Decoding Alopecia Areata – Where Genetics Meets Machine Learning

---

Abstract

Alopecia Areata (A.A) is an autoimmune disease that results in hair loss. Australia has a 2% incident rate, with slightly more proportion of female patients compared to male patients, where 40% of sufferers had symptoms by age 20. [1] Current methods of clinical diagnosis include imaging, medical history review, and physical examinations on the patient. Due to the subjective nature of these diagnosis techniques, and physical similarities with other hair loss conditions like Scarring Alopecia and Female-Pattern-hair-loss, around 36.2% of all A.A cases are misdiagnosed. [2] As a solution, my proposed framework combines machine learning with genetic testing to predict the presence of A.A. I used in-sample data retrieved from peer-reviewed dataset featuring A.A subtypes including AAP, AAP.T, AT, and AU. [3] This was followed by Explanatory Data Analysis, feature selection, and four classification techniques: SVM, KNN, LASSO, and Random Forest. My SVM model was then adapted into a Shiny App intended as a more precise dermatological and clinical tool for A.A diagnosis. The corresponding code used is stored in a qmd file along with the report.

Introduction

Hair loss is defined as the "falling of scalp hairs in sufficient quantities" [4]. Alopecia Areata (AA) is a prevalent autoimmune condition affecting hair follicles, where stages may range from getting patches of hair loss to complete baldness. It begins when the body's autoimmune system starts to target the hair follicles, disturbing their normal functioning and preventing subsequent hair growth [5]. Being more prevalent in patients with a family history of the condition, genetic testing through scalp biopsy will provide another layer of effective prediction beyond physical examinations for early stages of A.A, especially when symptoms have not shown yet.

Machine learning (ML) encapsulates the ability of prediction using computer algorithms that have demonstrated the ability to learn and adapt to both observable and unobservable patterns. ML techniques have recently shown effectiveness in predicting and classifying various diseases like CT scans for brain tumors [6], pulmonary diseases, and most recently – diagnosing COVID-19 severity levels [7]. In dermatology, notable models like SVM, Random Forest, and KNN have been used to classify dandruff and hair-growth disorders [8]. My model, inspired by this concept, incorporates the models along with another experimental LASSO model to compare predictive performance. I then built a predictive mechanism that accurately identifies genes and biomarkers associated with alopecia until a solid diagnosis of the disease can be made. The resulting product will be used for biomarker identification and diagnosis intended for clinicians to use, which will continuously be enhanced by the data collected from real patients and then logged into my App- GenoDerm's database.

In past cases, dermatological genetic models often helped develop targeted drug development like JAK inhibitors (Tofacitinib and Baricitinib) currently used to treat moderate to severe Alopecia [9].

Data Description

GSE68801, the dataset used for this project, is a dataset from the NCBI Gene Expression Omnibus (GEO). It contains gene activity measurements taken using the Affymetrix Human Genome U133 Plus 2.0 Array (GPL570). All samples come from small pieces of scalp skin (punch biopsies). These biopsies were stored in PAXgene containers and sent to the lab for processing:

- Patchy Alopecia Areata (AAP): 22 samples
- Patchy Alopecia Areata post-treatment (AAP.T): 6 samples
- Alopecia Totalis (AT): 9 samples
- Alopecia Universalis (AU): 23 samples
- Normal healthy controls: 36 samples

In total, 96 biopsy specimens were profiled, which helps in the comparison of different clinical subtypes of alopecia areata and healthy skin.

Exploratory Data Analysis

Exploratory data analysis is an essential step to undertake before the process of modelling. EDA was conducted to gain a better understanding of the data which can aid in not just gaining a better understanding of the data but also to uncover hidden patterns, verify data quality and detect outliers or batch effects.

A box plot of the expression values helps in seeing how the data is spread out, which can help reveal if some patients tend to have higher or lower typical values, how much the values vary from one patient to another, and whether there are any odd points that stick out. From analyzing the box plot (Appendix 1), we can observe that the expression distribution is consistent, with the median hovering at the same level which indicates that there are no outlier samples that can skew the analysis and that there are no batch effects.

Differential expression analysis is a way to find which genes (or probes) behave differently between two groups – in this case, between the “control” and the “patient” and how. The column focus for this analysis is the adjusted p-value, which helps understand whether the differences are due to chance or not. Observing a small p-value indicates that the observed differences are not by chance and this is significant because having a difference between the patient and the control group suggests that the gene’s expression change is real- likely linked to the disease process.

The MA plot is a scatter plot that helps in visualizing the relationship between the log ratio and mean values of two variables which shows which genes are up-regulated and down-regulated genes and shows how these fold-changes are dependent on the overall expression level. The up-regulated genes (marked in red) and the down-regulated genes (marked in blue) are not symmetrical, with the down-regulated genes being greater than the up-regulated genes (Appendix 3). This indicates that the disease is characterised more by repression rather than activation.

The volcano plot is a scatter plot (Appendix 4) that has the log2 values fold change and the -log10 of the adjusted p-values. The volcano plot also shows a clear skew toward the down-

regulated genes where there's a denser cloud of features on the left. This also reveals that the disease is characterized more by repression rather than activation.

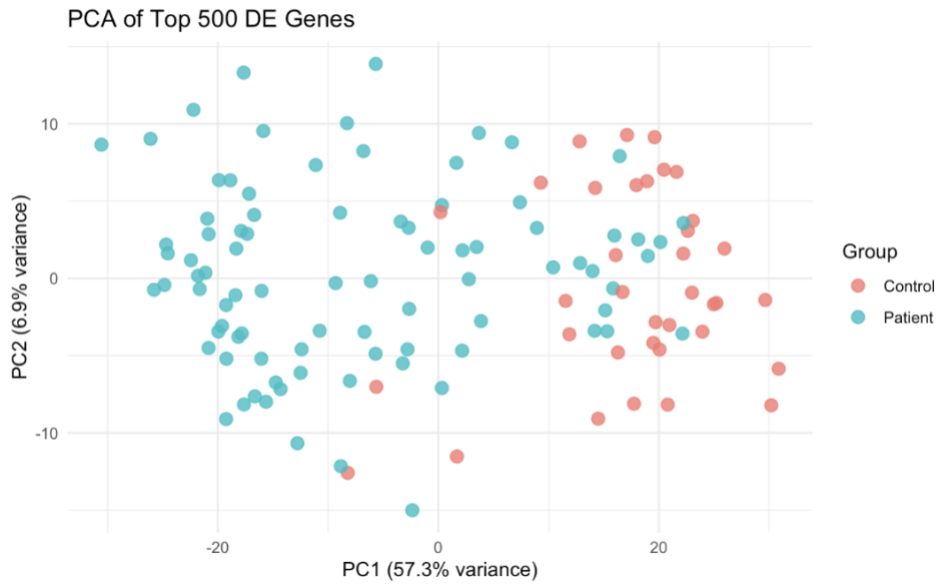


Figure 2: Principal Component Analysis

Principal Component Analysis reduces a high dimensional dataset into a smaller set that captures the greatest source of variation. In the PCA, the principal component 1, which accounts for 57.3% variance, clearly separates most patient samples from controls with a few individuals that do cross into the opposite cluster. This indicates that the disease status is the most significant variation in the data. This means that the dominant pattern in the data is the contrast between AA and healthy scalp, and this contrast accounts for over half the signal in these genes. The small overlap that we see can indicate that there can be additional that could still remain after accounting for the main disease effect. Principal component 2 only accounts for 6.9% variations and doesn't separate the disease from the control group which indicates that it reflects secondary factors like batch-to-batch differences or sex-specific expression patterns.

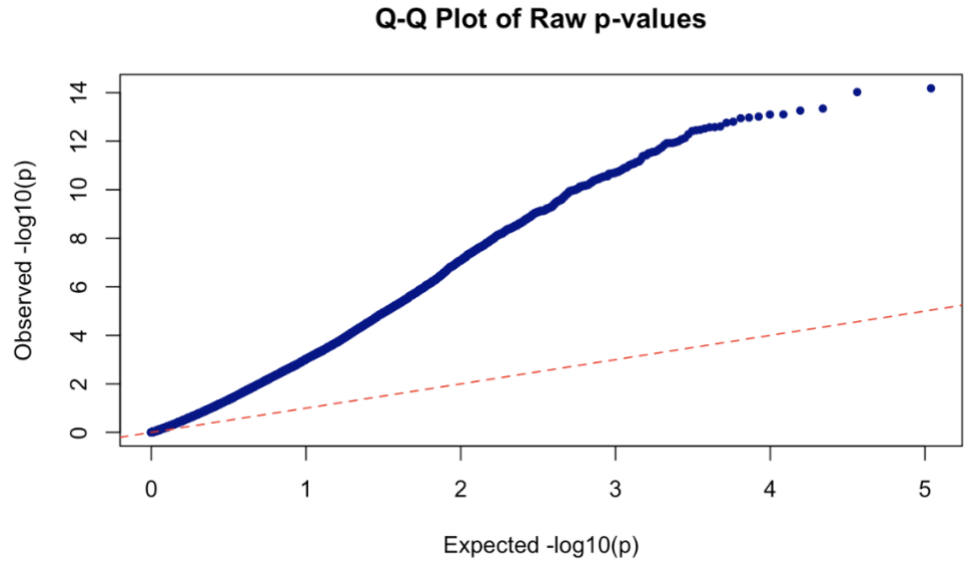


Figure 3: Quantile-Quantile Plot

The Quantile-Quantile plot helps understand if a dataset follows a theoretical distribution by comparing the observed distribution of features against what one would expect if no genes were truly differentially expressed. The upward deviation with almost all the genes above the red line (where y=x) indicates that more genes have smaller p-values than what would occur by chance. This is complimented by the gentle upward-convex shape across most of the middle quantiles and not just at the extreme tails that many genes have p-values smaller than expected under the null.

Since the boxplot and PCA did not show any outstanding outliers, there was no removal of them. Moreover, the box plot revealed that the data is standardised which meant that there was no need for processes like log2 transformation and quantile normalisation. However, for the modelling process, I accounted for gender and age as well as these factors influence the biological variance and therefore including them will make the model more well-rounded.

Methodology

The Tools used for data processing, EDA, and model training were from R-studio packages. Biobase used data structures (like ExpressionSet) to analyze high-throughput Genetic data like mines'. Bioconductor was a project that provided tools: GEOquery, limma, caret, Random Forest, e1071.

First, the dataset was queried from the Gene Expression Omnibus (GEO) public repository where the raw data was stored in arrays. The four subtypes were identified and counted. The expression distribution of my samples were visualized using box plots from the 'ExpressionSet' function via Biobase package, where all distributions appeared to follow a normal distribution with a right skew and no deviation from the range of Expression Value = 4 to 5. This indicates good quality data.

Exploratory Data analysis was then done to map out the nature of every sample's P Values via MA plot, PCA and Volcano plot of the top 500 most significantly differentially expressed genes. This was done to visually separate down and up-regulated genes to identify the most significant genes.

Data cleaning of the P-values was another step to ensure only the most significant genes were included in model selection and was done by filtering all P.Val < 0.05 to a new data frame: sig_genes for the 21 most significant genes. Feature selection was then done using LASSO for the genes to be used in prediction data, where a training and testing split of 80:20 was done to allow the model to learn patterns while also considering unseen data. Cross Validation was then used to integrate the most significant genes from the LASSO selection with the variables Gender and Age to form a complete feature matrix, resulting in 23 predictor variables.

For the quantified CV performance across all my selected models, I chose to use the Receiver-operating characteristic (ROC) curve detailing the model's Cross Validated Sensitivity and Specificity. Sensitivity measures the ability to correctly identify True Positives while Specificity measures the ability to identify True Negatives. Positive likelihood and negative likelihood ratios respectively measure how much a positive/ negative result is likely in someone with the condition than without. I chose a threshold of 0.5 to balance all factors to ensure minimal bias towards one metric and maximising performance. ROC curve maps the

tradeoff between the two metrics, where Area Under Curve (AUC) summarizes overall performance [10].

$$Sensitivity = \frac{TP}{(TP + FN)}$$

$$Specificity = \frac{TN}{(TN + FP)}$$

$$Positivelikelihoodratio(LR+) = \frac{sensitivity}{(1-specificity)}$$

$$Negativelikelihoodratio(LR-) = \frac{(1-sensitivity)}{specificity}$$

The first model was Random Forest, where the adjusted parameter was 'mtry' (number of trees) and the optimal mtry = 2, indicating that the model only requires a small number of features at each node to make the correct decision. The ROC curve showed that the curve was distant from the diagonal line, with most of the area nearing the upper left corner, indicating excellent performance (Appendix 4).

The second model was the Support Vector Machine- SVM (Radial) derived by the svm() function from e1071 package and was the best performing model and functions to determine the linear and non-linear separability with the aid of a hyperplane [11]. It uses a kernel method which transforms two-dimensional data into higher dimensions to separate the data. In my case, this model helps enforce a strict boundary that forces each diagnosis closer to either 0 or 1 classification results. The kernel multiplies kernel function k with a dot product $xi \,.\, xj$ as seen below:

$$\max_{\propto} \sum_{i=1}^{m} \propto_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \propto_i \propto_j y_i y_j K\left(x_i \cdot x_j\right).$$

I also applied the Radial Basis Function as the kernel function and contains a parameter sigma (γ) as shown below:

$$K\left(x_i.x_j\right) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right).$$

In my model, the C parameter is a penalty parameter controlling the regularization strength of the SVM. A smaller C allows for more misclassification but less features, and a larger C is stricter on misclassifications but may risk overfitting. Overall, a smaller sigma $\gamma = 0.01$ combined with a larger $C = 10$ performed the best, indicating that the model requires a higher complexity and a narrower decision boundary which resulted in the largest AUC (Appendix 5), meaning that the model has a strong ability to distinguish patients from the control group.

The next model was the K-Nearest Neighbour (KNN) where several values of K from 5 to 23 were used to test the best parameter. The smaller the K value, the more complex the model and the more prone it is to overfitting. I decided on $K = 3$ selected based on the maximum ROC value (Appendix 6).

My final trained and tested model was the LASSO which used lambda = 0.001, which is a relatively small regularization strength, allowing the model to retain more features (Appendix 7). This suggests I must consider the risk of overfitting when using this model.

To diagnose, I chose a classification model that gives binary results like 0 (no A.A present) and 1 (A.A present). The log2 transformation on each of the corresponding genes was derived from the gene expression samples from each scalp biopsy and was then fed into my Shiny App database. My data-pipeline process followed the schematic overview in figure 4 for my final product- GenoDerm app.

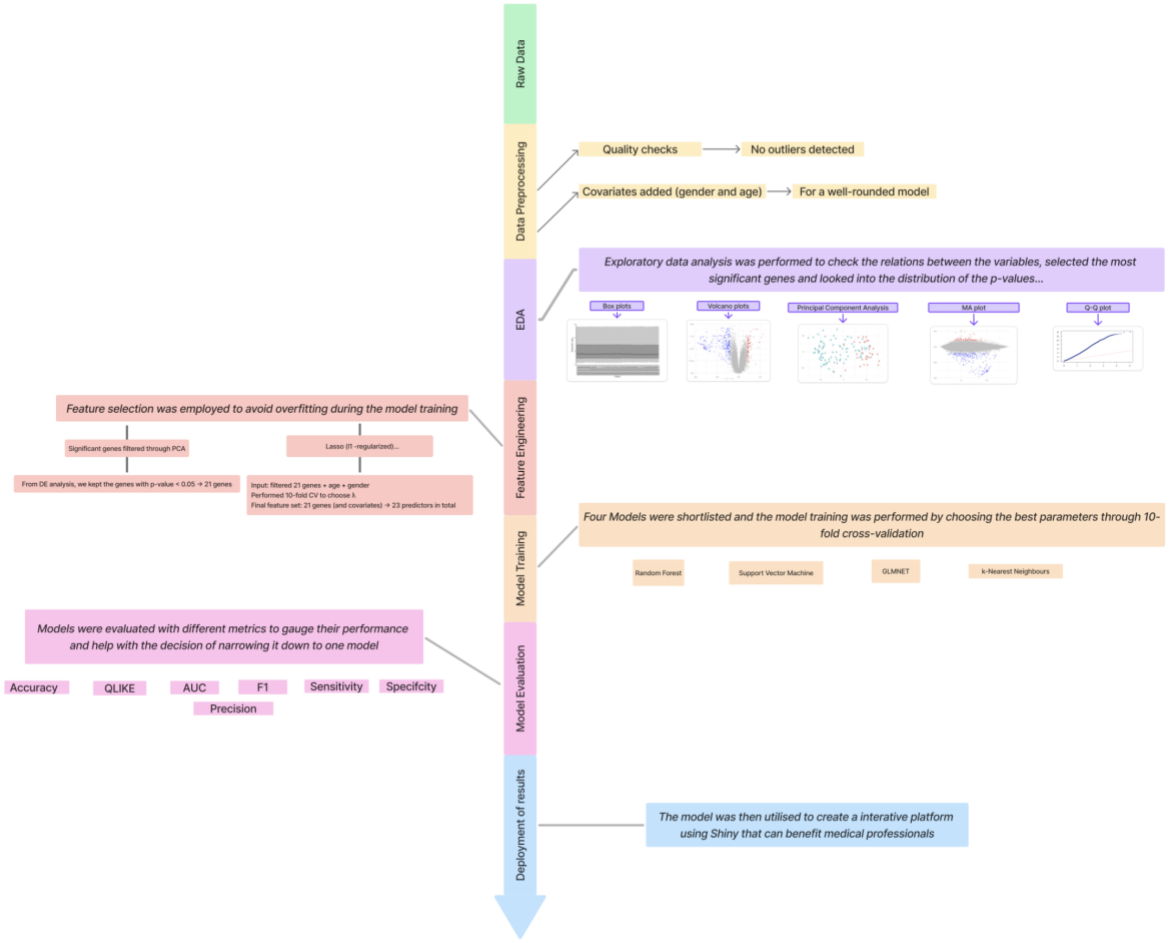


Figure 4. Schematic overview

Results

The performance results of ROC from all four models showed SVM (Radial) as having the highest performance (Appendix 8). Figure x shows the two predicted outcomes 0 and 1. When the actual value is 1 and the predicted value is also 1, the outcome is True Positive (TP), otherwise the outcome is False Negative (FN). The confusion matrix Figure 5 visualizes each outcome:

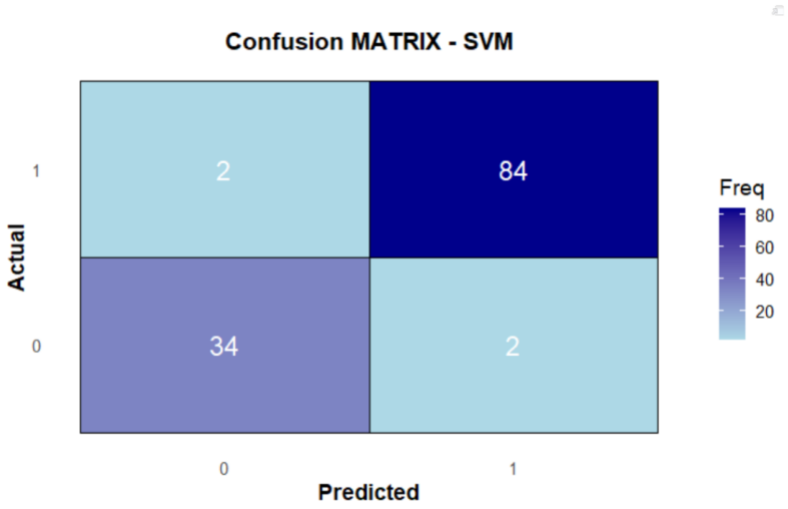


Figure 4. Schematic overview

The accuracy for each model's prediction is derived from dividing the number of truly classified results by the number of test results and multiplying by 100 to get a percentage:

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

The factor() function from baseR separated each class of True Positives and False Positives with True Negatives and False Negatives. Further classification techniques include Precision, Recall, and F1 scores. Recall represents the ability to class TP, and Precision calculates the ratio of TP over TP and FN to ensure no positive values were marked as negative.

$$\text{Pr}\,e\,cision = \frac{TP}{TP+FP} \times 100\%$$

$$Recall = \frac{TP}{TP+FN} \times 100\%$$

$$F1score = \frac{(2*precision*recall)}{precision+recall} \times 100\%$$

Q-like (Quasi-Likelihood Loss) measures how close the predicted variance is to the realized variance, where the lower the score, the smaller the loss resulting from under/overestimation penalties. While primarily used in financial forecasting, I decided to include this metric as SVM can use the predicted realized volatility via Q-like to assess the accuracy of variance forecasts. As LASSO intends to sparce linear model for volatility, Q-like also considers misclassifications in this context. KNN used Q-like to measure non-parametric regression on volatility and finally, Random Forest used it to capture non-linear volatility for forecasting. Figure 5 and Table 1 shows that SVM model with AUC = 0.99, Accuracy = 0.97, F1 = 0.98 and Q-like = 0.12 achieved the best scores for all the metrics mentioned.

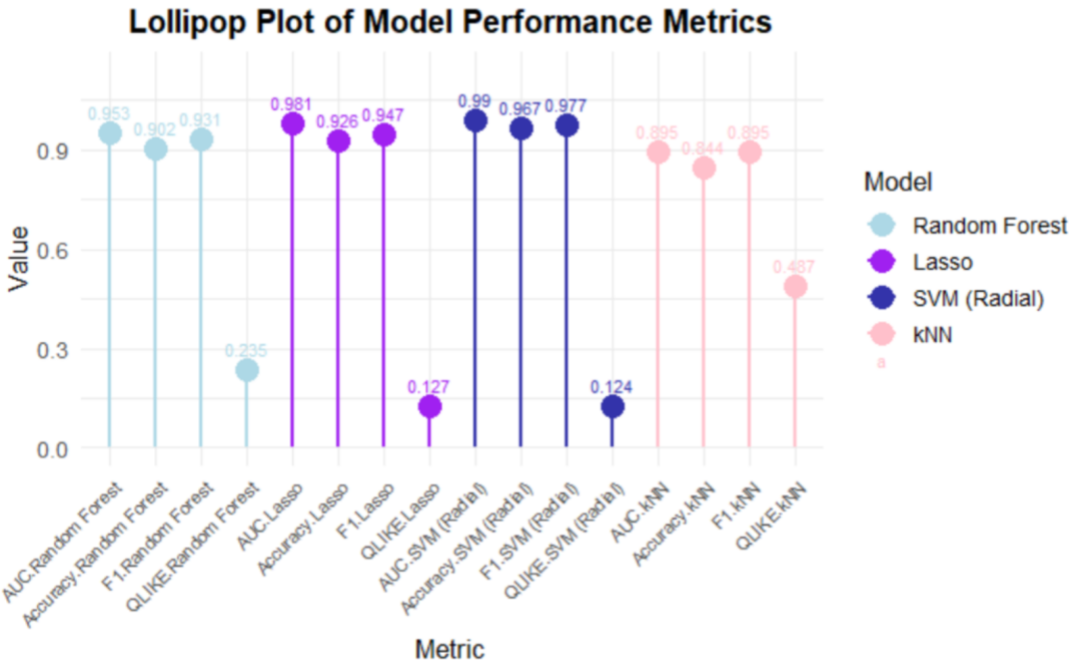


Figure 5. Performance Metrics Lollipop Plot

Model	AUC	Accuracy	F1-Score	Q-like
Random Forest	0.953	0.902	0.931	0.235
LASSO	0.981	0.926	0.947	0.127
SVM	0.990	0.967	0.977	0.124
kNN (k=5)	0.895	0.844	0.895	0.487

Table 1. Performance Metrics

Application: Shiny App

The first interface of the Alopecia Areata Smart Risk Calculator captures critical demographic variables, age (numeric) and gender (dropdown), that are incorporated as covariates in the predictive model. These are recognized contributors to gene expression variability and enhance the model's accuracy. Users may either input data manually or upload a .csv file structured according to the model's feature schema, allowing batch processing for cohort-level predictions (Appendix 9). The interface prioritizes usability and clinical practicality, with a clean layout and intuitive design. The "Next: Gene Expression" button advances users to the next step, maintaining a structured, low-error workflow that supports reproducibility. By inputting the logged (base 2) scalp biopsy gene expression sample data, the SVM pipeline generates a result of either very close to 1 or 0 for diagnosis (Appendix 10).

Batch predictions tab supports high-throughput predictions using bulk .csv uploads. Each row must include demographic and gene expression data matching the model's schema. By clicking "Run Batch Prediction," users can obtain predictions across multiple samples using the SVM model (Appendix 11). Results output as "Yes" or "No" and can be downloaded for integration into clinical or research workflows, enhancing the app's scalability and translational impact.

The app features a Dataset Overview section summarizing the GSE68801 dataset used to train the predictive model. It includes sample counts across Alopecia Areata subtypes and

controls, along with the dataset source, platform, and related publications. Download links for both processed and raw gene expression data are provided to support transparency and reproducibility. Additionally, the About Alopecia section offers a brief introduction to Alopecia Areata as an autoimmune condition, outlines its main subtypes (Patchy, Totalis, Universalis), and touches on common triggers, emotional impact, and general treatment approaches. Together, these sections provide essential context for users without overwhelming the interface.

Discussion

The objectives of this project were to develop an accurate and reproducible model for scalp biopsy analysis which will take genetic expression data for the diagnosis of Alopecia Areata. My SVM model classified linear and non-linear data using a line generated on a hyperplane which was then enhanced by the Radial kernal's accuracy. My choice of boundaries and parameters optimized the final model, which earned high performance results compared to other historically competent models like Random Forest, KNN, and LASSO. High AUC, Accuracy and F1 scores indicate the model performs very well in distinguishing between Alopecia-Positive and Alopecia-Negative cases, with strong overall classification, balanced precision and recall with robust discrimination ability. Paired with a low Q-like score, my model makes stable, confident predictions with low forecast error. My final model is thus accurate and stable in its predictions of Alopecia Areata.

A notable shortcoming of this project is that the model was not fully verified for robustness. Due to the limited research resources available related to Alopecia Areta, I could not find a dataset that matched the columns of genes that were used to train the model. Moreover, since the project uses an SVM model which produces a binary output, the app does not produce a probability value that measures the likelihood on a scale but rather produces a "yes" or a "no". I can, however, verify the high accuracy and performance metrics achieved by my model and use my product as a diagnostic support tool for dermatologists and clinicians to identify early genetic signs of alopecia. Due to SVM's use of a kernal, firm boundary parameters for outcomes and reproducibility, my model can be used to collect and fit clinical data in the future which satisfies my purpose for a clinical tool that accurately diagnoses based on genetic data.

Conclusion

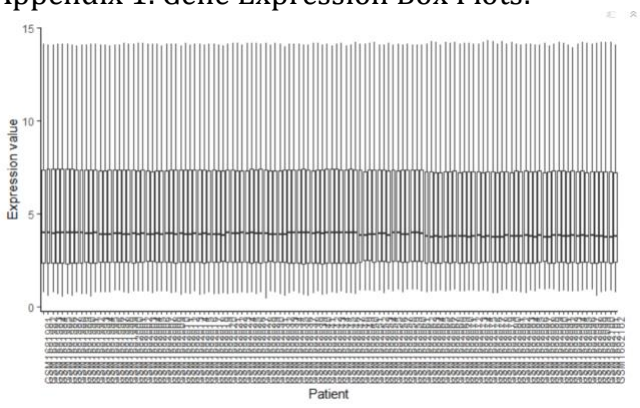
This project developed a gene-based prediction tool for Alopecia Areata (A.A) using my refined SVM model. After going through my data pipeline process, the model was integrated into a user-friendly Shiny app aimed at supporting medical researchers and clinicians who may use this tool to more accurately diagnose a patient with A.A beyond using just subjective visual trademarks. By using binary yes/no classification outcomes, I ensure practical ways to future model enhancements without limiting the model to subtypes and instead focusing on common genetic trademarks of A.A as a whole. This app serves as a foundational diagnostic aid and offers a promising step to early diagnosis of A.A before symptoms arise, which will help prevent lowered quality of life or mental wellbeing, especially for younger patients. Future work should focus on clinical validation, incorporation of probabilistic outputs, and expansion to larger, more diverse datasets to enhance the model's robustness and real-world applicability

References

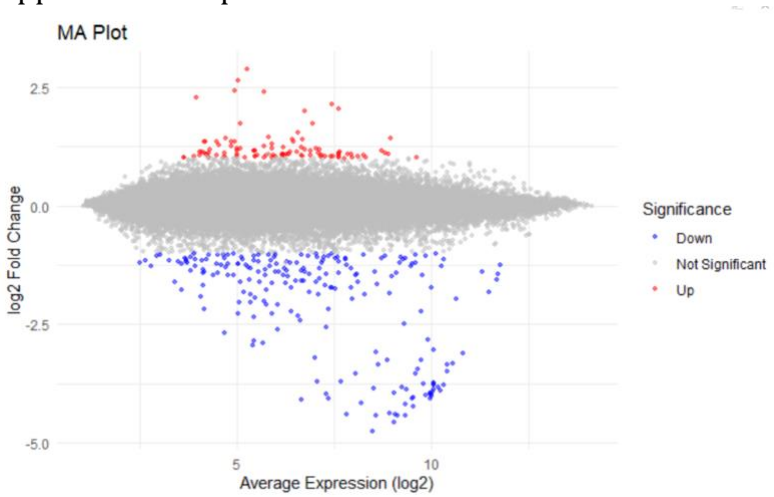
1. 'Prevalence and impact of cicatricial alopecia misdiagnosis' Journal of the American Academy of Dermatology, Volume 81, Issue 4, AB107
2. C. H. Pratt, L. E. King, A. G. Messenger, A. M. Christiano, and J. P. Sundberg, "Alopecia areata," *Nature Reviews Disease Primers*, vol. 3, no. 1, pp. 1–17, 2017.
3. Jabbari A, Cerise JE, Chen JC, Mackay-Wiggan J et al. Molecular signatures define alopecia areata subtypes and transcriptional biomarkers. *EBioMedicine* 2016 May;7:240-7. PMID: 27322477
4. N. S. Sadick, "New-generation therapies for the treatment of hair loss in men," *Dermatologic Clinics*, vol. 36, no. 1, pp. 63–67, 2018.
5. T. Simakou, J. P. Butcher, S. Reid, and F. L. Henriquez, "Alopecia areata: a multifactorial autoimmune condition," *Journal of Autoimmunity*, vol. 98, pp. 74–85, 2019.
6. G. Mathiyalagan and D. Devaraj, "A machine learning classification approach based glioma brain tumor detection," *International Journal of Imaging Systems and Technology*, vol. 31, pp. 1424–1436, 2021.
7. P. Wu, H. Ye, X. Cai et al., "An effective machine learning approach for identifying non-severe and severe coronavirus disease 2019 patients in a rural Chinese population: the Wenzhou retrospective study," *IEEE Access*, vol. 9, pp. 45486–45503, 2021.
8. W.-C. Wang, L.-B. Chen, and W.-J. Chang, "Development and experimental evaluation of machine-learning techniques for an intelligent hairy scalp detection system," *Applied Sciences*, vol. 8, no. 6, 2018.
9. Egeberg A, Linsell L, Johansson E, Durand F, Yu G, Vañó-Galván S. Treatments for Moderate-to-Severe Alopecia Areata: A Systematic Narrative Review. *Dermatol Ther (Heidelb).* 2023 Dec;13(12):2951-2991. doi: 10.1007/s13555-023-01044-5. Epub 2023 Oct 13. PMID: 37833617; PMCID: PMC10689337.
10. Çorbacıoğlu ŞK, Aksel G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turk J Emerg Med.* 2023 Oct 3;23(4):195-198. doi: 10.4103/tjem.tjem_182_23. PMID: 38024184; PMCID: PMC10664195.
11. Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. doi: 10.1007/3-540-44673-7_12.

Appendix

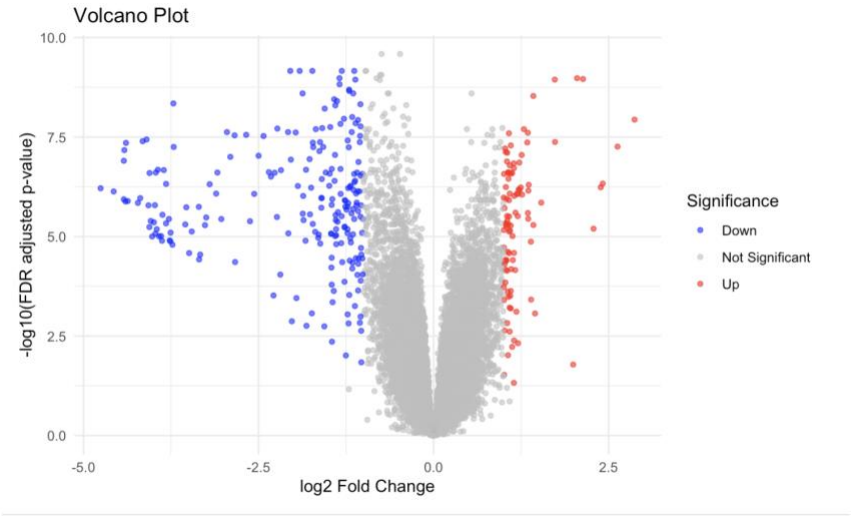
Appendix 1. Gene Expression Box Plots:



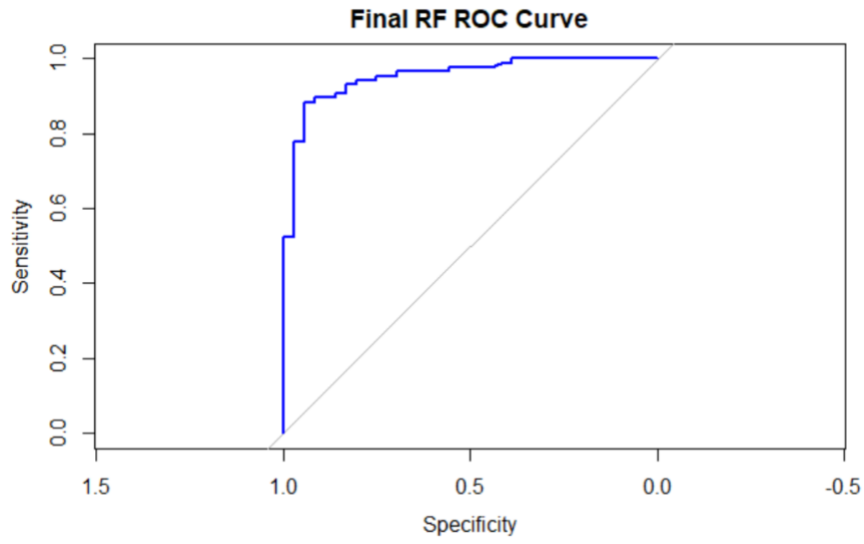
Appendix 2. MA plot:



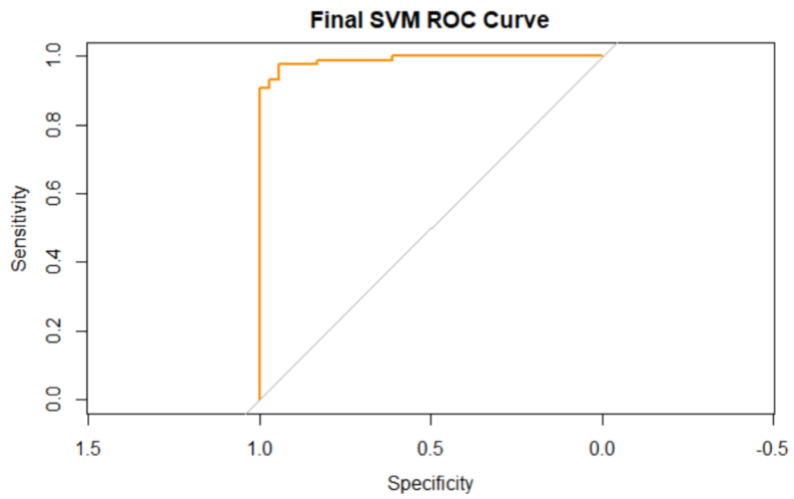
Appendix 3. Volcano plot:



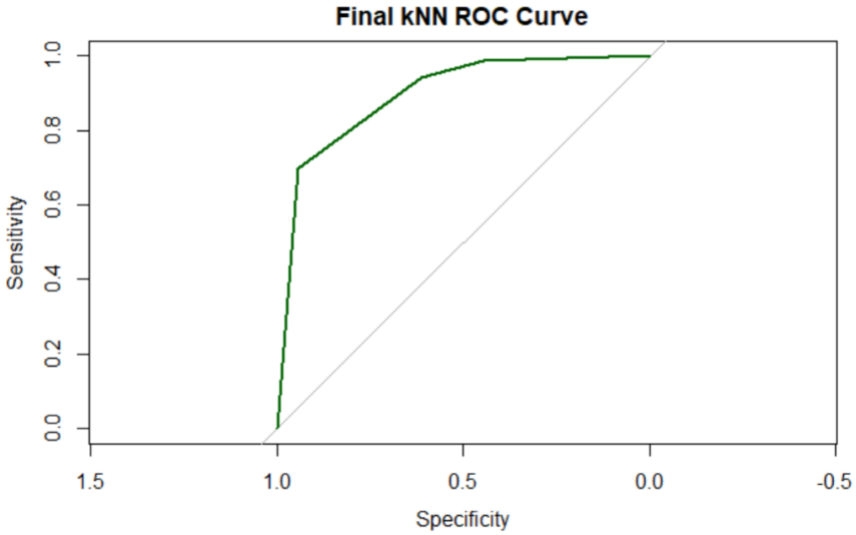
Appendix 4: ROC curve, RF:



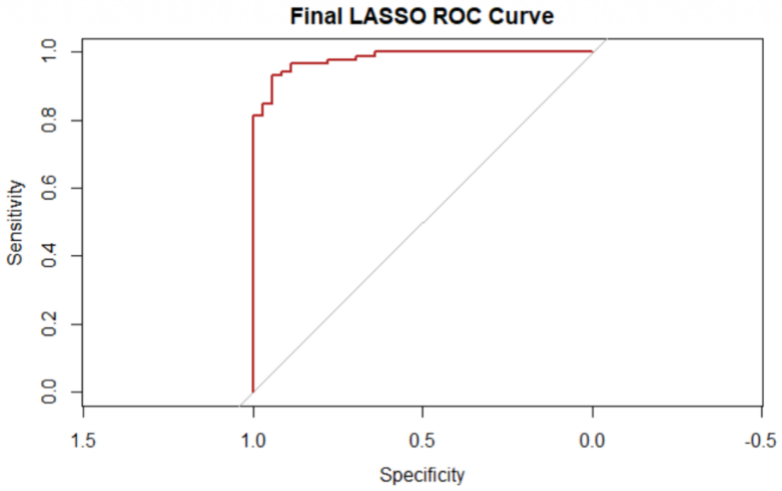
Appendix 5. ROC curve, SVM:



Appendix 6: ROC curve, KNN:



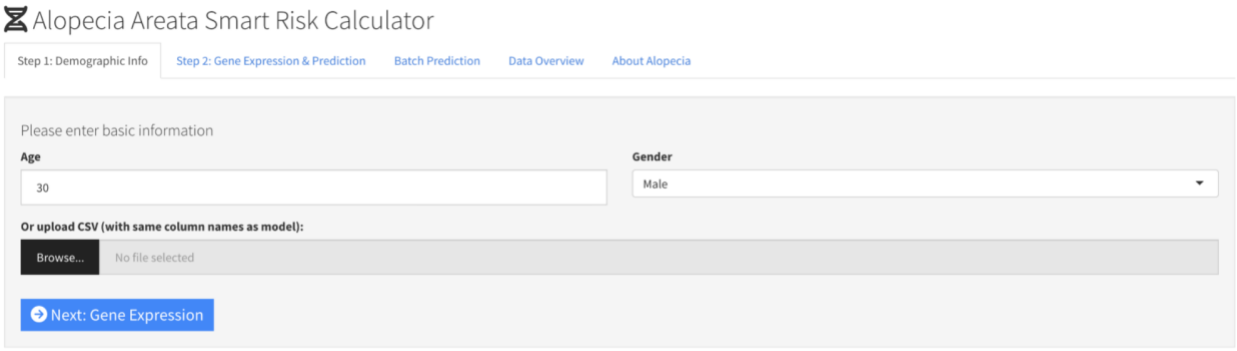
Appendix 7: ROC curve, LASSO:



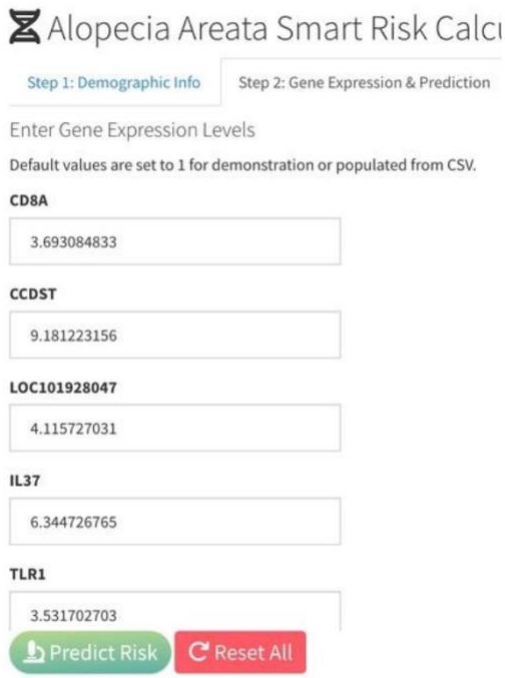
Appendix 8. Model Parameters:

Model <chr>	Parameters <chr>	ROC <dbl>
LASSO	lambda = 0.001	0.9663194444
Random Forest	mtry = 2	0.9604166667
SVM (Radial)	C = 10, sigma = 0.01	0.9824074074
kNN	k = 3	0.8851851852

Appendix 9. Demographic info:



Appendix 10. Gene Expression and Prediction:



Appendix 11. Batch Predictions: