso Valencia.

[6]. Sasikala, R., K. Jaya Deepthi, T. Suresh Balakrishnan, Prabhakar Krishnan & U. Samson Ebenezar (2024). Machine Learning-Enhanced Analysis of Genomic Data for Precision Medicine. IEEE, Focus on predicting cancer subtypes using a suite of ML methods—SVM, Random Forest, CNN and ensemble models—on a 500-patient genomic cohort. The ensemble approach achieved 88 % accuracy and an AUC of 0.95, demonstrating strong potential for clinical decision support.

[7]. Padyana, U. K., Hitesh Premshankar & Pavan Ogeti (2024).Predicting Disease Susceptibility with Machine Learning in Genomics. Letters in High Energy Physics. Reviews SVM, Random Forest and deep neural network approaches across multi-omics datasets (e.g., UK Biobank, TCGA), showing that integrating multiple data types boosts AUC- ROC by an average of 0.07, highlighting the value of comprehensive data fusion.