

Integrating WGCNA and Spread Sub-Sampling for Effective Feature Selection in Imbalanced Rheumatoid Arthritis RNA-Seq Classification

Ethan Wright^{1*}, Chloe Bennett², Jack Turner¹, Olivia Harris³

¹ Department of Rheumatology and Bioinformatics, Faculty of Medicine, University of Leeds, Leeds, United Kingdom.

² Department of RNA-Seq Analysis and Feature Selection, Faculty of Medicine, University of Sheffield, Sheffield, United Kingdom.

³ Department of Imbalanced Data and Machine Learning in Autoimmune Diseases, Faculty of Medicine, University of York, York, United Kingdom.

*E-mail ✉ ethan.wright@leeds.ac.uk

Abstract

An imbalanced classification problem arises when the distribution of samples across classes is uneven or skewed. Managing small and imbalanced training datasets remains a significant obstacle in machine learning, particularly in fields like bioinformatics and medical research. Such imbalances can produce biased models that perform poorly on minority classes, over-rely on certain features, and fail to identify the true underlying patterns in the data. This study introduces a feature selection method grounded in gene connectivity, combined with a class-balancing strategy, to construct a machine learning model from imbalanced gene expression data. The approach was evaluated on a rheumatoid arthritis dataset containing 28 normal samples and 152 rheumatic samples. Using weighted gene co-expression network analysis (WGCNA), the original 27,991 features were reduced to 601. These selected features served as input for machine learning classification models trained first on the imbalanced data and then on balanced data generated via the Spread Sub-Sample technique. Results showed that two classifiers achieved higher accuracy on the imbalanced dataset compared to the balanced version, highlighting the tendency of many classifiers to become biased when trained on imbalanced data. Logistic regression attained an improved accuracy of 95%, while the decision tree and IBK algorithms recorded reduced accuracies of 81% and 91%, respectively. In conclusion, combining feature selection with class-balancing techniques is essential to reduce model execution time and improve accuracy, especially when working with RNA-Seq gene expression data.

Keywords: Imbalanced data, Rheumatic arthritis, Machine learning, Spread sub-sample, RNASeq

Introduction

Rheumatoid arthritis (RA) is a complex systemic inflammatory disorder driven by intricate cellular pathways that vary across different disease stages. A defining feature is the presence of anti-citrullinated protein autoantibodies (ACPA) and rheumatoid factor, which characterize the seropositive RA subgroup [1].

This autoreactivity typically emerges during the pre-clinical phase before the onset of chronic joint inflammation. For complex machine learning (ML) algorithms to be widely trusted in decision-making, it is crucial to understand the specific features influencing their predictions. Small, imbalanced training datasets pose a major challenge in machine learning, especially in bioinformatics and medical research. These issues often lead to biased models that underperform on minority classes, overly emphasize particular features, and miss genuine data patterns [2].

Feature selection plays a critical role in RNA-seq data analysis because gene expression profiles are highly dimensional. In RNA-seq studies, thousands of genes are quantified, yet only a subset holds meaningful biological

Access this article online

<https://smerpub.com/>

Received: 06 February 2026; Accepted: 13 May 2026

Copyright CC BY-NC-SA 4.0

How to cite this article: Wright E, Bennett C, Turner J, Harris O. Integrating WGCNA and Spread Sub-Sampling for Effective Feature Selection in Imbalanced Rheumatoid Arthritis RNA-Seq Classification. J Med Sci Interdiscip Res. 2026;6(1):190-9. <https://doi.org/10.51847/ovVk1fBiKM>

relevance. Proper feature selection isolates key genes, improving interpretability, reducing computational demands, and boosting the effectiveness of subsequent analyses [3]. Selecting pertinent genes helps uncover biologically important markers, differentiate between conditions or classes, and clarify underlying biological mechanisms. An imbalanced classification problem emerges when sample counts across classes are disproportionately distributed. This imbalance can range from mild to extreme, with substantial disparities between minority and majority classes [4]. Most machine learning classification algorithms assume balanced class distributions. As a result, imbalanced datasets complicate predictive modeling [5]. Models trained on such data frequently show reduced prediction accuracy, especially for the minority class, which is often the more clinically relevant one. This increases susceptibility to errors in classification tasks. Ultimately, models tend to favor the majority class, producing artificially inflated accuracy scores [6, 7].

Weighted Gene Co-expression Network Analysis (WGCNA) is a widely applied computational method for identifying groups of highly correlated genes through a clustering-based strategy in high-throughput sequencing data analysis. Given a gene expression matrix $G_{n \times m}$, where n represents the number of samples under a specific condition and m denotes the number of transcripts, each element $g(i,j)$ corresponds to the expression level of the j -th transcript within the i -th sample. The WGCNA framework proposed by Langfelder and Horvath [8] constructs a weighted adjacency matrix to quantify the relationships between genes based on their expression similarity.

The type of network generated depends on the adjacency function. In a signed network, correlation values ranging from -1 to 1 are transformed into values between 0 and 1 , allowing the network to preserve information regarding co-regulation patterns, including both gene activation and repression. In contrast, an unsigned network converts negative correlations into positive values, thereby considering only the strength of gene associations while disregarding the direction of regulation. The adjacency value between two genes i and j is calculated as $a_{ij} = |cor(i,j)|^\beta$, where β is the soft-thresholding power parameter. This parameter controls the degree of emphasis placed on strong gene correlations and regulates the smoothness of the transition between weak and strong co-expression relationships within the network [9].

The primary contributions of this work are:

- A combined approach integrating gene connectivity-based feature selection with class balancing techniques to develop a machine learning prediction model for RNA-Seq gene expression networks.
- An effective strategy to mitigate bias arising from imbalanced datasets, where models typically overlook the minority class.

The remainder of the manuscript is structured as follows: Section 2 reviews related work, Section 3 details the proposed approach and methodology, Section 4 presents the results and findings, and Section 5 provides the study's conclusions.

Related work

Numerous studies have tackled class imbalance, with many relying on the Synthetic Minority Oversampling Technique (SMOTE) for data balancing. Sachdeva *et al.* [10] proposed a diagnostic model for hepatitis disease that applied SMOTE to correct class imbalances. They tested four classifiers—SVM, LR, KNN, and RF—finding that SMOTE improved performance across all models. The reported accuracies were SVM 90.43%, LR 93.18%, KNN 85.93%, and RF 92.52%, with LR achieving the best result with SMOTE. Trigka and Dritsas [11] investigated various ML models and assessed the effect of SMOTE. Using metrics including accuracy, precision, recall, and AUC, they determined that a stacking ensemble model with SMOTE and 10-fold cross-validation delivered the strongest performance: 90.9% accuracy, 96.7% precision, 87.6% recall, and 96.1% AUC. Hassan *et al.* [12] compared six machine learning algorithms (LR, DT, NB, KNN, SVM, and RF) on the UCI hepatitis dataset and found that Random Forest attained the highest accuracy of 85%. Gomes *et al.* [13] employed SMOTE with 3 nearest neighbors for class balancing. In the 3-class scenario, they noted substantial feature reductions: 77.6% for LeNet, 82.6% for ResNet, and 80.3% for VGG16. For the 5-class case, reductions were 73.2% for LeNet, 78.6% for ResNet, and 81.1% for VGG16 [12].

Several techniques for feature selection in gene expression data have also been proposed. Ref. [14] introduced a hybrid method that merges mutual information maximization (MIM) and an adaptive genetic algorithm (AGA) to remove redundant features and reduce dimensionality. Their MIMAGA approach achieved classification accuracies above 80% across four

classifiers when compared to other feature selection methods [14]. In 2015, García and Sánchez developed a two-phase model that combined ReliefF-based feature selection with dissimilarity-based representation across eight microarray datasets. They compared artificial neural networks, support vector machines, and Fisher's linear discriminant classifiers in both feature space and dissimilarity space, concluding that dissimilarity-based classifiers outperformed traditional feature-based approaches [15]. Alhenawi *et al.* [16] reviewed 132 studies from three publishers and found that hybrid feature selection methods were the most common research direction (34.9%), followed by others ranging from 13.6% to 3%. Consistent with another survey [17], they noted a growing emphasis on hybrid techniques, which enhance gene selection stability and classification accuracy in cancer studies. In bioinformatics, gene expression classification is central to cancer diagnosis [18]. The authors developed a neighborhood rough set-based feature selection method using neighborhood entropy uncertainty measures. Experiments on ten gene expression datasets showed superior efficiency in selected gene count and classification accuracy, particularly as gene numbers increased [18]. Saberi-Movahed *et al.* [19] proposed a feature selection method integrating matrix factorization and subspace learning with regularization. Tests on nine gene expression datasets confirmed its effectiveness in leveraging feature space information, with suggestions that manifold structures could yield further gains [19].

Most prior research has addressed feature selection and class imbalance independently. Therefore, the current study proposes WGCNA—a network-based method leveraging gene connectivity for feature selection—alongside the Spread Sub-Sample technique for class balancing to build robust machine learning models.

Materials and Methods

The proposed method consists of six main steps, illustrated in **Figure 1**. These include data acquisition and preprocessing, construction of a co-expression network for feature selection, training a classification model on the imbalanced dataset, balancing the data with the Spread Sub-Sample technique, developing a classification model on the balanced data, and finally evaluating the models.

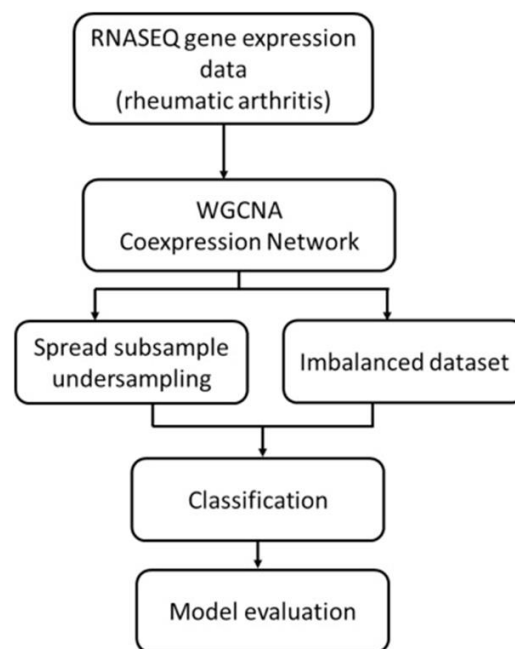


Figure 1. Overview of the methodology employed in the present investigation.

Data type and source

This study used the GSE89408 rheumatoid arthritis dataset, retrieved from the GEO database (<https://www.ncbi.nlm.nih.gov/gds/>). Preprocessing was performed via the back-end GEO RNA-seq experiments processing pipeline (GREP2), following the procedure outlined by Al Mahi *et al.* (2019). The data were divided into two groups: 28 healthy control samples and 152 samples from individuals with rheumatoid arthritis.

WGCNA for feature selection

Weighted Gene Co-expression Network Analysis (WGCNA) was applied to build a scale-free topology network using a soft-thresholding power (β). Several metrics were calculated to evaluate the connectivity among genes. In line with WGCNA recommendations and to meet the scale-free topology criteria, appropriate soft-thresholding powers were chosen to convert the gene expression data into adjacency matrices. Topology overlap matrices (TOM) were then generated from the adjacency matrices. Gene modules were detected through hierarchical clustering combined with the dynamic tree cut approach [20].

Class balancing

One experiment was conducted on the original imbalanced dataset containing 28 normal and 152

rheumatoid arthritis samples, with the outcomes documented. A second experiment used a balanced version of the dataset to assess the impact of class imbalance on machine learning model performance. The Spread Sub-Sample technique—an under-sampling method—was applied to decrease the number of instances in the majority class until balance was achieved. A class frequency setting of 2.1 was used for balancing. Classifier results regarding accuracy, precision, recall, and training time were compared between the balanced and imbalanced conditions.

Classification

Machine learning algorithms

Three classifiers—Logistic Regression, IBK, and J48 decision tree—were trained and assessed on both the balanced and imbalanced datasets.

Logistic Regression is a statistical technique that models the probability of a binary outcome as a function of explanatory variables. It analyzes situations in which independent variables affect a specific result, using a binary dependent variable (True/False, 1/0, or Yes/No). The model formulation is represented in the equations below:

$$x = c_0 + \sum_{i=1}^n c_i x_i \quad (1)$$

$$p(x) = \frac{e^x}{1 + e^x} \quad (2)$$

Where x denotes the level of variables x_i ($i=1, \dots, n$), c_i represents the regression coefficient derived by maximizing the likelihood relative to its standard errors, and $P(x)$ indicates the probability of the event occurring based on the predictor variables.

IBK

The K-Nearest Neighbors (KNN) method represents a supervised machine learning approach grounded in similarity-driven classification. As a non-parametric technique, it is commonly applied in instance-based or lazy learning scenarios, enabling immediate predictions when evaluating new data points. To classify an unseen instance, the algorithm identifies its category by measuring its proximity to the k closest existing examples (k -nearest neighbors), based on a chosen k parameter. To determine these nearest points, multiple distance metrics can be applied, including Euclidean,

Manhattan, Chebyshev, and Levenshtein (also known as Edit Distance). In the present research, Euclidean distance was selected for its proven performance and computational efficiency. The process of this algorithm is outlined in the equation that follows:

$$\text{dist}(X_1, X_2) = \sqrt{\sum_{i=1}^n (X_{1i} - X_{2i})^2} \quad (3)$$

Where $\text{dist}(X_1, X_2)$ is the distance between two samples.

J48

The J48 algorithm determines information gain by evaluating splits on standardized attributes. It builds decision nodes according to class distributions and can accommodate missing values, varying attribute costs, and specific feature characteristics. Pruning may be applied to improve predictive performance. The procedure for J48, as described by Saravanan and Gayathri [21], includes the following steps:

- Step 1: If every instance in the current set belongs to one class, create a leaf node labeled with that class.
- Step 2: Calculate the potential information for each attribute and determine the information gain achieved by testing each attribute.
- Step 3: Choose the attribute that yields the highest information gain based on Step 2. J48 incorporates key decision tree concepts such as entropy, information gain, and information gain ratio, as noted by Gulati *et al.* [22].

Model performance evaluation measures

Classification accuracy

Accuracy indicates the overall correctness of a classifier by showing the proportion of instances correctly assigned to their true classes. It is expressed as a percentage using the formula:

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP} 100\% \quad (4)$$

Precision

Precision quantifies the reliability of positive predictions by measuring the fraction of true positives among all predicted positives. It is computed as:

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

Recall

Recall assesses the completeness of the analysis by considering the total number of true/real positive instances in relation to the total number of false negative instances. Calculated as:

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

F-measure

The F-measure provides a single score that balances precision and recall. Its formula is:

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

Ten-fold cross-validation served as the evaluation approach. In this procedure, the dataset was partitioned into 10 distinct folds. For each iteration, one fold was held out as the test set, while the other nine folds were combined for model training. In the present work, k was fixed at 10. Subsequently, we applied the Kruskal–Wallis H-statistic to assess whether mean ranks differed significantly across groups.

Results and Discussion

Dataset class distribution

The dataset contained 28 normal samples (15.6%), while the remaining 152 samples (84.4%) belonged to the rheumatic arthritis group. This composition reflects a typical class imbalance commonly observed in RNA-seq gene expression datasets.

WGCNA results

Network topology was examined under various soft-thresholding powers. The left panel displays the scale-free topology fit index as a function of soft-thresholding power; power 7 was chosen because the fit index reached a sufficiently high value (> 0.89). The right panel shows

the mean connectivity (expressed in degrees) plotted against the soft-thresholding power (**Figure 2**).

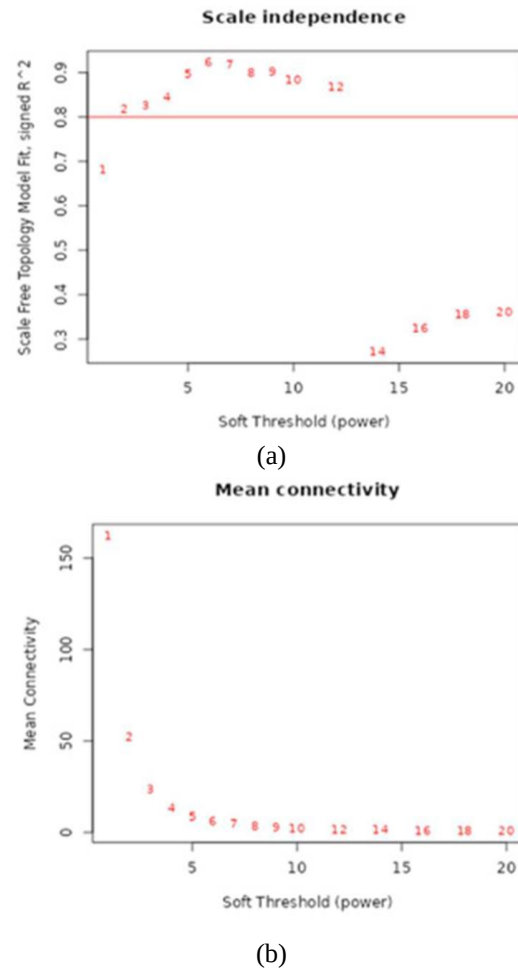


Figure 2. Soft-thresholding.

The constructed network of 601 genes was partitioned into seven modules (**Figure 3**). Networks corresponding to the genes in the turquoise and red modules were retained for downstream analysis because of their high connectivity. Genes in the other five modules showed no connectivity and were therefore excluded from subsequent analyses.

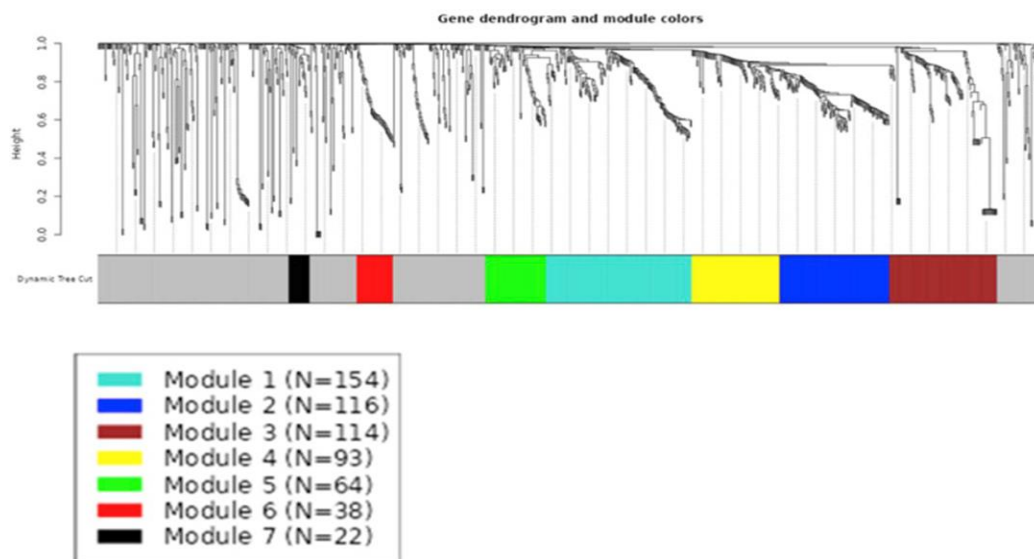
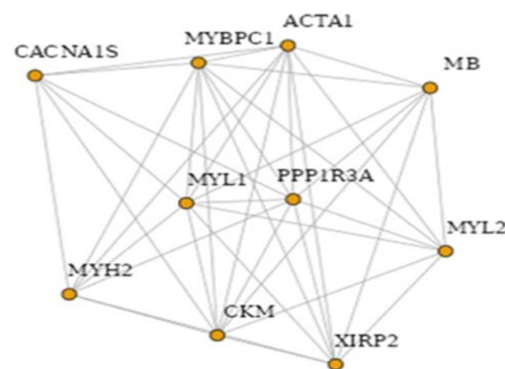


Figure 3. Gene dendrogram and module colors.

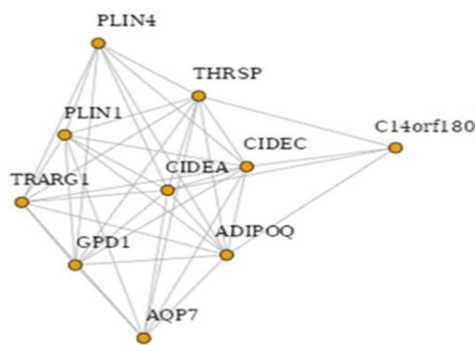
Figure 4 presents the top twenty genes ranked by network connectivity. Juyal *et al.* [23] identified C14orf180 as a key gene linked to rheumatoid arthritis. In addition, CIDEA and ADIPOQ were included during network construction; these genes have previously been associated with obesity and insulin resistance [24, 25]. PPP1R3A has been linked to diabetic nephropathy according to Ref. [26]. When processing large-scale gene expression datasets using bioinformatics approaches, the “curse of dimensionality” emerges as a major obstacle, often reducing performance in classification and prediction tasks [27]. To mitigate this problem, dimensionality reduction strategies such as gene co-expression networks are particularly valuable because they yield interconnected features with clear biological relevance.



Red module network

(b)

Figure 4. Gene network.



Turquoise module network

(a)

Table 1 and **Figure 5** compare three classification algorithms applied to rheumatoid arthritis prediction using imbalanced data. Both the IBK and Decision Tree models reached an identical accuracy of 92%. Nevertheless, building the Decision Tree required an additional 0.05 seconds compared to the Instance-Based Learning (IBK) approach. In contrast, logistic regression delivered the poorest performance in accuracy, precision, recall, and F-measure among the three algorithms. Furthermore, it took 8.49 seconds to construct the logistic regression model, making it the most sensitive to dataset imbalance in terms of training duration.

Table 1. Classification results for the imbalanced dataset.

Imbalanced dataset					
Classification algorithm	Accuracy	Precision	Recall	F-Measure	Time/s
Logistic regression	91	90	91	90	8.49
IBK	92	92	92	92	0
Decision Tree	92	92	92	92	0.05

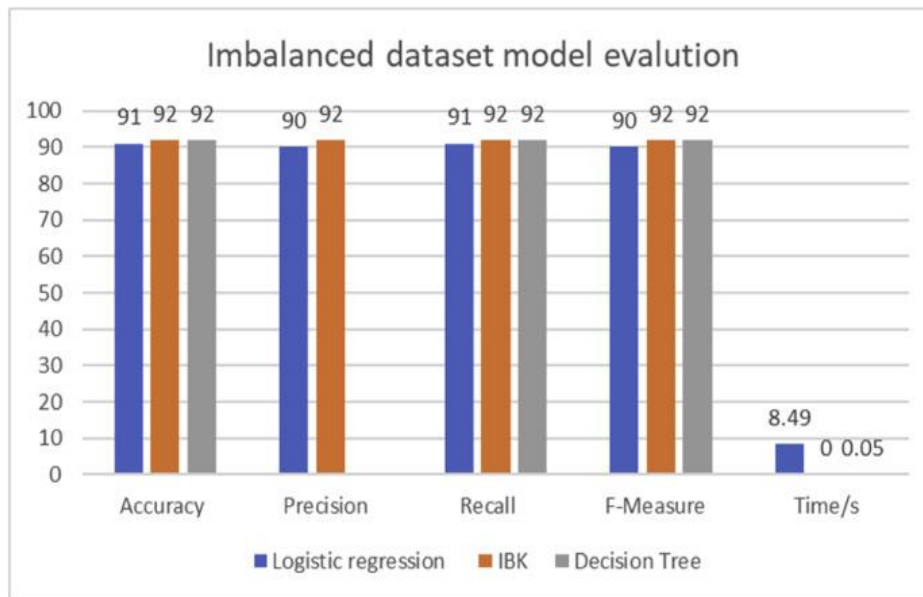
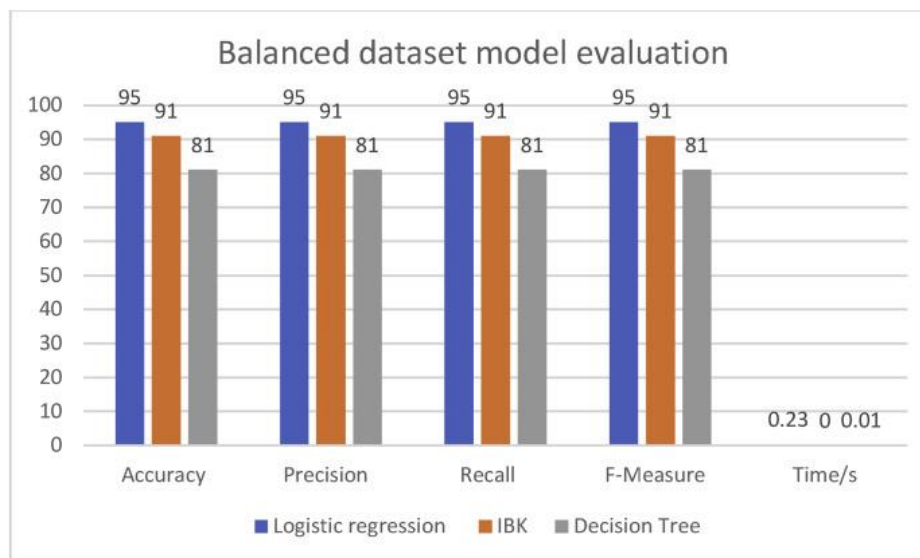
**Figure 5.** Model evaluation results for three algorithms on an imbalanced dataset.

Table 2 and **Figure 6** display a side-by-side comparison of the three algorithms for rheumatoid arthritis prediction. After applying the Spread Sub-Sample method—an under-sampling approach used to balance the dataset—both IBK and Decision Tree showed noticeable declines in accuracy, precision, recall, and F-measure. The IBK algorithm experienced a 1% drop in accuracy, while the Decision Tree saw an 11% reduction. These models generally favor the majority class, resulting in stronger predictions for the dominant group and inflated overall performance metrics before balancing. In this investigation, logistic regression was most affected by the original class imbalance, both in predictive accuracy and training duration. After balancing, its accuracy increased by 4% and model construction time dropped to 0.23 s. With a balanced dataset, logistic regression better identified patterns and relationships in the minority class, improving its effectiveness on underrepresented samples. However, statistical testing yielded a p-value of 0.93812, which

exceeds the 0.05 threshold, indicating that the observed difference in accuracy was not statistically significant. Beyond the gains in accuracy and precision, applying WGCNA to preprocess the gene expression data also substantially reduced the computational demands of running logistic regression. The reduced dataset size and complexity allowed the algorithm to execute more efficiently. The weaker results of logistic regression on the imbalanced data can be linked to the large data volume and its inherent complexity. Although processing speed is important, accuracy remains a key priority, especially when an appropriate trade-off between the two is necessary [28]. Logistic regression is a simple yet powerful method that can handle both binary and multi-class problems. In contrast, WGCNA is an effective tool for identifying biologically related, co-expressed genes through network connectivity. Consequently, integrating WGCNA with the Spread Sub-Sample under-sampling technique can improve accuracy and reduce execution time.

Table 2. Classification results for the balanced dataset.

balanced dataset					
Empty Cell	Accuracy	Precision	Recall	F-Measure	Time/s
Logistic regression	95	95	95	95	0.23
IBK	91	91	91	91	0
Decision tree	81	81	81	81	0.01

**Figure 6.** Model evaluation results for three algorithms on a balanced dataset.

Conclusion

This study introduced a pipeline that integrates WGCNA for dimensionality reduction, the Spread Sub-Sample technique to address class imbalance, and logistic regression as the primary classifier. This combination improved overall performance while significantly reducing model training time. We also tested two additional classifiers for predictive accuracy, first on the original imbalanced dataset and then after applying Spread Sub-Sample balancing. The rheumatic arthritis dataset was highly skewed, consisting of only 28 normal samples and 152 rheumatic arthritis samples. Through WGCNA, the number of features was reduced to 601 genes, which were grouped into seven modules. The red and turquoise modules were chosen for further analysis because they contained fully connected networks. Among the selected genes were several associated with rheumatic arthritis, including C14orf180, as well as CIDEA and ADIPOQ, which are also linked to obesity. On the imbalanced data, IBK and Decision Tree achieved high accuracy; however, their performance dropped considerably once the dataset was balanced. In summary, feature selection and class balancing before model

development are essential. This is because most machine learning algorithms are inherently biased toward the majority class, producing misleadingly high accuracy scores while failing to account for the minority class properly and resulting in poor representation across all categories.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Malmström V, Catrina AI, Klareskog L. The immunopathogenesis of seropositive rheumatoid arthritis: from triggering to targeting. *Nat Rev Immunol.* 2017;17(1):60-75. doi:10.1038/nri.2016.120
2. Yap M, Johnston RL, Foley H, MacDonald S, Kondrashova O, Tran KA, et al. Verifying explainability of a deep learning tissue classifier

- trained on RNA-seq data. *Sci Rep.* 2021;11(1):2641. doi:10.1038/s41598-021-81755-9
3. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: a data perspective. *ACM Comput Surv.* 2017;50(6):94. doi:10.1145/3136625
 4. Hany O, Abu-Elkheir M. Detecting vulnerabilities in source code using machine learning. In: *International Conference on Applied CyberSecurity*. Cham: Springer International Publishing; 2021. p. 35-41.
 5. Bader-El-Den M, Teitei E, Perry T. Biased random forest for dealing with the class imbalance problem. *IEEE Trans Neural Networks Learn Syst.* 2018;30(7):2163-72. doi:10.1109/TNNLS.2018.2875408
 6. Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inf Sci.* 2018;465:1-20. doi:10.1016/j.ins.2018.06.056
 7. Parsons Z, Banitaan S. Automatic identification of Chagas disease vectors using data mining and deep learning techniques. *Ecol Inf.* 2021;62:101270. doi:10.1016/j.ecoinf.2021.101270
 8. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf.* 2008;9(1):1-13. doi:10.1186/1471-2105-9-559
 9. Grassi L, Rechberger C. Practical low data-complexity subspace-trail cryptanalysis of round-reduced PRINCE. In: *Progress in Cryptology – INDOCRYPT 2016*. Cham: Springer International Publishing; 2016. p. 322-42. doi:10.1007/978-3-319-49890-4_18.
 10. Sachdeva RK, Bathla P, Rani P, Solanki V, Ahuja R. A systematic method for diagnosis of hepatitis disease using machine learning. *Innov Syst Softw Eng.* 2023;1-10. doi:10.1007/s11334-023-00534-7
 11. Trigka M, Dritsas E. Long-term coronary artery disease risk prediction with machine learning models. *Sensors.* 2023;23(3):1193. doi:10.3390/s23031193
 12. Hassan CAU, Khan MS, Shah MA. Comparison of machine learning algorithms in data classification. In: *2018 24th International Conference on Automation and Computing (ICAC)*. IEEE; 2018. p. 1-6. doi:10.23919/ICAC.2018.8748998
 13. Gomes JC, de Santana MA, Masood AI, de Lima CL, Dos Santos WP. COVID-19's influence on cardiac function: a machine learning perspective on ECG analysis. *Med Biol Eng Comput.* 2023;1-25. doi:10.1007/s11517-023-02835-6
 14. Lu H, Chen J, Yan K, Jin Q, Xue Y, Gao Z. A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing.* 2017;256:56-62. doi:10.1016/j.neucom.2016.12.069
 15. García V, Sánchez JS. Mapping microarray gene expression data into dissimilarity spaces for tumor classification. *Inf Sci.* 2015;294:362-75. doi:10.1016/j.ins.2014.10.006
 16. Alhenawi EA, Al-Sayyed R, Hudaib A, Mirjalili S. Feature selection methods on gene expression microarray data for cancer classification: a systematic review. *Comput Biol Med.* 2022;140:105051. doi:10.1016/j.combiomed.2021.105051
 17. Hambali MA, Oladele TO, Adewole KS. Microarray cancer feature selection: review, challenges and research directions. *Int J Cogn Comput Eng.* 2020;1:78-97. doi:10.1016/j.ijcce.2020.09.001
 18. Sun L, Zhang X, Qian Y, Xu J, Zhang S. Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. *Inf Sci.* 2019;502:18-41. doi:10.1016/j.ins.2019.05.089
 19. Saberi-Movahed F, Rostami M, Berahmand K, Karami S, Tiwari P, Oussalah M, et al. Dual regularized unsupervised feature selection based on matrix factorization and minimum redundancy with application in gene selection. *Knowl-Based Syst.* 2022;256:109884. doi:10.1016/j.knsys.2022.109884
 20. Ge SX, Son EW, Yao R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinf.* 2018;19(1):1-24. doi:10.1186/s12859-018-2339-2
 21. Saravanan N, Gayathri V. Performance and classification evaluation of J48 algorithm and Kendall's based J48 algorithm (KNJ48). *Int J Comput Trends Technol.* 2018;59(2):73-80.
 22. Gulati P, Sharma A, Gupta M. Theoretical study of decision tree algorithms to identify pivotal factors for performance improvement: a review. *Int J Comput Appl.* 2016;141(14):19-25. doi:10.5120/ijca2016909935
 23. Juyal G, Pandey A, Garcia SL, Negi S, Gupta R, Kumar U, et al. Stratification of rheumatoid arthritis cohort using Ayurveda based deep phenotyping

- approach identifies novel genes in a GWAS. *J Ayurveda Integr Med.* 2022;13(3):100578. doi:10.1016/j.jaim.2022.100578
24. Ross KA. Evidence for somatic gene conversion and deletion in bipolar disorder, Crohn's disease, coronary artery disease, hypertension, rheumatoid arthritis, type-1 diabetes, and type-2 diabetes. *BMC Med.* 2011;9:1-29. doi:10.1186/1741-7015-9-19
 25. Warodomwichit D, Shen J, Arnett DK, Tsai MY, Kabagambe EK, Peacock JM, et al. ADIPOQ polymorphisms, monounsaturated fatty acids, and obesity risk: the GOLDN study. *Obesity.* 2009;17(3):510-7. doi:10.1038/oby.2008.553
 26. Mansour A, Mousa M, Abdelmannan D, Tay G, Hassoun A, Alsafar H. Microvascular and macrovascular complications of type 2 diabetes mellitus: exome wide association analyses. *Front Endocrinol.* 2023;14:1143067. doi:10.3389/fendo.2023.1143067
 27. Han Y, Zhou Y, Li H, Gong Z, Liu Z, Wang H, et al. Identification of diagnostic mRNA biomarkers in whole blood for ankylosing spondylitis using WGCNA and machine learning feature selection. *Front Immunol.* 2022;13:956027. doi:10.3389/fimmu.2022.956027
 28. Awad FH, Hamad MM, Alzubaidi L. Robust classification and detection of big medical data using advanced parallel K-means clustering, YOLOv4, and logistic regression. *Life.* 2023;13(3):691. doi:10.3390/life13030691