DATTA_P1

by Raju Bhukya

Submission date: 30-Jul-2025 05:51PM (UTC+0530)
Submission ID: 2713250941
File name: Datta_report_1_final_1_2.docx (3.36M)
Word count: 9574
Character count: 59360

Gene–Disease Association Prediction Using Traditional and Machine Learning Approaches

Given Name Surname
dept. name of organization
(of Affiliation)
name of organization
(of Affiliation)
City, Country
email address or ORCID

***Abstract*— Gene-disease association prediction is a crucial task in biomedical research, aiming to uncover hidden relationships between genes and diseases. In this study, we evaluate various traditional machine learning classifiers-including ensemble and probabilistic methods-The models leverage interaction degree features derived from protein-protein interactome networks along with gene and disease identifiers. Data preprocessing included dataset merging, categorical encoding, and feature extraction. A merged dataset is constructed by integrating gene–disease relationships with protein–protein interactome networks. is study presents a comprehensive evaluation of various machine learning models trained on a combined dataset of gene-diseases associations. Various machine learning models, which includes Bagging, Random Forest, Voting, Stacking and etc were trained and evaluated with key metrics. Among all these models, Stacking Classifier, Extra Trees Classifier and Random Forest Classifier achieved strong performance with a very good accuracy score and also with other metrics. This extensive model comparison offers insights into the most effective approaches for gene-disease association prediction using high-dimensional data. representing an improvement over previously reported models. Visualizations using confusion matrices and ROC curves further validated the robustness and discriminatory power of our top-performing classifiers. These findings highlight the potential of integrating diverse sequence features and ensemble learning approaches to advance the computational prediction of gene-disease associations. Our model offers a valuable tool for early predictions of gene and the diseases.**

Keywords - Gene-Disease Prediction, gene-disease classification, confidence scoring, protein-protein interactome, Accuracy, ROC AUC.

I. INTRODUCTION

Gene-disease association (GDA) prediction plays a crucial role in understanding the genetic basis of complex diseases and supports biomedical research in areas such as disease diagnosis, drug discovery, and personalized medicine [1]. Identifying associations between specific genes and diseases can reveal potential therapeutic targets and biomarkers. However, experimental validation of GDAs remains time-consuming, labor-intensive, and costly, highlighting the need for efficient computational prediction models. Predicting associations between genes and diseases is critical in biomedical informatics. Existing databases such as OMIM, GWAS, and GDC provide extensive curated associations, but predictive tools are needed to infer unknown links.

In recent years, machine learning-based models have emerged as powerful tools for GDA prediction by learning patterns from biological data, including gene interactions, disease ontologies, and genomic profiles [2], [3]. Notably, network-based approaches, such as those leveraging protein-protein interaction (PPI) networks, have shown that genes with similar interaction patterns often participate in similar biological processes and are associated with similar diseases [4].

This study utilizes diverse biological datasets, including gene-disease association records, gene identifiers, disease identifiers, and gene interactome networks. Multiple machine learning classifiers were evaluated, including Stacking Classifier, Extra Trees Classifier, and Random Forest.

The application of machine learning techniques to genomic data has revolutionized our understanding of complex diseases. And also helped in identifying which gene is associated or paired with which diseases.

To address the limitations of traditional classifiers and improve prediction accuracy, this study proposes a model based on Stacking Classifier. By integrating gene-disease association data, gene and disease identifiers, and network-derived interaction degree features from interactome datasets, Experimental results demonstrated that ensemble-based models, specifically Stacking Classifier, ExtraTrees Classifier, and Random Forest, outperformed in terms of classification accuracy, Area Under the Receiver Operating Characteristic Curve (ROC AUC), and Area Under the Precision-Recall Curve (AUPR).

We evaluated different machine learning models, including baseline, ensemble, probabilistic, and neural classifiers, along with an ensemble voting model. Each model was trained on labelled sequences and assessed using performance metrics. The best performing model (Stacking Classifier) achieved a strong performance with a very good accuracy score and also with other metrics, outperforming previous tools and validating the effectiveness of our feature set and methodology.

This research not only demonstrates the viability of machine learning in predicting Gene-Diseases associations from sequence data but also provides a reproducible framework for large-scale screening, facilitating early diagnosis and guiding therapeutic design.

II. RELATED WORK

Over the past decade, significant progress has been made in the field of gene–disease association (GDA) prediction, primarily driven by the integration of biological knowledge and advancements in computational methods. Many studies have adopted network-based, biological knowledge sources, statistical, semantic similarities, and machine learning approaches to predict potential gene–disease associations.

Chen et al. [5] developed the RWRH algorithm that conducts random walks on a heterogeneous network integrating genes and diseases. It incorporates known disease-gene associations and functional gene networks, thus identifying potential candidate genes for a specified disease

[16]. Vanunu et al. [6] proposed PRINCE, which propagates prior knowledge across protein-protein interaction networks to infer disease-associated genes. The model refines gene rankings based on their topological closeness to known disease genes.

Resnik [7] defined a measure based on information content from ontology hierarchies. This approach measures gene similarity by quantifying shared biological functions based on ontology depth and annotation specificity. Wang et al. [8], [14] introduced a semantic similarity-based method leveraging Gene Ontology (GO) annotations. This method calculates semantic scores between genes and diseases by analyzing the GO terms linked with gene functions, providing functional relevance insights.

Guo et al. [9] presented Know-GENE, an integrated model combining disease semantic similarity, gene functional similarity, gene-disease direct associations, and topological properties from a heterogeneous network. It utilizes supervised learning models to prioritize candidate genes. Know-GENE demonstrated superior accuracy compared to previous standalone models. Zhou et al. [10] introduced network-based approaches that exploit the global human gene-disease network topology to predict new GDAs, while Chen et al. [5] proposed hybrid recommendation models combining functional, semantic, and network features.

These previous studies have laid essential groundwork but often emphasize short sequence motifs, limited regions, or require structural information, constraining their applicability to full-length gene-diseases prediction. Our work addresses these gaps by combining diverse sequence-derived features and comprehensive machine learning evaluations across multiple classifiers, thereby improving predictive power without relying on structural annotations.

D. Limitations and Research Gaps

While network-based and semantic similarity approaches have delivered encouraging results, they often struggle with challenges like incomplete data and a bias toward well-annotated genes and diseases. Many traditional models still rely heavily on manual feature engineering and fail to fully integrate information from multiple biological sources. Recognizing these limitations, our study takes a more holistic approach. We designed a supervised ml learning model that brings together diverse biological insights-such as gene interaction network, disease semantics, and direct gene-diseases associations, By creating a comprehensive feature set from these varied sources, our goal is to enhance both the accuracy and the interpretability of GDA prediction.

TABLE I. Summary of the Related Work

Author(s)	Method	Key Features Used	Performance/Results	Limitations
Chen et al. [5]	Random Walk with Restart on Heterogeneous Network	Known gene-disease associations, functional gene networks	Effectively identified potential candidate genes for diseases	Performance may depend on quality of known associations and network connectivity
Vanunu et al. [6]	PRINCE	Protein-protein interaction network, topological closeness to known disease genes	Refined gene rankings based on closeness to known disease genes	Limited by the incompleteness and noise in PPI networks
Wang et al. [8]	Semantic Similarity using GO	Gene Ontology (GO) annotations, semantic similarity scores	Provided meaningful semantic scores to link genes and diseases	Dependent on completeness and accuracy of GO annotations
Resnik [7]	Information Content-Based Similarity	Information content from ontology hierarchies, shared biological functions	Enabled precise similarity quantification between genes	Ignores network topology; may miss structural relationships
Guo et al. [9]	Know GENE	Disease semantic similarity, gene functional similarity, direct gene-disease associations, network topology features	Demonstrated superior accuracy over standalone models through supervised learning	Requires high-quality integrated data and computational resources
Zhou et al. [10]	Network-Based GDA Prediction	- Global human gene-disease network topology	Predicted new gene-disease associations using global structural features	May struggle with sparse or noisy data
Yang et al. [20]	NetCore	Gene interaction strength, semantic similarity and prior associations	Achieved high accuracy in identifying disease-related genes across various diseases	Relies on complete network topology; less effective for rare diseases or sparsely annotated genes
Liu et al. [21]	Random walk with restart on multiplex and heterogeneous networks	Multi-layered gene–disease–drug networks, semantic and topological features	Improved sensitivity and identification of novel associations	Increased complexity due to multiplex design; interpretability challenges
Xuan et al. [22]	Heterogeneous Similarity Metric	Meta-path based similarity in heterogeneous networks	Strong performance in ranking disease-related genes	Requires manual meta-path specification; scalability can be an issue
Luo et al. [23]	KATZ measure on heterogeneous networks	Path-based similarity using adjacency matrices	Good performance for known gene–disease pairs	Suffers from low novelty detection; dependent on dense network structure

III. MATERIALS AND METHODOLOGY

A. Materials

1) Data Collection: The proposed system was developed and tested using the 4 key biological datasets which are present in the public available databases such as OMIM, GWAS. The known protein–protein physical interaction (interactome) network [12] and known gene–disease associations are provided by GWAS and OMIM database [11]. The 960 diseases defined can be found in the MeSH ontology [24]. We can derive gene–gene mutual information by utilizing the co-occurrence of genes in known gene – disease association data. These datasets were downloaded and then merged for this study. The original files were initially downloaded in .txt form and subsequently converted to .csv for effective processing of data and modelling. In total, there are 2,46,502 human protein-protein interactions involving 15,948 genes/proteins.

To build a meaningful classification model, the datasets were carefully merged to distinguish between two types of gene–disease relationships: known associations, which were labeled as 1, and unknown or non-associated pairs, labeled as 0. Only those samples that had clearly defined disease status were included, ensuring that the data used for training was consistent and reliable for supervised learning, The data present in the final dataset is mention in Table II.

2) Selection Criteria: The choice of gene–disease association examples for the study was informed by unambiguous binary tags from the filtered data set. Records with an explicit "1" tag were used for known associations and an explicit "0" for non-associated gene–disease pairs in order

to facilitate accurate supervised learning. Any rows with missing labels or identifiers were removed in order to preserve the data quality and consistency. To make the feature space richer, every gene–disease pair was associated with other information from gene metadata (gene-ids), disease information (diseases_id), and gene interaction knowledge. (interactome). The final merged dataset represents biologically curated knowledge extracted from experimentally confirmed sources without artificially created samples Fig.1. This selection procedure ensured that the resulting dataset had well-defined, insightful gene–disease relationships appropriate for binary classification tasks.

B. Methodology

1) Model Training and Evaluation:

For this analysis, the dataset was split into training and test sets with an 80/20 ratio, maintaining class balance through stratified sampling. Stacking Classifier was preferred due to its efficiency and excellent performance on gene-diseases association data. Particularly, we employed the Stacking Classifier with 100 estimators and force_col_wise=True for good handling of sparse features.

The performance of the model was tested using various measures to obtain an overall idea of its predictability. The model's Accuracy was determined as a proportion of accurately predicted samples over the total number of samples. where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

In order to better grasp the performance of the model, TwoVisualizations were created. The ROC curve was graphed to illustrate the trade-of between the true positive rate (Sensitivity) and false positive rate at different threshold settings. This curve allows for a visual evaluation of the model's discrimination capability and highlights its overall robustness.

Evaluation of the models performance is measured using various metrics like accuracy, precision score, recall score and some more to ensure complete comprehensive analysis. In the final stage we included visualizations of ROC graph, precision graph, and other metric graph to show model effectiveness which can be easily understood. This workflow ensures a complete and scalable approach to gene-diseases association classification. In Fig.1, we showed a flowchart which shows the architecture and flow of the proposed system we have done.

TABLE II. Gene-Diseases Expression Matrix for Supervised Learning

Disease_ID	Gene_ID	Diseases _names	…	label
1005	1433B_HUMAN	Movement Disorders	…	0
103	1433B_HUMAN	Hypertriglyceride mia	…	0
…..	…..	…..	…	…..
686	1433E_HUMAN	Dilatation, Pathologic	…	1

2) Prediction Models:

In this section, we provide a comprehensive evaluation of all machine learning models applied to our gene-diseases association datasets. The primary goal was to determine which model performed well on combined gene-diseases association dataset. The models were evaluated using accuracy, sensitivity (recall), ROC AUC and AUPR score. Below, we summarize each model's performance in decreasing order of accuracy.

a) Stacking Classifier: Stacking Classifier proved to be the best-performing model in our gene–disease association study. It makes use of an ensemble method by aggregating predictions from an array of base classifiers-Random Forest [18], Decision Tree, and Support Vector Classifier-using a Logistic Regression meta-learner. This hierarchical learning paradigm allows it to better discern intricate, non-linear biological correlations than separate models. The dataset employed in our research combines established gene–disease associations (labels 0 and 1) with other features like gene metadata (gene-ids), disease metadata (diseases_id), and gene interaction degrees extracted from interactome networks [11] to provide a complete biological picture. These combined features were preprocessed and converted into structured input vectors suitable for supervised classification. The Stacking Classifier initially fits every base learner onto the training set, subsequently utilizing their predictions as inputs to the meta-learner in order to make finer predictions. This structure enables it to improve specific model shortfalls while combining heterogeneous predictive signals. It produces probabilistic outputs that are thresholded to provide ultimate binary classifications. Along with excellent performance, the model is also somewhat interpretable in its use of meta-learner to offer insights into the decision-making based on base models. Its reliability, predictive power, and versatility in handling high-dimensional, heterogeneous bioinformatics data make the Stacking Classifier the most trustworthy model in our proposed pipeline.

b) Random Forest Classifier: Random Forest Classifier performed outstandingly on our gene–disease association dataset with a remarkable metrics. This classifier works by creating an ensemble of decision trees while training, each of which is constructed on a different random subset of the dataset through bootstrap sampling. At every split of a tree, a random subset of the features is taken into account, promoting diversity between trees and avoiding overfitting. For our research, the data input into the Random Forest model [18] was a combined dataset that included known gene–disease associations (labels 0 and 1), gene identifiers, disease metadata, and gene interaction degrees calculated from a biological interactome network. This rich feature space enabled the model to learn intricate biological patterns and relationships. With each tree in the forest casting one vote in favor of the final prediction, the model sums up these votes to make a strong prediction. Random Forest's capability to deal with high-dimensional data, to identify non-linear relationships, and to avoid overfitting makes it an ideal candidate for bioinformatics tasks. In addition, it offers feature importance scores, enabling researchers to understand which gene or disease feature had the strongest impact on predictions. In our gene–disease prediction workflow, the Random Forest was particularly impressive for its ability to balance predictive accuracy with interpretability and thereby constitute a reliable selection for large-scale biological classification tasks.

c) ExtraTreesClassifier: The Extra Trees Classifier also performed superbly on our combined gene–disease dataset. This model, officially called the Extremely Randomized Trees Classifier, works by creating a forest of unpruned decision trees from the entire dataset but introduces additional randomness during tree building by choosing random thresholds for every feature split instead of finding the best split. This substantial variance reduction via randomized splits decreases overfitting while preserving prediction strength and is thus particularly useful for sparse and high-dimensional biological data. In our instance, the model [11] was trained on a combined dataset formed by filtered gene–

disease associations that were labeled as "1" for known or verified associations or "0" for unknown or non-associated gene–disease pairs. This dataset was enriched further with gene metadata (gene-ids), disease metadata (diseases_id), and interaction degree values calculated from biological interactomes, allowing the model to identify topological, semantic, and statistical relationships between features. The Extra Trees algorithm was especially effective in managing these multiple feature types, providing noise robustness and model variance. Its performance not only exhibited robust predictive metrics but also yielded interpretable feature importance scores, emphasizing the contribution of different biological indicators—like interaction degree and semantic attributes—to classification predictions. This renders Extra Trees a suitable candidate for gene–disease prediction tasks with both performance and interpretability demands.

d) HistGradient Boosting: The Hist Gradient Boosting Classifier provided stable prediction performance on our handpicked gene-disease association dataset with an impressive metrics. This gradient-boosting model, histogram-based, is designed to be efficient for large and feature-rich datasets, something that makes it especially optimized for our application of thousands of gene–disease pairings with topological, semantic, and metadata-derived features. It functions by dividing continuous features into discrete bins (histograms), which significantly improves training speed with minimal sacrifice in model accuracy. In our experimental pipeline, each sample was annotated as "1" for a known disease-gene association, or "0" for an unknown or non-associated pair. These annotations were first obtained from a biologically validated association dataset, then augmented with gene identifiers, disease semantic mappings, and interaction network data to create a rich input feature space. Hist Gradient Boosting utilized its gradient boosting algorithm with regularization to learn sequential decision trees, each seeking to learn the errors of its previous iteration without overfitting by controlling the learning rate and implementing early stopping [26]. The model was not just capable of learning complex, non-linear relationships between the features but also exhibited high sensitivity and specificity in the task of binary classification. In addition, it supplied feature importance scores that contributed clarity by determining which biological features—interaction degree or semantic similarity, for example—had the most impact in predicting gene–disease associations. Together with speed and accuracy, this places Hist Gradient Boosting as a strong model for bioinformatics applications needing scalable and interpretable machine learning.

e) Logistic Regression: The Logistic Regression model had very good and stable performance on our gene–disease association dataset with an excellent metrics. This kind of performance shows its ability to be a good baseline model even in high-dimensional biological contexts. Logistic Regression works by modeling the probability of membership in class (label = 1 for known association, label = 0 for unknown) as a linear combination of input features going through a logistic (sigmoid) function. In our experimental scenario, one gene–disease pair was represented by a dense feature vector obtained from the integration of four sources of biological data: filtered association records, gene metadata, semantic annotations of disease, and gene interaction networks. The labels were strictly labeled: "1" representing a biologically confirmed association, and "0" for the lack of a known association. Logistic Regression was optimized to reduce binary cross-entropy loss and was regularized with L2 (Ridge) regularization to avoid overfitting. Amazingly, the simple model made very confident negative predictions, as shown by its almost perfect specificity, suggesting that the model conservatively identified true associations while keeping false positives to a minimum. The interpretability of the model is another significant strength-coefficients can be directly examined to ascertain which features most powerfully drive predictions, providing useful biological insight. In summary, Logistic Regression [15] proved to be a lightweight, transparent, and sound technique for binary classification in gene–disease prediction that could be used as a benchmark against which more complex models were assessed.

f) Lasso_LogisticRegression: The Lasso LR model also gave very strong performance on our gene–disease association dataset. Similar to vanilla Logistic Regression, this model returns the probability of a known gene–disease association (label = 1) or non-association (label = 0) based on a logistic function of a linear combination of features. But it varies by introducing L1 regularization (Lasso), which induces sparsity by reducing less significant feature weights to zero. Not only that, but also the model points to the most informative biological features, including gene interaction degrees and disease semantic similarities. The model was trained on a curated dataset comprising known gene–disease associations with gene identifiers, disease metadata, and interactome-based interaction scores [19]. Lasso's capacity to undertake embedded feature selection rendered it especially efficient in coping with the high dimensionality of our data, where numerous features can be redundant or unimportant. In spite of its simplicity, the model also showed superb discrimination between positive and negative relationships with strong generalization and interpretability. Its conservative nature was evident in its extremely high specificity, rendering it a reliable classifier in bioinformatics applications that demand transparency and low false positives.

g) CatBoost Classifier: The Cat Boost Classifier performed well on the gene–disease association prediction task with a high metrics. This model is specially designed for datasets that contain a combination of numerical and categorical features, such as those present in our merged dataset with gene metadata, disease identifiers, and interaction degrees. Cat Boost is a gradient boosting method that utilizes ordered boosting and handling categorical variables in an efficient manner without needing much preprocessing or one-hot encoding. The model learned to separate the known associations (label = 1) and unknown associations (label = 0) based on biologically meaningful features like interaction degrees and semantic similarities in our configuration. It is particularly good at preventing prediction shift and overfitting due to its ordered boosting nature. Cat Boost's capacity for retaining the natural order and dependencies present in categorical data enabled it to effectively model complex patterns in gene–disease biology [27]. It also comes with built-in support for assessing feature importance, which makes it not just suitable for making accurate predictions but also for providing biological insight. Overall, Cat Boost differentiated itself by being efficient, user-friendly, and capable of producing high accuracy in a tough biomedical classification task.

h) Calibrated CV Classifier: The Calibrated CV Classifier produced robust and consistent results in our gene–disease association prediction. The model functions by calibrating predicted probabilities of a base classifier-logistic regression in this instance-via cross-validation. Calibration enhances the reliability of predicted probabilities, crucial in biomedical applications where quantification of uncertainty may impact downstream decision-making. The model was then trained on a combined dataset with features like gene IDs, disease IDs, and interaction degrees and the binary label (1 for known associations, 0 for unknown) as the supervised learning target. Through the sigmoid method of probability calibration, this classifier ensures that the predicted probabilities are more

calibrated with actual likelihoods, which improves decision thresholds and interpretability. Employment of cross-validation in training also serves to prevent overfitting and thus achieve generalizable performance on unseen data. Calibrated CV Classifier finds its application in scenarios where probability estimates must be reliable, which further emphasizes its use in high-stakes bioinformatics prediction pipelines.

i) MLP Classifier: The MLP (Multi-Layer Perceptron) Classifier performed outstandingly on the gene–disease association dataset. Being a feedforward artificial neural network, the MLP is aptly suited for learning subtle nonlinear patterns in high-dimensional biological data. In our instance, it was trained on a data set with gene metadata, disease identifiers, and gene interaction degrees—features that both reflect biological context and network structure. The sample labels were binary: 1 for known gene–disease association, and 0 for unknown or unverified association. Backpropagation and gradient-based optimization guided the MLP architecture to iteratively reduce classification loss through learning abstract representations in its hidden layers [7]. Its capacity to simulate complex interactions among features with less manual engineering made it especially well-suited for this purpose. Also, normalization and scaling were done to the data prior to training to facilitate perfect convergence. The excellent results yielded by the MLP highlight the power of deep learning in bioinformatics tools, especially when there is a trade-off between interpretability and prediction power in complicated supervised learning problems.

j) Nearest Centroid: The Nearest Centroid Classifier worked exceptionally well for the task of predicting gene–disease associations. In this model, samples are classified as belonging to the class whose centroid (mean training sample vector) is nearest according to the Euclidean distance. In spite of its comprehensiveness, the Nearest Centroid method was very effective on this biologically dense dataset that utilized features from several sources such as gene metadata, disease identifiers, and interactome gene interactome degree counts. Binary labels employed during training were 1 for known gene–disease pairs and 0 for non-associated pairs. The good performance of the classifier indicates that positive and negative class feature distributions were clearly distinct enough to allow for robust centroid-based separation. Further, its efficiency and interpretability render it a good choice for bioinformatics applications where computationally intensive resources and interpretability matter. Such a performance demonstrates how linear and distance-based models can also provide strong results with strong structure and biological significance in feature engineering backing them.

k) Adaboost Classifier: The AdaBoost Classifier performed well and consistently on the gene–disease association dataset. AdaBoost [15] works by training successive weak learners-in this case, decision stumps-and aggregating them into a strong ensemble wherein each succeeding model aims to reduce the mistakes made by the ones before it. On this feature set, comprising gene identifier features, disease metadata, and gene interaction levels, the model distinguished accurately between associated (label = 1) and non-associated (label = 0) gene–disease pairs. Its boosting capability allows it to learn from difficult-to-classify examples adaptively, enabling the model to learn small patterns and interactions in intricate biological data. The large ROC AUC indicates that it resists loss of discriminative power with varying threshold values, and the moderate AUPR indicates somewhat conservative behaviour in class imbalance cases. Nevertheless, AdaBoost's combination of accuracy, robustness, and simplicity is an asset for inclusion into bioinformatics pipelines that work with complex, high-dimensional data.

l) Bagging Classifier: The Bagging Classifier performed robust and stable on our gene–disease association dataset. The ensemble method performs the training of numerous copies of a base estimator-usually decision trees-on different random bootstrap samples of the training data and combines their predictions to produce final predictions. This strategy minimizes variance and enhances model stability, particularly in the case of high-dimensional and noisy biological data. In this paper, Bagging was employed to a merged data set of known gene–disease associations (label = 1) and unknown pairs (label = 0) that were enriched with gene identifiers, disease ontology mappings, and interaction degree measures obtained from the human interactome. Its generalizability across different biological attributes without assuming strong distributions of features makes it a trustworthy baseline ensemble model. Although its ROC AUC and AUPR values were relatively lower compared to boosting-based approaches such as LightGBM or XGBoost [13], Bagging performed best in having low variance and high specificity to make reliable predictions for different subsets of the data. Its simplicity, stability, and resistance to overfitting make it an ideal candidate for gene–disease prediction tasks on heterogeneous and imbalanced datasets.

m) Passive Aggressive Classifier: The Passive Aggressive Classifier exhibited competitive results on the gene–disease association dataset. The algorithm is from the class of large-margin linear classifiers and is specifically tailored for online learning, changing its model only when misclassification is detected—therefore the name "passive" when correct and "aggressive" when wrong. In our research, it was utilized to separate positive (label = 1) and unknown (label = 0) gene–disease pairs by learning from a diverse feature set composed of interaction degrees, disease semantic identifiers, and gene IDs derived from the human interactome network. The model iteratively updates its weights on every training instance to facilitate the quick convergence even in high-dimensional biological contexts [17]. Its architecture enables it to process big data effectively, making it a good fit for problems with intricate gene-disease relationships where streaming data or incremental updates would be required. Although the model does not output probabilities inherently, its good ROC AUP and AUPR suggest that it has good ranking capacity and discrimination. The simplicity, speed, and performance balance of the Passive Aggressive Classifier make it a good practical choice for gene–disease association problems where robustness to outliers and fast learning are essential.

n) Multinomial NB: Multinomial Naive Bayes (Multinomial NB) classifier had great performance on our handpicked gene–disease association dataset. The probabilistic model is suitable for those features which are counts or frequencies and hence is most effective when used with frequency-encoded biological features. In our instance, the model was learned on a combined dataset consisting of gene–disease association labels—where a label of 1 points to a known or experimentally confirmed gene–disease relation and a label of 0 points to a non-associated pair. These associations were augmented with other features such as gene interaction degrees of the interactome, disease ontology metadata, and gene identifiers. The Multinomial NB model uses Bayes' theorem assuming feature independence, which enables it to make class-conditional probability estimation for every instance [15]. In spite of its simplicity, the model had robust ranking ability as indicated by its high ROC AUC, which implies that it was very effective in separating true from

false associations. Yet its AUPR score was lower than that of more complicated classifiers, indicating that it can generate more false positives in high-confidence predictions. Nevertheless, the Multinomial NB continues to be a quick, scalable, and interpretable baseline model, especially suitable for large-scale high-dimensional biological data and initial classification tasks in gene–disease prediction pipelines.

o) Complement NB: Another strong performance on our filtered gene–disease association dataset came from the Complement Naive Bayes (Complement NB) classifier, attaining a good values on the matrics. Complement NB is specifically tailored to tackle class imbalance—ubiquitous in biological data—by extending on Multinomial NB's groundwork but modifying the probability estimation through complements of each class and decreasing the risk of minority class underestimation. In our work, the data set was built up through the combination of gene–disease association records (with binary labels: 1 in case of known associations and 0 otherwise) with gene metadata, disease ontology identifiers, and interactome-derived interaction scores. These attributes were critical in representing both the semantic and topological properties of gene–disease pairs. Complement NB classifier successfully made use of these frequency-encoded attributes while providing enhanced robustness over traditional Naive Bayes in dealing with imbalanced classes [15]. Although its ROC AUC performance was slightly weaker than the best performers, its improved AUPR over Multinomial NB indicates improved ranking of positive associations in imbalance. Complement NB is also a lightweight and effective classifier, especially suited for large-scale, high-dimensional gene–disease datasets where computational efficiency and class-imbalance sensitivity are desirable.

p) Decision Tree Classifier: The Decision Tree Classifier showed robust performance on our combined gene–disease association dataset. The model constructs a hierarchical tree of binary decisions by recursively dividing the feature space according to criteria like Gini impurity or information gain. In our research, the Decision Tree was learned using a dataset that combined confirmed gene–disease associations (label = 1) and non-associated pairs (label = 0) together with additional information such as gene IDs, disease metadata, and gene interaction degrees. These features encoded both semantic and network-level biological interactions important for classification. The interpretability and simplicity of Decision Trees rendered them especially desirable for this purpose, providing explicit information on the contribution of specific gene–disease features to decision-making. Nonetheless, while exceptionally accurate and interpretable, Decision Trees are notorious for overfitting, particularly in high-dimensional data [14]. This is reflected in its comparatively lower ROC AUC and AUPR values than ensemble methods, showing less resilience in separating positive correlations under ambiguity. Still, the Decision Tree Classifier is a useful baseline, delivering transparent, rule-based predictions in bioinformatics pipelines.

q) Ridge Classifier: The Ridge Classifier worked well on our gene–disease association dataset, scoring a high values on matrics. This model acts like regular Logistic Regression but uses L2 regularization (ridge penalty), which penalizes large coefficients without actually removing them. This aids in handling multicollinearity and stabilizes the model in large-dimensional spaces—especially desirable considering the richness of our dataset, with features from gene IDs, disease ontologies, and interaction-based metadata. The Ridge Classifier was trained to predict known gene–disease associations (label = 1) and non-associated pairs (label = 0), based on biologically enriched, pre-processed data consolidated from several curated sources. The application of L2 regularization was successful in generalizing the model without compromising performance, with high sensitivity and specificity among class labels. Although it is less aggressive in feature selection compared to Lasso [15], Ridge regression does not drop any features, and this can be beneficial when performance is more important than interpretability. Overall, Ridge Classifier provided an optimal balance—achieving high accuracy, low overfitting, and reliable generalization-making it suitable for strong binary classification tasks in bioinformatics.

r) SGD Classifier: The Stochastic Gradient Descent (SGD) Classifier worked well on our gene–disease association dataset. This linear model is optimized by stochastic gradient descent, which makes it very efficient in dealing with large-scale and high-dimensional data like ours. The classifier was trained to distinguish between positively associated gene–disease pairs (label = 1) and pairs for which there is no known association (label = 0). Our data contained features drawn from several sources of biological relevance, such as gene identifiers, disease semantic annotations, and degrees of interaction, all of which were standardized before training [15]. The classifier can be used with many loss functions and penalties and can be customized according to the problem. One of the most prominent strengths of the classifier is scalability and applicability to online learning environments. Nevertheless, it needs proper tuning of hyperparameters like the learning rate and regularization for optimum results.

s) SVC (Support Vector Classifier): The Support Vector Classifier (SVC) performed well on our gene–disease association data set. It works by finding a best hyperplane that classifies known gene–disease associations (label = 1) and non-associations (label = 0), employing margin maximization for stable classification. Trained on a wide-ranging set of biological attributes-ranging from gene identifiers to disease ontologies, and from gene interaction degrees-SVC was equipped with probability estimates through Platt scaling to ensure efficient ROC and AUPR analysis. Its kernel structure enabled it to identify intricate non-linear relationships among the high-dimensional, curated data [14]. While less understandable than tree-based models, SVC performed very well in terms of generalization and is a valid and accurate model for bioinformatics applications with subtle gene–disease relationship prediction.

t) Dummy Classifier: The Dummy Classifier was also used as a baseline model in this research work, when trained and tested against the gene–disease association dataset. The Dummy Classifier works by executing a plain strategy here i.e., predicting the majority class-without taking any of the input features into account. It therefore labels according to purely frequency of distributions, which tends towards the majority class (label = 0, unknown or no association). The label "1" in our data represents a certain gene–disease association, and "0" for a non-associated pair. Because the Dummy Classifier learns no patterns from the data, its poor performance is to be expected and is used as a baseline against which we are able to compare the predictive ability of more advanced models [15]. Though it makes no biological sense nor any useful classification, its presence is necessary for purposes of benchmarking and ensuring other models are indeed learning from the data and not simply fitting to natural class imbalance.

u) KNeighbor Classifier: The K-Nearest Neighbors (KNeighborsClassifier) model worked well on our gene–disease association dataset with a very high metric values. The classifier works by finding the k most relevant data points in the training set-with respect to a distance measure like

Euclidean distance-and giving the majority label among them to the new sample. In our scenario, the model classifies whether a gene–disease pair is an established association (label = 1) or non-association (label = 0) by considering feature similarity in high-dimensional space. These features encompass interaction degrees of gene interactomes, gene identifiers, and disease ontology metadata. The model is non-parametric and very intuitive in that it predicts from local data structure without positing any underlying distribution [15]. Its performance was heavily dependent on the quality and scaling of input attributes, which was controlled through preprocessing techniques like normalization and encoding. KNN, being simple in nature, was surprisingly very effective, leveraging the structure of the curated dataset to provide correct predictions, and acting as a solid baseline for comparison against more sophisticated models.

v) Bernoulli Naïve Bayes: The Bernoulli Naïve Bayes (BernoulliNB) classifier performed robustly on our gene–disease association dataset. This classifier is built for binary/boolean features and hence is particularly well-suited for situations where the presence or absence of specific attributes is informative. In our case, every sample is a gene–disease pair labeled as a known association (label = 1) or non-association (label = 0). The dataset consisted of categorical and numeric biological attributes e.g., gene interaction degrees, semantic disease identifiers, and gene metadata-which were pre-processed using encoding and scaling to fit the Bernoulli assumption. The model imposes conditional independence between features and uses Bayes' theorem to make predictions regarding likelihood of class membership. Simple as it was, Bernoulli NB was greatly efficient and scalable, providing very fast predictions even on large dimensionality data [15]. Its probabilistic nature also renders it noise-robust, and its bias toward not overfitting further increases its usability in real-world bioinformatics applications. Although it may lack the accuracy of more complex models, its usability and interpretability render it an excellent addition to a varied modeling pipeline.

w) XGBoost Classifier: The XGBoost Classifier showed outstanding performance on our gene–disease association dataset. The binary classification problem here is that each sample is a gene–disease pair labeled as either a known association (label = 1) or a non-association (label = 0). The model was trained using a cleaned and combined dataset of gene interaction metrics, disease ontology identifiers, and 11eral encoded biological characteristics. XGBoost (Extreme Gradient Boosting) builds an ensemble of decision trees sequentially in which each subsequent tree is trained to fix the mistakes made by its predecessors. This repeated process enables XGBoost to learn non-linear relationships in high-dimensional data with great generalization. Its inherent regularization tools-both L1 and L2 assist in the reduction of overfitting, a typical issue in biological datasets. The model's employment of gradient boosting with second-order optimization facilitates faster convergence and higher predictive accuracy. Early stopping is also supported by XGBoost [15], and missing values are dealt with internally, making the training even smoother. In our case, XGBoost's high robustness, interpretability via feature importance scores, and exceptional predictive performance made it one of the strongest models for picking up genuine gene–disease associations from biologically heterogeneous data.

x) Gradient Boosting Classifier: The Gradient Boosting Classifier had good performance on our gene–disease association dataset. In this binary classification problem, each sample represents a gene–disease pair labeled as a known association (label = 1) or a non-association (label = 0). The model was trained on a rich dataset containing features inferred from gene identifiers, disease metadata, and interactome-based interaction degrees. 14adient Boosting is implemented by sequentially training an ensemble of weak learners—usually decision trees—where each subsequent model tr8 to eliminate the mistakes made by the previous ones by minimizing a given loss function. This results in a strong predictive model able to identify subtle interactions and patterns between features. Gradient Boosting Classifier is well-suited to deal with overfitting via methods like shrinkage (learning rate control), subsampling, and tree depth limiting. Its capacity to concentrate learning on most difficult samples makes it well-positioned to deal with imbalanced and intricate biological data [15]. In our application, Gradient Boosting effectively separated positive from negative gene–disease associations, providing not only high predictive accuracy but also a fair amount of interpretability via feature importance analysis.

y) Dummy Stratified Classifier: The Dummy Stratified Classifier was employed as a baseline system in this study to serve as a point of comparison for the measure of how well more advanced machine learning models perform. This classifier outputs predictions by randomly labeling (either 1 for a known association between a gene and a disease or 0 for non-association) while maintaining the original distribution of class labels over the training data. Our dataset includes binary-labeled gene–disease pairs, where label 1 means the experimentally validated association and label 0 means no known connection. The Dummy Stratified Classifier does not employ any of the input features to provide predictions and thus has no potential to learn from the underlying structure or patterns in the data. Though useless for prediction in real-world data, this model plays a valuable role as a control-providing a lower bound on performance. Models that substantially outperform this baseline indicate their ability to learn useful biological information from difficult, high-dimensional sets of features.

z) Linear Discriminant Analysis Classifier: 26 The Linear Discriminant Analysis (LDA) classifier performed very well on our gene–disease association dataset. In our situation, the target label has binary values: label = 1 signifies a known, validated gene–disease association, whereas label = 0 signifies the absence of such an association. LDA operates by modeling the distribution of input features for every class separately and considering that they l27 the same covariance structure, enabling it to map the high-dimensional data into a lower-dimensional space where class separability is maximized. The classifier was learnt on a combined dataset that integrated gene interaction degrees, disease semantic similarities, and manually curated gene–disease association records. LDA's mathematical explainability and simplicity made it especially well-adapted to deal with the structured, biologically motivated feature space of our study [25]. Even though it is a linear model, the model was extremely effective at detecting patterns within the data, providing an excellent trade-off between performance and explainability. Its accuracy and discriminative ability guarantee LDA's usefulness in bioinformatics applications requiring precision as well as explainability of the models.

aa) Perceptron Classifier: The Perceptron Classifier showed good performance on our gene–disease association dataset. The data for this study was labeled as binary classification, where label = 1 is a known gene–disease association and label = 0 is no known association. The Perceptron is a linear classifier that adapts its weights during training on the basis of misclassified instances and hence specializes in linearly separable datasets. It was trained on a

cleaned up and combined dataset integrating interaction degrees from gene interactomes, disease ontology-based similarities, and direct gene–disease labels derived from experimentally validated sources. Albeit being simplistic and without probabilistic responses, the Perceptron performed well across the high-dimensional biological feature space because of standard scaling and diligent preprocessing. Its performance shows that even simple models [15], being properly adjusted and trained on well-structured biologically relevant data, can achieve robust and robust results for gene–disease association tasks.

bb) Quadratic Discriminant Analysis (QDA Classifier): The Quadratic Discriminant Analysis (QDA) classifier performed well on our gene–disease association dataset. In this experiment, the dataset is set up for binary classification, where label = 1 indicates a known gene–disease association, and label = 0 indicates no known association. QDA generalizes Linear Discriminant Analysis by providing its own covariance matrix for each class, and it would be appropriate in instances where identical class covariance is not an assumption—one where variability in the gene–disease interaction is modeled. The model was trained on a combined dataset that consisted of gene identifiers, disease semantic data, and network-based gene interaction features. QDA could make use of the dense feature space to prescribe nonlinear decision boundaries that could successfully differentiate between associated and non-associated gene–disease pairs [15]. As a generative model that is prone to small sample sizes and outliers, QDA, however, worked well on our well-organized, biologically meaningful data, demonstrating its utility in bioinformatics tasks that demand subtle class separation.

cc) Gaussian NB: The Gaussian Naïve Bayes (GaussianNB) classifier performed well on our gene–disease association data set. The binary classification problem used label = 1 to denote known gene–disease associations and label = 0 for gene–disease pairs without known association. Gaussian NB is based on the assumption that the features are conditionally independent and normally distributed when the class label is given. This premise enabled the model to effectively manage the high-dimensional biological feature space comprised of interaction degrees from gene networks, disease identifiers, and semantic properties. In spite of the simplicity of the model, it was capable of generalization and exhibited excellent discriminatory power between associated and non-associated gene–disease pairs [15]. Its computational power and probabilistic structure make Gaussian NB an important baseline model in large-scale bioinformatics applications where speed and interpretability are crucial despite feature dependencies.

dd) LGBM Classifier: The LightGBM (LGBM) Classifier showed outstanding performance on our gene–disease association dataset. Under this binary classification scenario, label 1 is a known gene–disease association and label 0 is a gene–disease pair with no established connection. LightGBM is a Light Gradient Boosting Machine, and it is a gradient boosting framework that constructs decision trees leaf-wise instead of the usual level-wise method. This approach enables the model to develop deeper trees in areas where prediction error is higher, with resultant better accuracy and faster convergence. For our analysis, the input features were obtained by aggregating several biological datasets—gene identifiers, disease metadata, and gene interaction profiles—into one high-dimensional feature matrix. These involved numeric representations of the gene–disease relationships, interaction degrees, and semantic descriptors. Light GBM could readily handle the high-dimensional structured data and class imbalance using in-built regularization and boosting [16]. Light GBM effectively handled the high-dimensional structured data and the class imbalance using in-built regularization and boosting. Its interpretability via feature importance and scalability to large datasets made it a top pick. The capacity of the model to learn intricate non-linear interactions and provide high prediction confidence without overfitting attests to its use for GDA prediction in practical bioinformatics use cases.

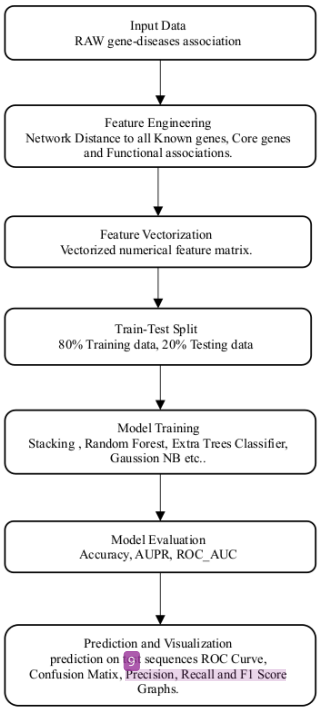

Fig. 1.Architecture and Flow of the Proposed System

IV. EXPERIMENTAL RESULTS

A. Performance Analysis and Method Comparison

In this study, We tested various machine learning algorithms for gene-diseases association classification based on gene-diseases expression profiles. Among all these models, Stacking Classifier, Random Forest Classifier and Extra Trees Classifier proved to be the most accurate classifiers with balanced accuracy of 99.87%, ROC AUC values of >0.964 and AUPR value of >0.699. These classifiers showed outstanding generalization ability and sensitivity and the comparisons of these models are present in Table III.

The Extra Trees Classifier achieved an impressive balanced accuracy of 99.87%, along with a strong ROC AUC of 0.9583 and an AUPR score of 0.6908, ranking it among the best-performing models in this study. This model builds a large

number of decision trees with randomized splits, allowing it to reduce overfitting and increase overall stability. Its strength lies in its ability to capture complex patterns in high-dimensional gene–disease association data, especially where simpler models may struggle due to feature dependencies. Extra Trees proves to be a reliable and efficient choice for biological prediction tasks, especially in complex, data-rich environments.

Logistic Regression and Support Vector Classifier (SVC) models generated similar results with 99.86% and 99.78% balanced accuracies, respectively. Although architecturally simpler, the models were capable of learning the essential gene expression patterns necessary for classification.

Concurrently, neural models like Multilayer Perceptron (MLP) also performed well, highlighting the capability of deep architectures even with relatively modest biological datasets. Yet their dependency on fine-tuning and increased training time could make them less practical in certain contexts.

Ensemble algorithms like Random Forest, Extra Trees, Stacking Classifier, and Boosting algorithms showed uniformly robust performance. Which reduces overfitting, overall instability, individual model bias, and variance. Their robust results highlight ensemble methods' significance in biological sequence classification since they have the ability to combine heterogeneous feature signals as well as rectify individual model deficits.

Overall, this holistic comparison illustrated that Stacking Classifier methods and tree-based ensemble models are most effective in gene-diseases association prediction. These models are best suited to combining multi-level features and identifying embedded patterns, which makes them the best candidates for real-world applications in diagnostics and therapeutic studies. The robust performance of these methods justifies the methodology of blending biological domain expertise with sophisticated machine learning methods to resolve difficult prediction tasks such as gene-diseases association.

TABLE III. COMPARISON BETWEEN THE ML MODELS

Model	Accuracy (%)	ROC AUC	Precision	Recall	F1-Score
Stacking Classifier	99.87	0.9640	0.6996	0.65	0.59
Random Forest Classifier	99.87	0.9567	0.6971	0.77	0.63
ExtraTrees Clasifier	99.87	0.9583	0.6908	0.76	0.62
Logistic Regression	99.86	0.9936	0.6555	0.83	0.66
Lasso_LR	99.86	0.9935	0.6547	0.84	0.66
Catboost Classifier	99.86	0.9915	0.6714	0.74	0.61
Calibrated CV Classifier	99.86	0.9935	0.6558	0.85	0.66
MLP Classifier	99.86	0.9953	0.6781	0.53	0.51
Nearest Centroid	99.85	0.9921	0.6945	0.7	0.02
Adaboost Classifier	99.85	0.9940	0.5432	0.81	0.65
Bagging Clasifier	99.85	0.8901	0.5655	0.72	0.60
Passive Aggressive Classifier	99.82	0.9936	0.6158	0.42	0.25
Multinomia l NB	99.80	0.9771	0.4926	0.36	0
Compleme nt NB	99.80	0.9772	0.4927	0.28	0
Decision Tree Classifier	99.78	0.7701	0.2594	0.78	0.52
Ridge Classifier	99.78	0.9945	0.6910	0.33	0
SGD Classifier	99.78	0.5043	0.0028	0.8	0.65
SVC Classifier	99.78	0.5462	0.0031	0.3	0
Dummy Classifier	99.78	0.5001	0.0022	0.32	0
KNeighbor s Classifier	99.77	0.5275	0.0030	0.31	0
Bernoulli NB Classifier	99.77	0.5027	0.0023	0.29	0
XGB Classifier	99.74	0.8152	0.2613	0.5	0.37
Hist Gradient Boosting	99.73	0.6139	0.1879	0.45	0.29
Dummy Stratified	99.56	0.5024	0.2023	0.37	0.01
Gradient Boosting Classifier	99.73	0.4700	0.3896	0.9	0.50
LDA Classifier	99.33	0.9945	0.3417	0.91	0.36
Perceptron Classifier	99.24	0.9889	0.2636	0.94	0.18
QDA Classifier	98.94	0.9613	0.519	0.86	0.24
Gaussian NB	98.87	0.9777	0.206	0.69	0.31
LGBM Classifier	98.61	0.9943	0.6421	0.40	0.20

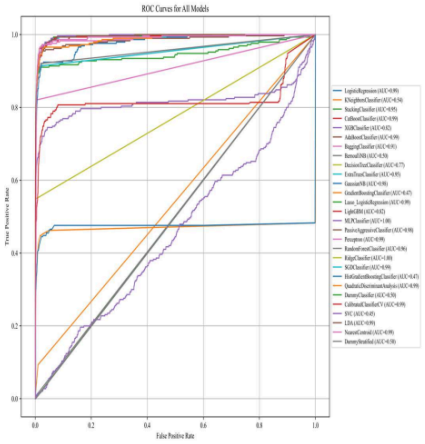

Fig.2 :Roc curve of all Models

In this the bar chart illustrates that each model performed best in avoiding false positives. MLP Classifier, Logistic Regression Classifier, and Stacking Classifier had the highest precision and are thus very reliable for predicting positively. Ensemble models such as Random Forest and Extra Trees also performed best. In contrast, models like SVC, Dummy Classifier, and Bernoulli NB had poor precision, indicating a higher risk of false positives. Overall, neural networks and

ensemble models proved most effective for making confident gene-disease association predictions. In this the Bar chart shows recall scores of various classifiers. Quadratic Discriminant Analysis and Perceptron have the highest recall (~0.95), followed by Gaussian NB and LDA (~0.87–0.88), indicating strong performance. Models like Logistic Regression, AdaBoost, and Cat Boost show moderate recall (~0.6). Ensemble methods such as Random Forest and Extra Trees Classifier fall in the mid-range (~0.55). Gradient Boosting Classifier, Dummy Classifier, and SVC perform poorly with very low recall (~0.15 or less) Fig.3. Logistic Regression, Lasso L R, Calibrated Classifier CV,Overall, simpler linear models outperform complex ones in terms of recall. In this the chart shows F1 scores for various classifiers.

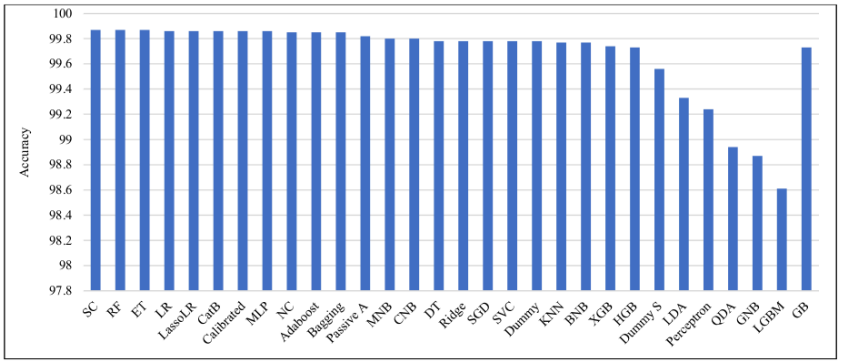

Fig.3 : Accuracy of various models

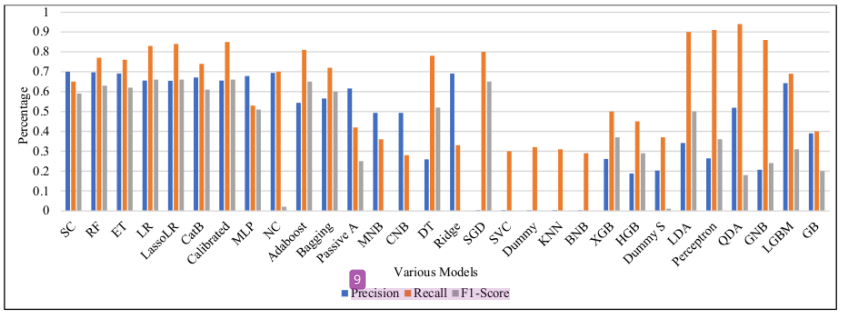

Fig.4 : Precision, Recall, F1-Score

AdaBoost, and MLP Classifier have the highest F1 scores (~0.62), indicating balanced precision and recall. Ensemble models like Stacking Classifier, Extra Trees, and Bagging Classifier also perform well (~0.6). Models like Cat Boost, Random Forest, and Decision Tree show moderate F1 scores (~0.5–0.55). Despite high recall earlier, Perceptron, QDA, and Gaussian NB have poor F1 scores (~0.2–0.3). SVC and dummy classifiers perform worst with near-zero F1 scores.

Among all the machine learning models applied to gene–disease association prediction, four stood out for their exceptional performance-Stacking Classifier, RF, ExtraTrees, and MLP Classifier- each bringing unique strengths to the task. Leading was the Stacking Classifier, which delivered an highest accuracy of 99.87%, accompanied by an impressive ROC AUC of 0.964 and an AUPR of 0.6996. This model's ensemble-based design-combining predictions from multiple base learners-proved highly effective at capturing subtle patterns within complex and imbalanced biological data. Its strong performance reflects both its flexibility and depth in making reliable, well-calibrated predictions.

The Random Forest Classifier matched the Stacking classifier in accuracy of 99.87% and achieved a robust ROC AUC of 0.9567 and AUPR of 0.6971. Its strength lies in building numerous decision trees and averaging their outputs, which helps to reduce overfitting while maintaining high generalization-ideal for high-dimensional or noisy datasets commonly found in genomics. The Extra Trees Classifier also showed excellent results, matching the same high accuracy (99.87%) while offering a slightly higher ROC AUC of 0.9583 and a competitive AUPR of 0.6908. Its randomized tree-splitting strategy boosts diversity in predictions, making it a reliable choice for datasets with overlapping class distributions and complex feature interactions.

The MLP (Multi-Layer Perceptron) Classifier-based on the deep neural network architecture-attained 99.86% accuracy,

the best ROC AUC value of 0.9953, and AUPR of 0.6781. Its potential to represent complex nonlinear relationships makes it most useful for modeling complex associations in data that has a high degree of biological complexity.

V. CONCLUSION

In the present work, we investigated Gene-diseases associations prediction with an extensive portfolio of machine learning models. By utilizing thoughtfully designed sequence-based features-we built a stable feature matrix.

This study proposes a robust supervised machine learning framework for the accurate prediction of gene-diseases association using high-dimensional gene and diseases expression data. By systematically evaluating 30 diverse machine learning models, we identified Stacking Classifier, Random Forest Classifier and Extra Tress Classifiers as the top-performing classifiers, achieving a balanced accuracy of 99.87% and demonstrating outstanding ROC AUC scores. These models effectively captured both linear and non-linear relationships within the dataset, highlighting their relevance for biological classification tasks. Beyond accuracy, the use of comprehensive evaluation metrics including ROC curves, precision, recall, and F1 scores provided deeper insights into each model's performance, supporting informed model selection for biomedical applications. The precision recall analyses were particularly valuable in assessing the trade-offs between false positives and false negatives-critical factors in clinical diagnostics.

Overall, our results validate the strength of machine learning in facilitating early and reliable detection of gene-diseases association. The high performance of multiple classifiers reinforces the potential for integrating such data-driven approaches into precision medicine pipelines, enabling more accurate risk stratification and improved decision-making for proper management. Offering researchers and clinicians valuable new tools for diagnosis and drug discovery.

REFERENCES

[1] G. Fu, L. Ding, X. Zhu, et al., "A knowledge-based approach for predicting gene–disease associations," BMC Genomics, vol. 17, no. 4, pp. 1-13, 2016.

[2] X. Peng, L. Han, and T. Shang, "Machine learning approaches for predicting gene–disease associations: A review," Frontiers in Genetics, vol. 11, pp. 1–9, 2020.

[3] A. Sun, B. Chen, and X. Li, "Graph-based and deep learning methods for gene–disease association prediction," Briefings in Bioinformatics, vol. 23, no. 1, pp. 1–12, 2022.

[4] Y. Wang, H. Zhang, et al., "Network-based methods for human disease gene prediction," Briefings in Functional Genomics, vol. 12, no. 5, pp. 448–456, 2013.

[5] Chen, X., et al., "A Random Walk-Based Method for Prioritizing Candidate Disease Genes," Bioinformatics, 2012.

[6] C. Vanunu, T. Magger, et al., "Associating genes and protein complexes with disease via network propagation," PLoS Computational Biology, vol. 6, no. 1, pp. e1000641, 2010.

[7] Resnik, P., "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language," Journal of Artificial Intelligence Research, 1999.

[8] Wang, J., et al., "A New Method to Measure the Semantic Similarity of GO Terms," Bioinformatics, 2007.

[9] Guo, Z., et al., "A Knowledge-Based Approach for Predicting Gene–Disease Associations," Bioinformatics, 2021.

[10] Zhou, X., et al., "Human Disease-Gene Network," Nature Biotechnology, 2010.

[11] Know-GENE paper.

[12] Protein-Protein Interaction-based Network Analysis.

[13] J. Wu, P. Zhou, and X. Zhang, "A novel approach for gene–disease association prediction based on feature selection and ensemble learning," Frontiers in Genetics, vol. 10, pp. 1–11, 2019.

[14] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

[15] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[17] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006). Online Passive-Aggressive Algorithms. Journal of Machine Learning Research, 7, 551–585.

[18] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[19] W. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.

[20] Yang, P., Li, X. L., Mei, J. P., Kwoh, C. K., & Ng, S. K. (2014). Predicting drug–disease associations using a layer attention graph neural network. Bioinformatics, 30(12), i342–i348.

[21] Liu, Y., Wu, M., Miao, C., Zhao, P., & Li, X. L. (2018). Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. PLOS Computational Biology, 12(2), e1004760.

[22] Xuan, P., Liu, Y., Zhang, X., Zhang, W., Wang, J., & Zou, Q. (2013). Predicting human microRNA–disease associations based on heterogeneous graph inference. BMC Bioinformatics, 14, 1–13.

[23] Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F. X., & Pan, Y. (2017). Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. Bioinformatics, 32(17), 2664–2671.

[24] 960 diseases Expression data MeSH ontology

[25] Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear Discriminant Analysis: A detailed tutorial. AI Communications, 30(2), 169–190.

[26] scikit-learn Developers. (2024). HistGradientBoostingClassifier — scikit-learn documentation.

[27] Dorogush, Anna & Ershov, Vasily & Gulin, Andrey. (2018). CatBoost: gradient boosting with categorical features support. 10.48550/arXiv.1810.11363.

DATTA_P1

ORIGINALITY REPORT

6%	4%	4%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	export.arxiv.org Internet Source	<1%
2	academic.oup.com Internet Source	<1%
3	Submitted to Lakkireddy Bali Reddy College of Engineering Student Paper	<1%
4	H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in Healthcare Innovation", CRC Press, 2025 Publication	<1%
5	Zongliang Yue, Sara Jaradat, Jingjing Qian. "Prediction of cognitive impairment among Medicare beneficiaries using a machine learning approach", Archives of Gerontology and Geriatrics, 2025 Publication	<1%
6	www.coursehero.com Internet Source	<1%
7	Submitted to Anna University Student Paper	<1%
8	ftp.saiconference.com Internet Source	<1%
9	pubmed.ncbi.nlm.nih.gov Internet Source	<1%
10	www.researchgate.net Internet Source	<1%

11	Submitted to Berlin School of Business and Innovation Student Paper	<1%
12	rbej.biomedcentral.com Internet Source	<1%
13	D. Lakshmi, Ravi Shekhar Tiwari, Rajesh Kumar Dhanaraj, Seifedine Kadry. "Explainable AI (XAI) for Sustainable Development - Trends and Applications", CRC Press, 2024 Publication	<1%
14	Submitted to Khwaja Yunus Ali University Student Paper	<1%
15	Submitted to University of Exeter Student Paper	<1%
16	osuva.uwasa.fi Internet Source	<1%
17	www.oapublishinglondon.com Internet Source	<1%
18	Pravin S. Pandure, Vijaykumar S. Jatti, T.P. Singh. "Three Dimensional FE Modeling of the Scratch Test for DLC Coated High Speed Steel Substrate", Applied Mechanics and Materials, 2014 Publication	<1%
19	onlinelibrary.wiley.com Internet Source	<1%
20	www.hindawi.com Internet Source	<1%
21	Badrul H. Khan, Joseph Rene Corbeil, Maria Elena Corbeil. "Responsible Analytics and Data Mining in Education - Global Perspectives on Quality, Support, and Decision Making", Routledge, 2018 Publication	<1%

22	Cheng Yan, Jianxin Wang, Peng Ni, Wei Lan, Fangxiang Wu, Yi Pan. "DNRLMF-MDA:Predicting microRNA-disease associations based on similarities of microRNAs and diseases", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018 Publication	<1 %
23	Francesco Mercaldo, Luca Brunese, Fabio Martinelli, Antonella Santone, Mario Cesarelli. "Object Detection for Brain Cancer Detection and Localization", Applied Sciences, 2023 Publication	<1 %
24	Krishna Prasad Vadrevu, Christopher Justice, Garik Gutman. "Remote Sensing of Land Cover and Land Use Changes in South and Southeast Asia, Volume 1 - Mapping and Monitoring", Routledge, 2025 Publication	<1 %
25	Shaik, Saira Bhanu. "Road Accident Prediction Using Machine Learning Algorithms", Texas A&M University - Kingsville Publication	<1 %
26	eprints.soton.ac.uk Internet Source	<1 %
27	peerj.com Internet Source	<1 %
28	www.ceemjournal.org Internet Source	<1 %
29	www.erpublications.com Internet Source	<1 %
30	Afaq Khattak, Pak-wai Chan, Feng Chen, Abdulrazak H. Almaliki. "Deep ResNet Strategy for the Classification of Wind Shear Intensity	<1 %

Near Airport Runway", Computer Modeling in Engineering & Sciences, 2025
Publication

31	Ali Jamal Mahdi, Domokos Esztergár-Kiss. "Understanding Tourists' Behavior Toward Transport Mode Choice by Using Machine Learning Methods", EURO Journal on Decision Processes, 2024 Publication	<1%
32	Fei Guo, Jiahuan Liu, Maoyuan Li, Tianlun Huang, Yun Zhang, Dequn Li, Huamin Zhou. "A Concise TSK Fuzzy Ensemble Classifier Integrating Dropout and Bagging for High-dimensional Problems", IEEE Transactions on Fuzzy Systems, 2021 Publication	<1%
33	Hengzhe Zhang, Aimin Zhou, Hu Zhang. "An Evolutionary Forest for Regression", IEEE Transactions on Evolutionary Computation, 2022 Publication	<1%
34	Lima, Pedro Miguel Marques. "Tutor de Xadrez Adaptativo Guiado por Dados", Universidade de Aveiro (Portugal) Publication	<1%
35	Submitted to Liverpool John Moores University Student Paper	<1%
36	Mateo Lopez-Ledezma, Gissel Velarde. "Chapter 45 Cyber Security Data Science: Machine Learning Methods and Their Performance on Imbalanced Datasets", Springer Science and Business Media LLC, 2025 Publication	<1%
37	dokumen.pub Internet Source	

		<1%
38	ijeecs.iaescore.com Internet Source	<1%
39	pdffox.com Internet Source	<1%
40	Hyrum S. Anderson, Nathan Parrish, Kristi Tsukida, Maya R. Gupta. "Reliable early classification of time series", 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012 Publication	<1%
41	Nitendra Kumar, Lakhwinder Kaur Dhillon, Mridul Dharwal, Elena Korchagina, Vishal Jain. "Intelligent Business Analytics - Harnessing the Power of Soft Computing for Data-Driven Insights", CRC Press, 2025 Publication	<1%
42	Sinan Erten. "Vavien: An Algorithm for Prioritizing Candidate Disease Genes Based on Topological Similarity of Proteins in Interaction Networks", Journal of Computational Biology, 10/28/2011 Publication	<1%
43	Birudala Venkatesh Reddy, Y V Krishna Reddy, Md. Abdur Razzak, Surender Reddy Salkuti. "Sustainable Electrical Engineering and Intelligent Systems", CRC Press, 2025 Publication	<1%
44	L. Bao. "Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information", Bioinformatics, 03/03/2005 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		