scientific reports

Check for updates

OPEN

Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine

William DeGroat[1], Habiba Abdelhalim[1], Kush Patel[1], Dinesh Mendhe[1], Saman Zeeshan[2] & Zeeshan Ahmed[1,3]✉

Personalized interventions are deemed vital given the intricate characteristics, advancement, inherent genetic composition, and diversity of cardiovascular diseases (CVDs). The appropriate utilization of artificial intelligence (AI) and machine learning (ML) methodologies can yield novel understandings of CVDs, enabling improved personalized treatments through predictive analysis and deep phenotyping. In this study, we proposed and employed a novel approach combining traditional statistics and a nexus of cutting-edge AI/ML techniques to identify significant biomarkers for our predictive engine by analyzing the complete transcriptome of CVD patients. After robust gene expression data pre-processing, we utilized three statistical tests (Pearson correlation, Chi-square test, and ANOVA) to assess the differences in transcriptomic expression and clinical characteristics between healthy individuals and CVD patients. Next, the recursive feature elimination classifier assigned rankings to transcriptomic features based on their relation to the case–control variable. The top ten percent of commonly observed significant biomarkers were evaluated using four unique ML classifiers (Random Forest, Support Vector Machine, Xtreme Gradient Boosting Decision Trees, and k-Nearest Neighbors). After optimizing hyperparameters, the ensembled models, which were implemented using a soft voting classifier, accurately differentiated between patients and healthy individuals. We have uncovered 18 transcriptomic biomarkers that are highly significant in the CVD population that were used to predict disease with up to 96% accuracy. Additionally, we cross-validated our results with clinical records collected from patients in our cohort. The identified biomarkers served as potential indicators for early detection of CVDs. With its successful implementation, our newly developed predictive engine provides a valuable framework for identifying patients with CVDs based on their biomarker profiles.

Abbreviations

AF	Atrial fibrillation
AI	Artificial intelligence
AML	Acute myeloid leukemia
AUC	Area under the curve
ANOVA	Analysis of variance
CAD	Coronary atherosclerosis heart disease
CVD	Cardiovascular disease
CIGT	Clinically integrated genomic and transcriptomic
HER	Electronic health record

[1]Health Care Policy and Aging Research, Rutgers Institute for Health, Rutgers University, 112 Paterson St, New Brunswick, NJ 08901, USA. [2]Rutgers Cancer Institute of New Jersey, Rutgers University, 195 Little Albany St, New Brunswick, NJ, USA. [3]Department of Medicine/Cardiovascular Disease and Hypertension, Robert Wood Johnson Medical School, Rutgers Biomedical and Health Sciences, 125 Paterson St, New Brunswick, NJ, USA. ✉email: zahmed@ifh.rutgers.edu

GWAS	Genome-wide association studies
HF	Heart failure
IRB	Institutional review board
K-NN	K-nearest neighbor
MI	Myocardial infarction
ML	Machine learning
NGS	Next-generation sequencing
RF	Random forest
RFE	Recursive feature elimination
ROC	Receiver operating characteristic
SVC	Soft voting classifier
SVM	Support vector machine
WES	Whole exome sequencing
WGS	Whole genome sequencing
XGBoost	Xtreme gradient boosting

Artificial intelligence (AI) and machine learning (ML) encompasses a plethora of supervised and unsupervised methodologies for scrutinizing genomics data, culminating in the formation of multivariate statistical instruments[1]. The proficient implementation of AI/ML techniques holds the promise of fostering an augmented comprehension of diseases at the systemic level, unveiling the intricacies of genomic regulatory networks. By leveraging AI/ML approaches, clinical and genomics data can undergo statistical analysis and classification, enabling the prediction of high-risk patients. AI/ML can be deployed to capture genetic sequences associated with chronic diseases, categorize phenotypes based on knowledge about human diseases and establish population dimensions for rare diseases[1,2]. Genetic studies have facilitated disease prognosis[3,4], the identification of genetic regions and variants that influence disorders, and the functional assessment of these regions[5–7]. While holding great prospects, the formidable task at hand lies in analyzing the immense magnitude of recognized (and unrecognized) genetic variations and leveraging this knowledge to facilitate diagnosis, ascertain risk, and forecast treatment responses among heterogenous human populations[8]. This challenge is being addressed through precision medicine which encompasses the integration of clinical and genomics data to enable predictive treatment within a diverse cardiovascular disease (CVD) population[5]. The primary objective of personalized medicine is to analyze a patient's genetic makeup to identify crucial biomarkers and enhance comprehension of the underlying pathophysiology of intricate disorders such as CVD[6].

The American Heart Association states that approximately 82.6 million individuals in the U.S. presently suffer from one or more types of CVDs, establishing it as a primary factor behind mortality in both males and females[9]. Common types of CVDs include stroke, congestive heart failure, coronary heart disease, and hypertension[10,11]. Considering the intricate nature, risk factors, inherent genetic composition, and trajectory of CVD, personalized treatment is considered indispensable[12]. Moreover, progress in genomics has significantly contributed to comprehending the molecular pathways linked to the prevalence of CVDs[3]. These advancements were propelled by next-generation sequencing (NGS), which enabled the discovery of novel genetic correlations and the capacity to assess genetic diversity among patients[13]. Recent developments in the field of genomics and bioinformatics have greatly aided in better understanding the complex nature of CVD etiology. However, the development of an AI/ML predictive engine that utilizes genetic biomarkers to assess the risk of CVD in patients is still in its early stages[14–16]. Recent studies have explored the potential of employing AI/ML algorithms on whole genome and whole exome sequencing (WES/WGS) data for statistical and prognostic analyses for a wide variety of diseases including but not limited to Crohn's disease[17], inflammatory bowel disease[18], breast cancer[19], colon cancer[20] and Alzheimer's disease[21].

Previously, we have created AI/ML models to investigate and identify genes associated with heart failure (HF), atrial fibrillation (AF), and other CVDs and successfully predict these diseases with high accuracy[22]. However, one of the major limitations of our and most of the other published disease specific research using AI/ML and bioinformatics approaches is the focus on genes known to be associated with disease[2,22,23]. In this study, we propose a new AI/ML model that adapts an innovative nexus of algorithms to predict CVDs using critical transcriptomic biomarkers determined using our comprehensive statistical analysis (Fig. 1). Our model is trained on an AI/ML ready dataset of whole transcriptome-based gene expression and clinical data of consented individuals. We observed novel as well as known biomarkers that were associated with CVDs, relative to our previous model[22]. We demonstrate that our current model can produce accurate predictions regarding CVD diagnosis. By identifying specific biomarkers, we have unveiled a crucial set of potential indicators for the early detection of CVDs. These biomarkers provide essential clues in identifying at-risk patients before symptoms manifest, allowing for timely intervention and improved patient outcomes. With the successful implementation of our newly developed predictive engine, healthcare professionals now have access to a valuable framework that utilizes biomarker profiles to accurately identify patients at risk of CVDs.

Material and methods

Our study is divided into two major steps: (I) identification of significant biomarkers, and (II) implementation of nexus AI/ML models for predictive analysis (Fig. 1).

Identification of significant biomarkers

We utilized a convergence of statistical algorithms to evaluate the variations in expression levels and clinical characteristics between individuals with CVDs and those that are healthy. The proposed feature selection model

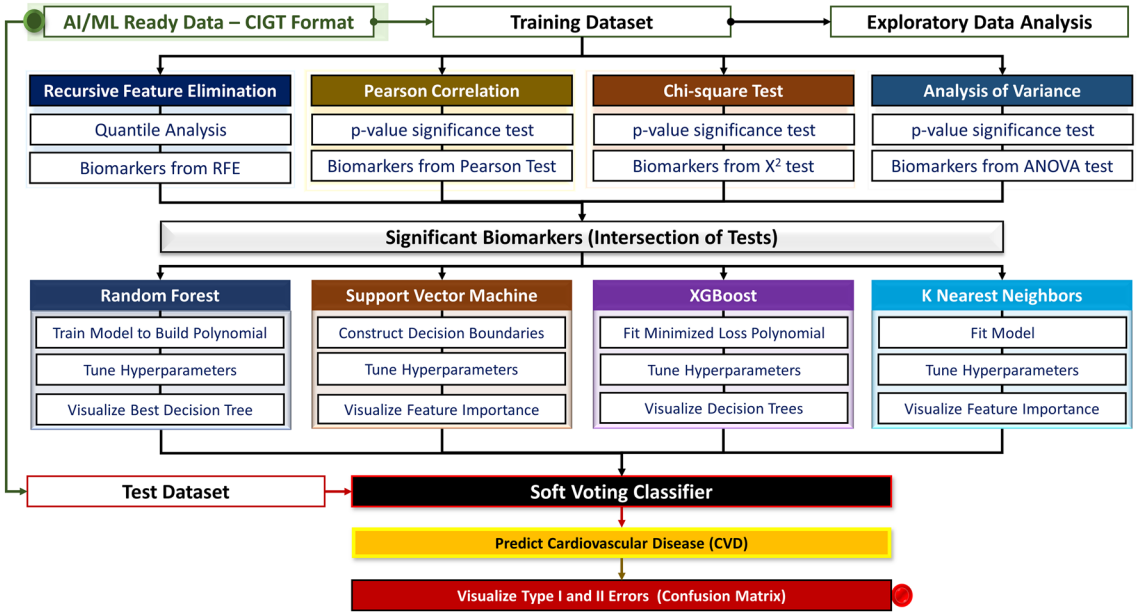


Figure 1. Methodology and study design, workflow, and bioinformatics. This figure presents implemented statistical tests (Recursive Feature Elimination, Pearson correlation, Chi-square test, and Analysis of Variance) for the exploratory data analysis to assess the differences in genomics and phenotypic features between healthy individuals and patients with CVD and observe significant biomarkers. Next, applied a nexus of Machine Learning (ML) algorithms (Random Forest, Support Vector Machine, Xtreme Gradient Boosting Decision Trees, and k-Nearest Neighbors) to predict CVD. In addition, it includes Training Dataset, Test Dataset, Soft Voting Classifier, and Visualization of Type I and II errors.

uses four distinct algorithms: (I) Recursive Feature Elimination (RFE)[24], (II) Pearson Correlation[25], (III) Chi-Square Test[26], and IV) Analysis of Variance (ANOVA)[27]. A combination of these tests allows the model to adapt to different matrix sizes, distributions, and attributes. All these algorithms used our CIGT dataset to compute the statistical significance of supported biomarkers by means of a *p* value significance test.

To eliminate biomarkers that do not have high significance to CVD and reduce the computational load for the analysis downstream, we applied the RFE algorithm[28]. In our study, we chose the scoring metric to be based on decision trees with top 10% number of features to be from the original list of biomarkers. The correlation coefficient plays a crucial role in ranking: the higher the coefficient, the higher the rank assigned to the gene, implying a stronger association between the gene and CVD. It is important to note that a higher rank corresponds to a lower integer value. To determine each biomarker's linear relationship to disease, we applied the Pearson correlation test where each biomarker was assigned a correlation coefficient. Subsequently, to examine the dependence between the test variable and the significant biomarkers, we employed the chi-square test. The chi-squared test has been applied widely in genomics for feature selection due to its application in multi-disease classification for genome-wide association studies (GWAS)[29]. The SelectKBest function is used to select the top 'k' (k = 10) features on univariate statistical tests, in this case, the chi-squared test. Next, we implement the ANOVA procedure, which uses a five-step approach to compute a f-statistic that determines the significance of a biomarker. We chose selectors that could easily be merged into a single scoring metric to select supported biomarkers for downstream analysis. Statistical methods that produce *p* values and ML selectors which provide rankings were favored to methods like principal component analysis, uniform manifold approximation, and projection, and t-distributed stochastic neighbor embedding that do not offer feature importance.

There are documented limitations associated with each testing algorithm utilized in our study. To address these challenges, we have merged these algorithms to satisfy different requirements. RFE cannot quantify the correlation between biomarkers and lacks the ability to compute multivariate significance. Furthermore, due to its iterative nature, RFE has a high time complexity[25]. One of the main limitations of the Pearson correlation test is the sensitivity to range differences between the biomarkers and their relation to disease. However, we have accounted for this by increasing the volume of data to reduce range differences between biomarkers. The main challenge associated with the chi-square test is the number of Type I and II errors in small sample sizes. However, the rationale for implementing this algorithm was to make our overall system predict better in larger matrix sizes. A challenge that arises with ANOVA testing is the fact that if two groups of samples are of different sizes, then there is a direct issue with the strength and validity of the test. Due to the inclusion of all the other algorithms that can handle imbalances in sample size, this limitation is not of concern to this study. In our merged function, we select the statistically significant biomarkers for the ANOVA, chi-square, and Pearson correlation test and show up in the top 10% of significant biomarkers in RFE.

Implementation of a nexus AI/ML models for predictive analysis

The biomarkers selected were predictive for patient diagnosis and classification. We selected four algorithms for this task: Random Forest (RF)[30], Support Vector Machine (SVM)[31], K-nearest neighbors (k-NN)[32], and Extreme Gradient Boosting Decision Trees (XGBoost)[33]. We applied hyperparameter tuning to all algorithms, which were then ensembled using a Soft Voting Classifier to curate a powerful predictive engine that can perform accurate classification specific to user-specified matrices.

We started with RF, which is a meta-classifier that combines the output of multiple decision trees to categorize individuals based on their disease state. The algorithm computes a decision tree to classify patients based on their biomarker profile. The best decision tree from the forest was considered which highlights the decision boundary (i.e., polynomial) that the algorithm uses to sort patients. To classify patients based on their biomarker profile, we implemented SVM that computes support vectors. The most important classification feature highlights the relative significance of each biomarker. To further classify patients based on their biomarker profile and address limitations associated with SVM, we used the XGBoost algorithm. This algorithm computes a decision tree to highlight biomarkers that were of significance in the classification process. Finally, we applied the k-NN algorithm to determine the classification of a datapoint by majority voting amongst its 'k' nearest neighbors. The k-value was chosen based on iterating through all possible values of k and selecting the model with the highest accuracy.

Employing this nexus of ML algorithms helped us in navigating shortcomings that might arise from individual algorithms. The main limitation of SVMs is their inability to perform well when the data set is large[31]. However, through a combination of algorithms, SVMs can be an integral part of an ML system when the input set is small. Another limitation arises in the implementation of XGBoost where the performance is greatly diminished on sparse and unstructured data[33]. However, due to our robust data pre-processing function, we have been able to address this issue. The main limitation of k-NN is the sensitivity to feature scaling[32]. KNN calculates distances between instances to determine their similarity. If features have different scales, those with larger values can dominate the distance calculation, leading to biased results. It is essential to normalize or scale the features appropriately before applying KNN. However, KNN can adapt to changes in the training data without requiring complete retraining of the model, which is why it was selected for our analysis.

All four algorithms were ensembled using the Soft Voting Classifier, the class with the highest average probability of success was chosen as the final prediction. By combining each algorithm in this manner, the positives are accentuated while neutralizing the downsides for each algorithm.

Ethical approval and consent to participate

Informed consent was obtained from all subjects. All human samples were used in accordance with relevant guidelines and regulations, and all experimental protocols were approved by the Institutional Review Board.

Results

Building suitable cohorts

Substantiating our approach towards discovering disease-relevant biomarkers effectively to predict patients' diagnostic status necessitated creating a comprehensive dataset to represent our patient cohort. The cohort consisted of 61 CVD patients, including 40 males and 21 females, aged 45–92. The participants self-identified their race as follows: 42 were white, 7 were black or African American, 1 was Asian, and 11 were of unknown race. These individuals were clinically diagnosed with CVDs, specifically Heart Failure (HF), and Atrial Fibrillation (AF). In addition, we constructed a control group comprising 10 healthy individuals, evenly split between males and females. Among them, 9 identified as white, and 1 did not disclose their race. The age range of this group was 28–78 years. A persistent challenge in multi-genomic data analysis lies in the integration and standardization of large volumes of sequence data[2]. Currently, processed gene expression and variant data available through genomic pipelines are not available in AI/ML ready formats[2]. With its availability as AI/ML input, it can be used directly for predictive analysis[2,34,35]. To address this challenge, we propose the Clinically Integrated Genomics and Transcriptomics (CIGT) format, which integrates heterogeneous clinical, demographic, genomic and transcriptomic patient data. Due to the limited clinical history of our cohort, we focused on patient information such as age, gender, racial, and ethnic background, and gene expression data derived from RNA-seq. These attributes have shown their effectiveness in the development of genotype–phenotype studies[34]. In the future, attributes in the CIGT dataset could be expanded to integrate variant data as well as include more clinical attributes including but not limited to medications and risk factors such as smoking and alcohol consumption.

All procedures involving human participants were in accordance with the ethical standards of the institution and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. All human samples were used in accordance with relevant guidelines and regulations, and all experimental protocols were approved by the Institutional Review Board (IRB) of Rutgers. Utilizing our proposed CIGT format, we integrated transcriptomics, clinical, and demographics data of each patient (Supplementary Material 1). Data pre-processing increased our cohort's strength through the elimination of non-ubiquitous patient attributes; features present in 80% of the cohort were included and the less occurring were eliminated from the CIGT dataset to avoid extrapolation from ML classifiers downstream. Resulting from this filtration, 751 transcriptomic and clinical biomarkers remained suitable. The CIGT dataset was subset into training and testing sets, with a testing size of 30%.

Discovering supported biomarkers

Statistical algorithms were applied on the training dataset to retrieve highly significant biomarkers. To assess the differences in expression levels and clinical characteristics across CVD patients and healthy individuals, we employed a convergence of four statistical algorithms: (I) Recursive Feature Elimination (RFE), (II) Pearson Correlation, (III) Chi-Square, and IV) Analysis of Variance (ANOVA) (Fig. 2). To ascertain the statistical

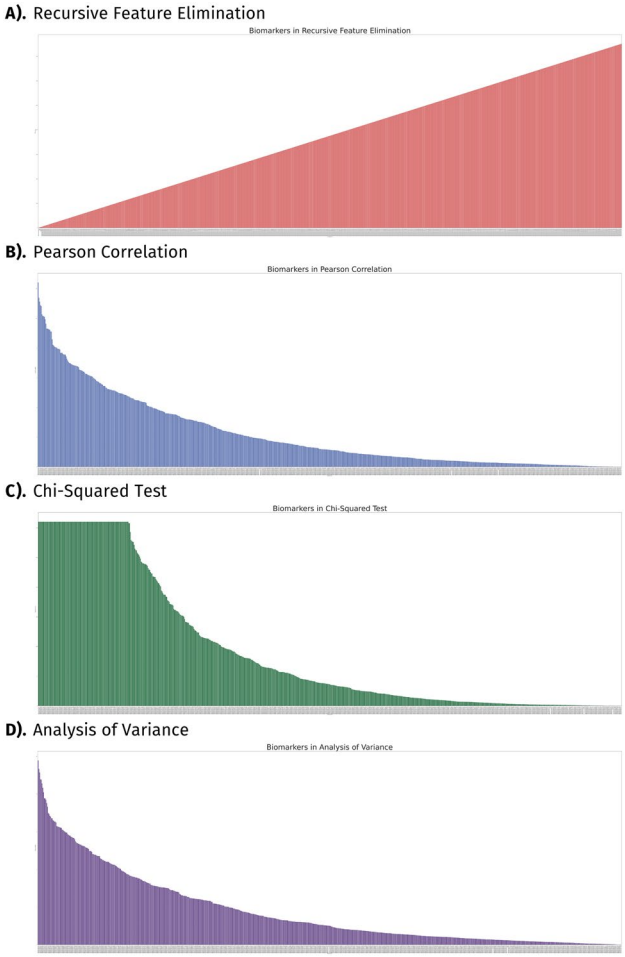


Figure 2. Feature selection of biomarkers. This figure presents the statistical significance test to determine the importance of each gene according to the algorithm used. The y-axis represents the *p* values as a logarithmic expression while the x-axis displays distinct biomarkers. Features are displayed from (**A**) Recursive Feature Elimination; (**B**) Pearson Correlation; (**C**) Chi-Squared test; and (**D**) Analysis of Variance.

significance of each algorithm, we conducted a *p* value significance test and recorded the obtained *p* values in a list together with the raw scores generated by each algorithm (Supplementary Material 2). We exercised the scientific standard of 0.05 as a threshold for our statistical significance test and utilized the logarithmic function, with a base of 10, for easier interpretation.

RFE systematically eliminated the least informative features, which enabled the identification of the strongest correlations between biomarkers and CVD. The RFE algorithm assigned scores to each feature, reflecting their relative importance, with higher scores indicating lesser significance. These scores were then utilized to rank the features based on their relevance to CVD diagnosis (Fig. 2A). Next, the Pearson correlation test was applied to quantitively assess the magnitude of linear association between biomarkers and CVD. In our study, we observed the correlation coefficient, which ranges from −1 to 1, with larger absolute values indicating a more pronounced association. However, to assess the statistical significance of the findings, we also examined the negative logarithm of the *p* value for both transcriptomic and clinical features (Fig. 2B). Notably, higher bars in the graph indicate greater significance to CVD diagnosis.

We applied the chi-square test to investigate the independence among categorical factors on CVD detection and discern any significant relationships that may exist (Fig. 2C). We calculated the chi-square statistic to quantify this independence (Supplementary Material 2). We utilized the ANOVA test to discern the difference in the distribution of gene expression patterns between healthy individuals and those afflicted with CVD (Fig. 2D). We computed the F-statistic to measure this variability (Supplementary Material 2). We found 313 biomarkers to be supported across three of our algorithms (Pearson correlation, chi-square test, and ANOVA). The presence of high outliers, such as genes *HBA1* and *HBA2*, which are beneficial in traditional selection methods but detrimental to predictive model training, diminishes importance within our RFE classifications. To counterbalance precursory approaches to subset our biomarkers, we implemented RFE. Biomarkers classified within the top 10% were endorsed for further predictive analysis (Table 1).

Ensembl ID	Recursive feature elimination score	Correlation coefficient	Pearson correlation (*p* value)	Chi-square statistic	Chi-square test (*p* value)	F-statistic	Analysis of variance (*p* value)
ENSG00000266422	8	0.573204861	1.42E−07	6099.039146	0	18.4616809	8.41E−05
ENSG00000242574	27	0.468662916	3.30E−05	1182.198479	4.51E−259	17.33140061	0.000129594
ENSG00000256618	1	−0.498577843	8.30E−06	425.0570428	1.94E−94	15.44622846	0.000271697
ENSG00000265150	10	0.501748748	7.12E−06	5570.193207	0	14.6231818	0.000378483
ENSG00000234745	41	0.44430813	9.24E−05	21,800.54816	0	13.25033749	0.000665893
ENSG00000241553	29	0.437526155	0.000121446	967.241151	2.37E−212	12.82521109	0.000795751
ENSG00000256514	13	−0.422350763	0.000219405	97.15855608	6.40E−23	12.5163011	0.000906631
ENSG00000231389	46	0.415749505	0.000281356	2762.649364	0	12.50820118	0.000909747
ENSG00000239998	35	0.437466127	0.000121737	467.7152611	1.01E−103	11.28761072	0.001536451
ENSG00000234741	42	0.38109307	0.000957704	250.754169	1.78E−56	10.14411903	0.002543671
ENSG00000247596	20	0.378112312	0.00105766	169.360342	1.02E−38	10.13146467	0.002558096
ENSG00000215845	66	0.318411748	0.006413323	324.0418477	1.91E−72	9.419225469	0.003526625
ENSG00000269858	5	0.393315171	0.000631198	199.9036854	2.19E−45	9.331682275	0.003670018
ENSG00000233276	43	−0.38130551	0.000950918	286.051535	3.61E−64	6.823535203	0.011973983
ENSG00000245910	21	0.290124517	0.013431239	146.3023238	1.11E−33	6.440924863	0.01445292
ENSG00000227097	53	0.256310109	0.029761901	3696.999979	0	5.590552265	0.022150113
ENSG00000254999	14	0.271571684	0.021022304	105.5014956	9.48E−25	5.208092813	0.026955423
ENSG00000260592	11	0.314078232	0.007215015	45.01668698	1.95E−11	4.491244041	0.039268284

Table 1. Statistical analysis of significant biomarkers. Table 1 includes rankings based on Recursive Feature Elimination scores, Pearson correlation, chi-square, and Analysis of Variance test. All raw scores for are included (correlation co-efficient, chi-square statistic, and f-statistic) as well as *p* values that were utilized in the visualization and artificial intelligence/machine learning (AI/ML) analysis of the data.

Predicting cardiovascular disease

Transcriptomic attributes serve as our predictive engine's training dataset. This engine consists of five unique classifiers to forecast case/control predictions for our testing dataset: Random Forest (RF), Support Vector Machine (SVM), Xtreme Gradient Boost (XGBoost), k-Nearest Neighbor (k-NN), and Soft Voting Classifier (SVC). Metrics, including weighted-average F1 scores and receiver operating characteristic curves (ROC), were calculated for each classifier. Weighted-average F1 scores evaluate models in circumstances where categorical predictors are not balanced. ROC-AUC provides an additional approach to ML performance evaluation, showing a probability of a binary classifier to make true predictions rather than false positives. Values approaching 1.0 in each measure suggest high performance. Exact metrics such as accuracy, ROC-AUC and weighted average F1 scores for each algorithm are provided in Supplementary Material 3.

RF has demonstrated practical usage within transcriptomics[23]. Optimizing RF with GridSearchCV (Fig. 3A), using dataset-standard parameters, the decision tree classifier made the most accurate predictions. RF selected case/control correctly in 95% of testing patients. Important features involved in RF prediction include *RN7SL593P, LILRA2*, and *HLA-B* (Fig. 3A). ROC-AUC for our RF classifier was 0.95. The weighted-average F1 score was 0.96. SVM, a classifier suited for single-diagnosis case/control predictions, performed satisfactorily. Optimized using GridSearchCV using dataset-standard parameters (Fig. 3B), the SVM classifier succeeded with 91% of predictions. *MTRNR2L1, GPX1*, and *AP003419.11* are the SVM classifier's most essential features. This model's ROC-AUC was the highest, 0.99. The SVM classifier's weighted-average F1 score was 0.91. XGBoost, another decision tree-based approach, provides an accessible approach to classification. The performance of XGBoost rivals our SVM classifier, scoring 91% on predictions. XGBoost was optimized with GridSearchCV using dataset-standard parameters (Fig. 3C). XGBoost's best tree functioned using *MTRNR2L1* as its sole feature. XGBoost's ROC-AUC was 0.94. The XGBoost classifier's weighted-average F1 score is 0.91. k-NN's performance was feeble compared to RF, SVM, and XGBoost. Tuned with GridSearchCV using dataset-standard parameters (Fig. 3D), the k-NN classifier hit 91% of predictions. This pairs with 0.85 ROC-AUC and 0.91 weighted-average F1 score. k-NN is a resource-intensive algorithm, producing worse performance at extended runtimes compared to our previous classifiers. k-NN used *MTRNR2L1, BRK1*, and *ARPC4* most when forming predictions.

RF and XGBoost classifiers proved most applicable to transcriptomic datasets. SVM performance is sufficient for case/control classifications, but diverse problems engaging multiple diseases and disorders will lead to substantial performance declines[5]. k-NN is the least appropriate for such datasets. *MTRNR2L1* was the best transcriptomic marker for CVD predictions, with top-three importance for our SVM, XGBoost, and k-NN classifiers. We employed hyperparameter tuning to each algorithm and combined them through a Soft Voting Classifier to create a robust predictive engine capable of accurately classifying data based on user-defined criteria. Our ensemble model was able to accurately classify seventeen individuals as CVD patients and three individuals as healthy. It also had two incorrect classifications where one was a false positive and the other a false negative (Fig. 3E). Identifying the intersectionality between the four classifiers' (RF, SVM, XGBoost and k-NN) most important biomarkers, we generated a non-traditional Venn diagram (Fig. 3F). The five most significant biomarkers were extracted from each classifier. Methods that relied on less than five biomarkers (RF, 4; XGBoost,

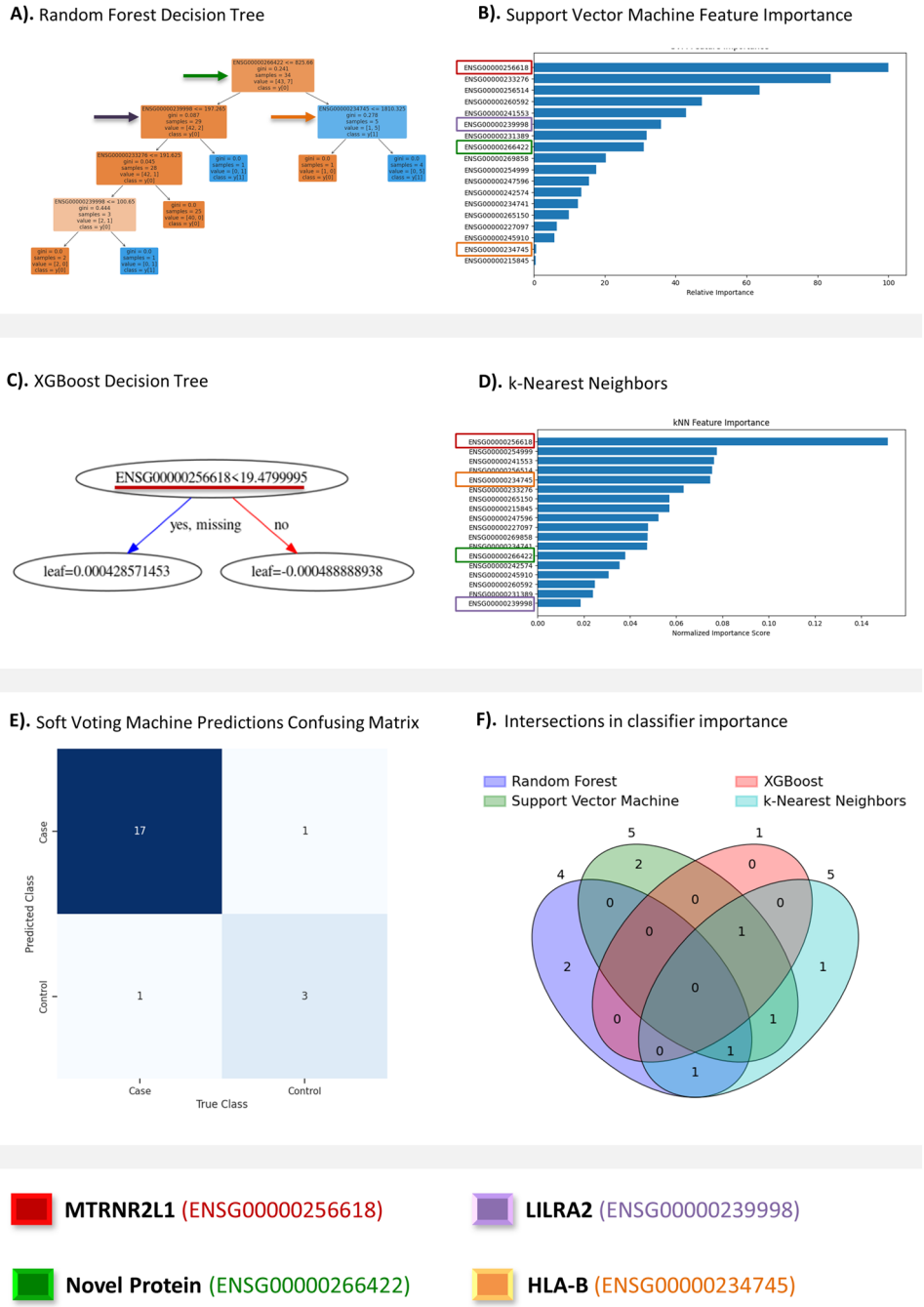


Figure 3. Results of predictive analysis. This figure presents results of AI/ML based predictive analysis and that include, (**A**) Random Forest decision tree; (**B**) Support vector machine feature importance; (**C**) XGBoost decision tree; (**D**) k-Nearest neighbors; (**E**) soft voting classifier predictions confusing matrix; and (**F**) Venn diagram detailing the counts of overlapping between the top five biomarkers from each classifier.

1) had only those included. This visualization illustrates which classifiers relied on similar biomarkers to others to make their predictions.

Examining transcriptomic predictors

Validating the detected biomarkers' relevance to our cohort's diagnoses necessitated an in-depth inspection of their function in prediction and prominence in previous literature. Alongside a thorough review of previous scientific findings, biomarkers correlations are reported and tied to their roles in disease classification. The literature review revealed 14 transcriptomic biomarkers linked with CVDs and a variety of other diseases and disorders within our cohort. *HLA-DMB* and *HLA-B* are associated with cardiomyopathy. *RN7SL2* and *GPX1* are associated with stroke. *ARPC4* and *LILRA2* are associated with atherosclerosis. Transcriptomic markers

(Fig. 4A) found within the supported list are also associated with various types of chronic diseases) and disorders (cancers, rheumatoid arthritis, and diabetes. Visualizations displaying clustered profiles of transcriptomic expression (Fig. 4B) and their associations with biomarker's intercorrelation (Fig. 4C) indicate the mechanisms of disease classification. This correlation metric was supported using literature as well. Genes *TWF2* and *ARPC4* scored perfect correlations.

Pseudogene *MTRNR2L1* was the observed feature in all three classifiers: SVM, XGBoost, and k-NN. *MTRNR2L1* presented fluctuating expression across patients and failed to surpass a correlation above 0.5 with

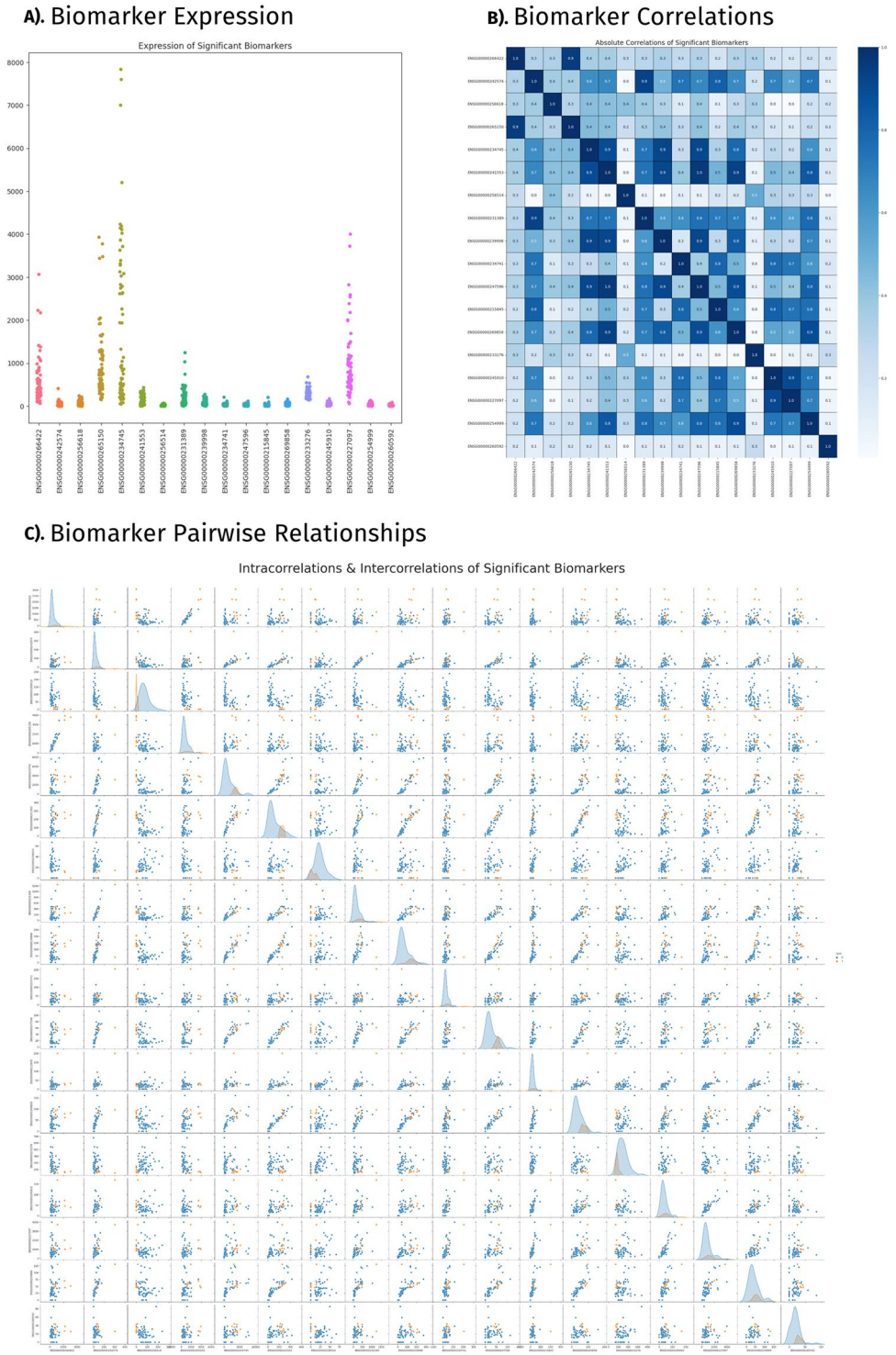


Figure 4. Significant biomarkers. This figure presents results of the statistical analysis and that include, (**A**) Biomarker expression; (**B**) Biomarker correlations; and (**C**) Biomarker pairwise relationships.

other transcriptomic biomarkers. *GPX1, AP003419.11*, and *CTA-363E6.6* were the three most important features of the SVM classifier beside the previously mentioned *MTRNR2L1*. *MTRNR2L1* and *GPX1* have been linked to CVDs, while *AP003419.11* and *CTA-363E6.6* have not been previously reported. These three transcriptomic markers are the least correlated with each other, the most independent function biomarkers within our list. The SVM classifier, more than others, is reliant upon independent-acting transcriptomic factors whose expression is not tied to that of another biomarker within the selected list. A cluster of highly correlated biomarkers identified, *RPS28P7, SNHG6*, and *TSTD1*, did not perform well with SVM classifier. The k-NN classifier did not display similar patterns regarding the correlation of transcriptomic biomarkers.

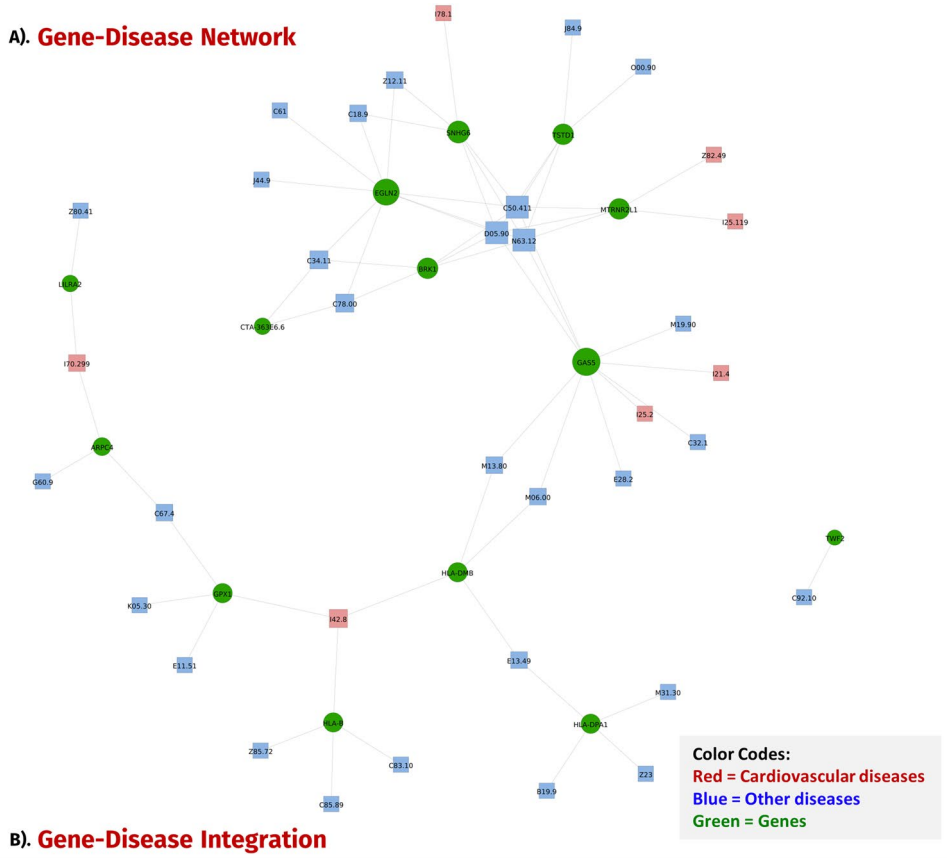
The XGboost classifier was reliant solely on *MTRNR2L1*, indicating the strongest association to CVDs of any transcriptomic biomarker. This algorithm ties the under expression of the lncRNA with CVD status. The RF classifier relied most prominently on the *RN7SL593P* biomarker, classifying patients below the threshold of 825.66 TPM as CVD cases. The overexpression of *RN7SL593P* has been linked to normal platelet function, a non-direct implication with CVDs. The RF classifier also placed heavy importance on *LILRA2, HLA-B,* and *GPX1* with direct links to CVDs. The decision tree algorithms contained only elements previously associated with CVDs within their optimized tree using our hyperparameter tuning metrics.

MTRNR2L1, RN7SL593P, LILRA2, and *HLA-B* showed the most distinct variety in their importance throughout the different classifiers. *MTRNR2L1*, scored the most important across three classifiers, but was not found in RF's decision tree. *LILRA2* and *HLA-B* scored a correlation of 0.9, near perfect. *HLA-B* ranked as the fifth most important feature in our k-NN classifier and the second least important in the SVM classifier. *LILRA2* placed as the sixth most important feature in our SVM classifier and last in our k-NN classifier. *RN7SL593P,* the workhorse of random forest, served average throughout the remaining classifiers. These incongruencies are algorithmically dependent but may offer some understanding of underlying biological interactions between these biomarkers and CVD.

Discussion

A persistent challenge in genomic data analysis lies in the handling and integration of large volumes of sequencing data[36]. With the implementation of our novel CIGT AI/ML ready dataset, we have begun to make significant progress to standardize heterogenous data types (genomic and clinical) for more accurate and reliable data analysis and interpretation[37]. Our novel AI/ML methodology uncovered eighteen transcriptomic biomarkers to be linked to CVDs, three of which were novel (*RN7SL593P, AP003419.11*, and *CTA-363E6.6*) and will require further analysis to understand the correlation between them and disease etiology. To further investigate gene-disease relationships for these significant biomarkers, we performed a literature review correlating these genes to CVDs and developed a gene-disease network (created using the 'igraph' Python package[38]) (Fig. 5). Genes such as *HLA-DMB*[39], *HLA-B*[40], and *GPX1*[41] were found to be profoundly expressed in cardiomyopathy. While other biomarkers such as *RN7SL2*[42], *LILRA2*[43], *GAS5*[44], *TWF2*[45], *EGLN2*[46], *SNHG6*[47–49], and *BRK1*[50] have all been previously associated with phenotypic variations linked to CVD, there is limited literature associating protein-coding genes such as *RPS28P7 and CTA-363E6.6* to other known CVDs. No direct links were recorded between *RN7SL593P* and *AP003419.11* and known CVDs as well as other non-CVD-related diseases. Additional validation of these biomarkers was conducted utilizing the patients' clinical records to elaborate on the associations between secondary diseases and their possible effect on CVD prognosis. Upregulation in *RN7SL2* can lead to ischemic stroke[42] and an increase in *LI*LRA2 expression can lead to coronary atherosclerosis heart disease (CAD) due to suppression of the immune response contributing to chronic inflammation, a hallmark sign of CAD[43]. *GAS5* regulates the proliferation, cell cycle and proliferation of myocardial infarction (MI) cells and its overexpression can lead to increased susceptibility to MI[44]. *TWF2* is strongly expressed in cardiac muscles and binds actin which contributes to the morphology of cardiomyocytes[45]. Additionally, the overexpression of *EGLN2* can lead to erythrocytosis; however, the mechanism by which it impacts the pathways is still unknown[46]. *SNHG6* can aggravate hypoxia/reoxygenation induced cardiomyocytes[47–49], while another significant biomarker, *BRK1*, is associated with heart development and its under expression can lead to obstructive heart defects[50]. A significant number of biomarkers were associated to other diseases diagnosed reported for CVD patients' clinical records. We created a network of overlapping diseases linked to the eighteen biomarkers in the highly diagnosed conditions from EHRs (Electronic Health Records) as well as those reported earlier in our comparative review (Fig. 5). We observed that most genes were interconnected through a CVD including but not limited to cardiomyopathy, stroke, and atherosclerosis. The most common non-CVD diagnosis within our patient cohort was breast cancer, and we found *GAS5*[51], *TSTD1*[52], *EGLN2*[53], *SNHG6*[54], *BRK1*[55], and *MTRNR2L1*[56] to be indicative biomarkers. As stated earlier, cardiomyopathy was the next prevalent disease in our network corroborating our claims that our innovative AI/ML model can accurately predict CVDs. Other diseases that were shared between the genes included coronary artery disease, myocardial infarctions, lung cancer, and type 1 diabetes among others (Fig. 5 and supplementary material 4).

In this study, we analyzed the complete transcriptome of patients based on the RNA-seq drive gene expression values allowing for an unbiased exploration of gene expression patterns, uncovering unexpected gene associations and novel biomarkers that might have been missed with a more targeted approach. While small sample sizes can prevent generalizability, statistical significance (p value) should be considered when interpreting a study's results[57]. Recent AI/ML analyses have focused on utilizing high-quality datasets as input for their predictive models[58,59]. A previous study comparing various ML algorithms for the identification of high-risk genes in colon cancer utilized transcriptomic, age and gender data from a cohort of 62 individuals (40 patients and 22 healthy controls)[58]. Similar to our analysis, this study followed a two-level investigation: feature selection for biomarker identification and choosing an optimum ML classifier to accurately stratify patients. Additionally, another novel framework identified gene markers for the precise and targeted treatment of acute myeloid leukemia (AML)[59].



B). Gene-Disease Integration

Diagnosis	ICD9	ICD10	Gene					
Type 2 or unspecified diabetes mellitus with peripheral circulatory disorder [Type 2 Diabetes]	250	E11.51	GPX1					
Osteoarthritis [Osteoarthritis]	715	M19.90	GAS5					
History of non-Hodgkins lymphoma [Diffuse Large B-cell Lymphoma]**	V10	Z85.72	HLA-B					
Malignant neoplasm of upper-outer quadrant of right female breast. unspecified estrogen receptor status (CMS/HCC) [Breast Cancer]	174	C50.411	MTRNR2L1	GAS5	TSTD1	EGLN2	SNHG6	BRK1
Seronegative arthritis [Rheumatoid Arthritis]**	716	M13.80	HLA-DMB	GAS5				
Mass of upper inner quadrant of right breast [Breast Cancer]	611	N63.12	MTRNR2L1	GAS5	TSTD1	EGLN2	SNHG6	BRK1
Coronary artery disease involving native heart with angina pectoris. unspecified vessel or lesion type (CMS/HCC) [Coronary Artery Disease]	414	I25.119	MTRNR2L1					
Special screening for malignant neoplasms. colon [Colorectal Cancer]	V76	Z12.11	EGLN2	SNHG6				
Seronegative rheumatoid arthritis (CMS/HCC) [Rheumatoid Arthritis]	714	M06.00	HLA-DMB	GAS5				
Family history of ovarian cancer [Ovarian Carcinoma]	V16	Z80.41	LILRA2					
Malignant neoplasm of upper lobe. right bronchus or lung (CMS/HCC) [Lung Cancer]	162	C34.11	EGLN2	BRK1	CTA-363E6.6			
Other malignant lymphoma of extranodal or solid organ sites [Diffuse Large B-cell Lymphoma]**	202	C85.89	HLA-B					
Other diabetic neurological complication associated with other specified diabetes mellitus (CMS/HCC) [Type 1 Diabetes]	249	E13.49	HLA-DMB	HLA-DPA1				
NSTEMI (non-ST elevated myocardial infarction) (CMS/HCC) [Myocardial Infarction]	410	I21.4	GAS5					
Obscure cardiomyopathy of Africa (CMS/HCC) [Cardiomyopathy]	425	I42.8	HLA-DMB	HLA-B	GPX1			
Other atherosclerosis of native artery of extremity [Atherosclerosis]	440	I70.299	ARPC4	LILRA2				
Family history of ischemic heart disease [Coronary Artery Disease]	V17	Z82.49	MTRNR2L1					
Wegeners granulomatosis (CMS/HCC) [Granulomatosis with Polyangiitis]	446	M31.30	HLA-DPA1					
Mantle cell lymphoma (CMS/HCC) [Diffuse Large B-cell Lymphoma]**	200	C83.10	HLA-B					
Viral hepatitis [Chronic Hepatitis B Virus]**	070	B19.9	HLA-DPA1					
Hereditary and idiopathic peripheral neuropathy [Neurodevelopmental Disorders]**	356	G60.9	ARPC4					
Malignant neoplasm of posterior wall of bladder (CMS/HCC) [Bladder Cancer]	188	C67.4	ARPC4	GPX1				
Carcinoma in situ of breast [Breast Cancer]	233	D05.90	MTRNR2L1	GAS5	TSTD1	EGLN2	SNHG6	BRK1
Need for prophylactic vaccination and inoculation against viral hepatitis [Chronic Hepatitis B Virus]**	V05	Z23	HLA-DPA1					
Polycystic ovaries [Polycystic Ovary Syndrome]	256	E28.2	GAS5					
Malignant neoplasm of colon (CMS/HCC) [Colorectal Cancer]	153	C18.9	EGLN2	SNHG6				
Telangiectasia [Hereditary Haemorrhagic Telangiectasia]**	448	I78.1	SNHG6					
Malignant neoplasm of prostate (CMS/HCC) [Prostate Cancer]	185	C61	EGLN2					
Interstitial lung disease (CMS/HCC) [Connective Tissue Disease-Associated Interstitial Lung Disease]**	515	J84.9	TSTD1					
Chronic periodontitis [Periodontitis]	523	K05.30	GPX1					
Secondary malignant neoplasm of lung (CMS/HCC) [Lung Cancer]	197	C78.00	EGLN2	BRK1	CTA-363E6.6			
Squamous cell cancer of epiglottis (CMS/HCC) [Oral Squamous Cell Carcinoma]	161	C32.1	GAS5					
Chronic obstructive pulmonary disease. unspecified COPD type (CMS/HCC) [Chronic Obstructive Pulmonary Disease]	496	J44.9	EGLN2					
Chronic myeloid leukemia (CMS/HCC) [Acute Myeloid Leukemia]**	205	C92.10	TWF2					
Ectopic pregnancy without intrauterine pregnancy [Development of Ectopic Pregnancy]	633	O00.90	TSTD1					
Old myocardial infarction [Myocardial Infarction]	412	I25.2	GAS5					

Figure 5. Gene-disease network. This figure presents a gene-disease network including linked ICD-9 and ICD-10 codes.

Gene expression data was collected from 30 AML patients for this analysis and the model was accurately able to organize genes based on their potential to drive cancer[59]. Similarly, our study introduces a novel methodology that has the potential to be extrapolated to larger and more diverse datasets. Additionally, we performed a two-tiered cross-validation on our findings through literature review as well as clinical records collected from patients in our cohort. Our small sample size does not limit the validity of our model as we have employed a nexus of statistical and ML algorithms that aided in managing the restrictions that could emerge from single algorithms. For instance, SVMs play a crucial role in ML systems when the dataset is constrained; however, k-NN provides more accurate predictions on larger cohort sizes[2]. Utilizing these approaches, we have ensured that our model can handle complex and rare disease predictions by accounting for sample size disparities.

We believe that synergistic use of multiple AI algorithms provides more accurate results, draws insightful conclusions, and precise predictions about real-world problems compared to single AI algorithm on its own.

Recently, we published a study in the Briefing in Bioinformatics (Oxford)[2], evaluating and comparing various ML approaches using the gene-variant and expression data for statistical and predictive analysis of a wide variety of disorders. Our study concluded that SVM and RF are the most applied and successful ML algorithms used to make high-accuracy predictions and solve regression and classification problems. The major differences between these two include adjusting hyperparameters (a parameter whose value is used to control the learning process) in SVM to prevent over and underfitting compared to no adjustment in RF[2]. SVM has been implemented to distinguish genetic susceptibility factors and identify previously unknown features that corresponded to common disease[57,60] when RF has been applied to identify differentially expressed genes that played an important role in disease prognosis by acting as a potential biomarker[61–63]. We also established that a multitude of other predictive ML algorithms are employed but less utilized including but not limited to k-NN and XGBoost[2]. Alternative AI/ML approaches exist, however, their adoption for the analysis of multi-genomic data remains limited[2]. Our approach combines the best aspects of multiple machine learning algorithms into a single model. It does not only hold the potential for personalized early detection of common and rare diseases in individuals, but also opens avenues for broader research using novel ML methodologies, ultimately leading to personalized interventions and novel treatment targets. A limitation of our current study is that experimental validation is needed to support the outcomes of our AI/ML model. We addressed this constraint by utilizing clinical records and comparative literature to support our findings. Currently, our methodology only suits binary disease prediction. Prospective multiclass classification tasks require novel methodologies; integrating patient demographics, transcriptomics, variants, and epigenomics can facilitate an unsupervised clustering approach that will allow mapping diseases onto patients through the extraction of these clusters' most important features.

We have proposed a unique combination of classical statistical methods and state-of-the-art ML algorithms to identify novel biomarkers and predict diseases. By integrating these approaches, we outperformed single algorithms, resulting in deeper insights and more precise predictions, essential for personalized early disease-risk detection in individuals[63]. Our AI/ML model can be implemented in the clinical setting to aid in early disease diagnosis and improve prognosis. It has the potential to be generalized to investigate non-CVDs with intricate characteristics such as breast cancer, diabetes, and Alzheimer's disease among many others. To foster these downstream applications, we have made source code openly available and freely accessible. This cutting-edge technology enhances the precision of diagnoses and empowers clinicians to tailor personalized treatment plans, ultimately leading to more effective and targeted healthcare interventions. Our findings validate the effectiveness and reliability of the model in the medical domain, offering promising prospects for improved healthcare outcomes. In the future, we look forward to advancing our methodology by curating an unsupervised learning study that removes the labels to indicate status of health and allows the algorithm to cluster data points based on integrated gene expression and variant data along with clinical, demographics, and longitudinal data.

Data availability

We anticipate that this study will serve as a future resource for the genomics community. The dataset, list of biomarkers, classifier metrics, gene-disease-ICD codes, and exploratory analysis details are attached in the supplementary material.

Code availability

All source code used to compute the results described in the study and generate the figures are available at: https://github.com/drzeeshanahmed/AI_ML_Analysis_Source_Code.

Received: 13 October 2023; Accepted: 21 December 2023
Published online: 02 January 2024

References

1. Ahmed, Z., Mohamed, K., Zeeshan, S. & Dong, X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database* https://doi.org/10.1093/database/baaa010 (2020).
2. Vadapalli, S., Abdelhalim, H., Zeeshan, S. & Ahmed, Z. Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. *Brief. Bioinform.* **23**(5), bbac191. https://doi.org/10.1093/bib/bbac191 (2022).
3. O'Donnell, C. J. & Nabel, E. G. Genomics of cardiovascular disease. *N. Engl. J. Med.* **365**(22), 2098–2109. https://doi.org/10.1056/NEJMra1105239 (2011).
4. Ganesh, S. K. *et al.* Genetics and genomics for the prevention and treatment of cardiovascular disease: update: A scientific statement from the American Heart Association. *Circulation* **128**(25), 2813–2851. https://doi.org/10.1161/01.cir.0000437913.98912.1d (2013).
5. Seo, D., Ginsburg, G. S. & Goldschmidt-Clermont, P. J. Gene expression analysis of cardiovascular diseases: Novel insights into biology and clinical applications. *J. Am. Coll. Cardiol.* **48**(2), 227–235. https://doi.org/10.1016/j.jacc.2006.02.070 (2006).
6. Lee, D. S. *et al.* Association of parental heart failure with risk of heart failure in offspring. *N. Engl. J. Med.* **355**(2), 138–147. https://doi.org/10.1056/NEJMoa052948 (2006).
7. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**(2), 95–108. https://doi.org/10.1038/nrg1521 (2005).
8. Ahmed, Z., Renart, E. G. & Zeeshan, S. Genomics pipelines to investigate susceptibility in whole genome and exome sequenced data for variant discovery, annotation, prediction and genotyping. *PeerJ* **9**, e11724. https://doi.org/10.7717/peerj.11724 (2021).
9. Roger, V. L. *et al.* Heart disease and stroke statistics–2011 update: A report from the American Heart Association. *Circulation* **123**(4), e18–e209. https://doi.org/10.1161/CIR.0b013e3182009701 (2011).
10. Ahmed, Z., Zeeshan, S. & Liang, B. T. RNA-seq driven expression and enrichment analysis to investigate CVD genes with associated phenotypes among high-risk heart failure patients. *Hum. Genomics* **15**(1), 67. https://doi.org/10.1186/s40246-021-00367-8 (2021).
11. Roth, G. A. *et al.* Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **70**(1), 1–25. https://doi.org/10.1016/j.jacc.2017.04.052 (2017).

12. Doran, S. *et al.* Multi-omics approaches for revealing the complexity of cardiovascular disease. *Brief. Bioinform.* **22**(5), bbab061. https://doi.org/10.1093/bib/bbab061 (2021).
13. Krittanawong, C. *et al.* Artificial intelligence and cardiovascular genetics. *Life* **12**(2), 279. https://doi.org/10.3390/life12020279 (2022).
14. Leopold, J. A. & Loscalzo, J. Emerging role of precision medicine in cardiovascular disease. *Circ. Res.* **122**(9), 1302–1315. https://doi.org/10.1161/CIRCRESAHA.117.310782 (2018).
15. Leopold, J. A., Maron, B. A. & Loscalzo, J. The application of big data to cardiovascular disease: Paths to precision medicine. *J. Clin. Investig.* **130**(1), 29–38 (2020).
16. Antman, E. M. & Loscalzo, J. Precision medicine in cardiology. *Nat. Rev. Cardiol.* **13**(10), 591–602. https://doi.org/10.1038/nrcardio.2016.101 (2016).
17. Baumgart, D. C. & Sandborn, W. J. Crohn's disease. *Lancet* **380**(9853), 1590–1605. https://doi.org/10.1016/S0140-6736(12)60026-9 (2012).
18. Khor, B., Gardet, A. & Xavier, R. J. Genetics and pathogenesis of inflammatory bowel disease. *Nature* **474**(7351), 307–317. https://doi.org/10.1038/nature10209 (2011).
19. Pearce, L. Breast cancer. *Nurs. Stand.* **30**(51), 15. https://doi.org/10.7748/ns.30.51.15.s16 (2016).
20. Cappell, M. S. Pathophysiology, clinical presentation, and management of colon cancer. *Gastroenterol. Clin. N. Am.* **37**(1), 1–v. https://doi.org/10.1016/j.gtc.2007.12.002 (2008).
21. Eratne, D. *et al.* Alzheimer's disease: Clinical update on epidemiology, pathophysiology and diagnosis. *Australas. Psychiatry* **26**(4), 347–357. https://doi.org/10.1177/1039856218762308 (2018).
22. Venkat, V., Abdelhalim, H., DeGroat, W., Zeeshan, S. & Ahmed, Z. Investigating genes associated with heart failure, atrial fibrillation, and other cardiovascular diseases, and predicting disease using machine learning techniques for translational research and precision medicine. *Genomics* **115**(2), 110584. https://doi.org/10.1016/j.ygeno.2023.110584 (2023).
23. Patel, K. K. *et al.* Genomic approaches to identify and investigate genes associated with atrial fibrillation and heart failure susceptibility. *Hum. Genomics* **17**(1), 47. https://doi.org/10.1186/s40246-023-00498-0 (2023).
24. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinform.* **7**, 3. https://doi.org/10.1186/1471-2105-7-3 (2006).
25. Benesty, J., Chen, J., Huang, Y. & Cohen, I. Pearson correlation coefficient. In *Noise Reduction in Speech Processing* 37–40 (Springer, 2009).
26. McHugh, M. L. The chi-square test of independence. *Biochem. Med.* **23**(2), 143–149. https://doi.org/10.11613/bm.2013.018 (2013).
27. Kaufmann, J. & Schering, A. G. Analysis of variance ANOVA. Wiley Encyclopedia of Clinical Trials. https://doi.org/10.1002/9781118445112.stat06938 (2007).
28. Kwak, S. K. & Kim, J. H. Statistical data preparation: Management of missing values and outliers. *Korean J. Anesthesiol.* **70**(4), 407–411. https://doi.org/10.4097/kjae.2017.70.4.407 (2017).
29. Chen, Z., Huang, H. & Ng, H. K. Design and analysis of multiple diseases genome-wide association studies without controls. *Gene* **510**(1), 87–92. https://doi.org/10.1016/j.gene.2012.07.089 (2012).
30. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995).
31. Mucherino, A. *et al.* K-nearest neighbor classification. *Data Min. Agric.* https://doi.org/10.1007/978-0-387-88615-2_4 (2009).
32. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794. https://doi.org/10.1145/2939672.2939785 (2016).
33. Wilczewski, C. M. *et al.* Genotype first: Clinical genomics research through a reverse phenotyping approach. *Am. J. Hum. Genet.* **110**(1), 3–12 (2023).
34. Mhatre, I. *et al.* Functional mutation, splice, distribution, and divergence analysis of impactful genes associated with heart failure and other cardiovascular diseases. *Sci. Rep.* **13**(1), 16769 (2023).
35. Bacchetti, P. Small sample size is not the real problem. *Nat. Rev. Neurosci.* **14**(8), 585 (2013).
36. Tang, L. Informatics for genomics. *Nat. Methods* **17**(1), 23. https://doi.org/10.1038/s41592-019-0709-z (2020).
37. Abdelhalim, H. *et al.* Artificial intelligence, healthcare, clinical genomics, and pharmacogenomics approaches in precision medicine. *Front. Genet.* **13**, 929736. https://doi.org/10.3389/fgene.2022.929736 (2022).
38. Isakov, O., Dotan, I. & Ben-Shachar, S. Machine learning-based gene prioritization identifies novel candidate risk genes for inflammatory bowel disease. *Inflamm. Bowel Dis.* **23**(9), 1516–1523 (2017).
39. Ji, X. *et al.* Single-cell sequencing combined with machine learning reveals the mechanism of interaction between epilepsy and stress cardiomyopathy. *Front. Immunol.* **14**, 1078731. https://doi.org/10.3389/fimmu.2023.1078731 (2023).
40. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18**(1), 76. https://doi.org/10.1186/s13059-017-1207-1 (2017).
41. Lei, C., Niu, X., Wei, J., Zhu, J. & Zhu, Y. Interaction of glutathione peroxidase-1 and selenium in endemic dilated cardiomyopathy. *Clin. Chim. Acta* **399**(1–2), 102–108. https://doi.org/10.1016/j.cca.2008.09.025 (2009).
42. Iwasa, N. *et al.* Gene expression profiles of human cerebral organoids identify PPAR pathway and PKM2 as key markers for oxygen-glucose deprivation and reoxygenation. *Front. Cell. Neurosci.* **15**, 605030. https://doi.org/10.3389/fncel.2021.605030 (2021).
43. Peng, W., Sun, Y. & Zhang, L. Construction of genetic classification model for coronary atherosclerosis heart disease using three machine learning methods. *BMC Cardiovasc. Disord.* **22**(1), 42. https://doi.org/10.1186/s12872-022-02481-4 (2022).
44. Zhang, Y. *et al.* lncRNA GAS5 regulates myocardial infarction by targeting the miR-525-5p/CALM2 axis. *J. Cell. Biochem.* **120**(11), 18678–18688. https://doi.org/10.1002/jcb.29156 (2019).
45. Li, Q. *et al.* Attenuation of microRNA-1 derepresses the cytoskeleton regulatory protein twinfilin-1 to provoke cardiac hypertrophy. *J Cell Sci.* **123**(Pt 14), 2444–2452. https://doi.org/10.1242/jcs.067165 (2010).
46. Camps, C. *et al.* Gene panel sequencing improves the diagnostic work-up of patients with idiopathic erythrocytosis and identifies new mutations. *Haematologica* **101**(11), 1306–1318. https://doi.org/10.3324/haematol.2016.144063 (2016).
47. Lang, Z. *et al.* Silencing of SNHG6 alleviates hypoxia/reoxygenation-induced cardiomyocyte apoptosis by modulating miR-135a-5p/HIF1AN to activate Shh/Gli1 signalling pathway. *J. Pharm. Pharmacol.* **73**(1), 22–31. https://doi.org/10.1093/jpp/rgaa064 (2021).
48. Tørring, P. M. *et al.* Long non-coding RNA expression profiles in hereditary haemorrhagic telangiectasia. *PloS One* **9**(3), e90272. https://doi.org/10.1371/journal.pone.0090272 (2014).
49. Chu, P. M., Yu, C. C., Tsai, K. L. & Hsieh, P. L. Regulation of oxidative stress by long non-coding RNAs in vascular complications of diabetes. *Life* **12**(2), 274. https://doi.org/10.3390/life12020274 (2022).
50. Edwards, J. J. *et al.* Systems analysis implicates WAVE2 complex in the pathogenesis of developmental left-sided obstructive heart defects. *Basic Transl. Sci.* **5**(4), 376–386. https://doi.org/10.1016/j.jacbts.2020.01.012 (2020).
51. Zhao, Z., Chen, C., Liu, Y. & Wu, C. 17β-Estradiol treatment inhibits breast cell proliferation, migration and invasion by decreasing MALAT-1 RNA level. *Biochem. Biophys. Res. Commun.* **445**(2), 388–393. https://doi.org/10.1016/j.bbrc.2014.02.006 (2014).
52. Ansar, M. *et al.* Promoter hypomethylation and overexpression of TSTD1 mediate poor treatment response in breast cancer. *Front. Oncol.* **12**, 1004261. https://doi.org/10.3389/fonc.2022.1004261 (2022).
53. Zheng, X. *et al.* Prolyl hydroxylation by EglN2 destabilizes FOXO3a by blocking its interaction with the USP9x deubiquitinase. *Genes Dev.* **28**(13), 1429–1444. https://doi.org/10.1101/gad.242131.114 (2014).
54. Jafari-Oliayi, A. & Asadi, M. H. SNHG6 is upregulated in primary breast cancers and promotes cell cycle progression in breast cancer-derived cell lines. *Cell. Oncol.* **42**(2), 211–221. https://doi.org/10.1007/s13402-019-00422-6 (2019).

55. Limaye, A. J. *et al.* In silico optimized stapled peptides targeting WASF3 in breast cancer. *ACS Med. Chem. Let.* **13**(4), 570–576. https://doi.org/10.1021/acsmedchemlett.1c00627 (2022).
56. Zhou, K., Arslanturk, S., Craig, D. B., Heath, E. & Draghici, S. Discovery of primary prostate cancer biomarkers using cross cancer learning. *Sci. Rep.* **11**(1), 10433. https://doi.org/10.1038/s41598-021-89789-x (2021).
57. Maniruzzaman, M. *et al.* Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. *Comput. Methods Progr. Biomed.* **176**, 173–193 (2019).
58. Lee, S. I. *et al.* A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat. Commun.* **9**(1), 42 (2018).
59. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Int. J. Complex Syst.* **1695**(5), 1–9 (2006).
60. Kegerreis, B. *et al.* Machine learning approaches to predict lupus disease activity from gene expression data. *Sci. Rep.* **9**(1), 9617 (2019).
61. Zhao, S. *et al.* Identification of diagnostic markers for major depressive disorder using machine learning methods. *Front. Neurosci.* **15**, 645998 (2021).
62. Schaack, D., Weigand, M. A. & Uhle, F. Comparison of machine-learning methodologies for accurate diagnosis of sepsis using microarray gene expression data. *PloS One* **16**(5), e0251800 (2021).
63. Degroat, W. *et al. IntelliGenes*: A novel machine learning pipeline for biomarker discovery and predictive analysis using multi-genomic profiles. *Bioinformatics* **39**, btad755 (2023).

Acknowledgements

We appreciate great support by the Department of Medicine, Rutgers Robert Wood Johnson Medical School (RWJMS); Rutgers Institute for Health, Health Care Policy, and Aging Research (IFH); Rutgers Biomedical and Health Sciences (RBHS), at the Rutgers, The State University of New Jersey. We thank members and collaborators of Ahmed Lab at Rutgers (RWJMS and IFH) for their support, participation, and contribution to this study.

Author contributions

Z.A. designed and led this study. Z.A. participated in sample collection, cohort building, and RNA-seq data generation. Z.A. performed processing, quality checking, and gene-disease data annotation and expression analysis. Z.A. generated AI/ML ready dataset and supported W.D. in designing methodology and implementing AI/ML techniques. W.D., H.A., D.M., and S.Z. supported the pre- and post-computational analysis, evaluation of results and preparation of the supplementary material. H.A. and Z.A. drafted the manuscript. All authors have participated in writing and review and have approved it for publication.

Funding

This study was supported by the Department of Medicine / Cardiovascular Disease and Hypertension, Division of General Internal Medicine, Rutgers Robert Wood Johnson Medical School, and Institute for Health, Health Care Policy and Aging Research which is the part of Rutgers Biomedical and Health Sciences at Rutgers, The State University of New Jersey.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-50600-8.

Correspondence and requests for materials should be addressed to Z.A.