International Journal of Artificial Intelligence in Medical Issues

Volume 3 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v3i1.360

Research Article

Predicting Hair Loss with Machine Learning: A Multi-Factor Analysis

M Ikbal Siami [1,*]; Huzain Azis [2]
[1] Institut Teknologi dan Bisnis Stikom Ambon, Kota Ambon, Maluku, Indonesia, siami25@gmail.com
[2] Universitas Kuala Lumpur, 50250 Kuala Lumpur, Malaysia, huzain.azis@s.unik.edu.my
Correspondence should be addressed to M. Ikbal Siami; siami25@gmail.com

Received 10 February 2025; Revised 15 February 2024; Accepted 30 April 2024; Published 30 May 2025

Abstract:
Hair loss is a multifactorial condition influenced by genetics, hormonal imbalance, lifestyle choices, and environmental factors. This study investigates the potential of machine learning (ML) to predict hair loss using a diverse dataset comprising categorical and numerical indicators related to these contributing variables. We applied an extensive data preprocessing pipelineincluding handling missing values, frequency encoding, and engineered interaction featuresto improve model input quality. Five ML algorithms (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost) along with an ensemble voting classifier were trained and evaluated on a balanced dataset. While performance metrics such as accuracy and F1-score remained modest, with the highest values around 50%, the analysis revealed the prominent role of age, stress, and nutritional deficiency in hair loss. Despite the limited predictive capability of the current feature set, this study presents a reproducible framework for ML-driven health diagnostics and identifies key directions for future work. Enhancing data granularity and incorporating richer clinical inputs could significantly boost prediction accuracy in subsequent studies.

Keywords: Hair Loss Prediction, Machine Learning, Feature Engineering, Ensemble Learning, Medical Informatics, Classification, Data Pre-processing.

Dataset link: https://www.kaggle.com/datasets/amitvkulkarni/hair-health

1. Introduction

Hair loss is a multifactorial condition affecting individuals across age groups and genders, with causes ranging from genetics and hormonal imbalances to lifestyle and environmental stressors. The psychosocial and emotional impact of hair loss is well-documented, particularly in younger adults, leading to a growing interest in early detection and preventive strategies [1]. While clinical diagnostics rely heavily on physical examination and history-taking, recent advances in artificial intelligence (AI) have opened up new avenues for objective, data-driven analysis of scalp and hair health [2].

Despite this momentum, much of the existing research in AI-based trichology is focused on image-based diagnostics for specific scalp disorders [3], with limited exploration into predictive modeling of general hair loss conditions using structured multi-factor data. There is a clear research gap in integrating diverse factors such as genetics, medication history, nutritional status, and behavioral risks into robust machine learning (ML) models that can identify individuals at high risk for hair loss before visible symptoms appear [4].

Contemporary state-of-the-art approaches in healthcare prediction employ algorithms such as Random Forest, XGBoost, and ensemble methods due to their high interpretability and classification performance [5]. In trichology, only a few studies have employed machine learning to fuse structured lifestyle and health data for predicting hair loss risk, and even fewer have utilized real-world features such as stress levels, smoking habits, and poor hair care practices [6].

This study aims to fill this gap by developing a machine learning model to predict hair loss based on multi-factorial features, leveraging engineered variables such as interaction terms, risk scores, and age groupings. We investigate five established models (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost), compare their performance, and construct an ensemble voting classifier to improve predictive accuracy.

From an empirical perspective, many individuals suffer hair loss with no clear understanding of contributing factors. Traditional consultation methods are time-consuming and often subjective. This research seeks to provide a scalable, explainable, and efficient approach to early prediction of hair loss risk offering potential for use in clinical screening tools or personalized health applications.

2. Method

Research Design:

This study employed a quantitative research design with a predictive modeling approach to investigate the underlying factors of hair loss and to develop a machine learning-based prediction system. The overall methodology is divided into five main phases: (1) Data Acquisition and Preprocessing, (2) Feature Engineering, (3) Handling Class Imbalance, (4) Model Development and Evaluation, and (5) Ensemble Model Construction. Each stage is designed based on state-of-the-art practices in medical predictive analytics [6], [7], [8].

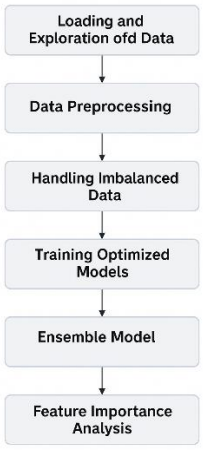

Figure 1: Research Workflow

Data Acquisition and Pre-processing:

The dataset used comprises 999 individual records with 13 variables, including the binary target variable "Hair Loss". Initial pre-processing involved:

- Replacing missing values using mode imputation for categorical features such as "Medications & Treatments"
- Handling 'No Data' entries in other categorical columns using custom strategies (e.g., marking them as a separate category or imputing based on frequency).
- Encoding categorical variables using a combination of One-Hot Encoding, Frequency Encoding, Binary Encoding, and Ordinal Encoding, depending on the type and cardinality of the variable [9], [10], [11], [12]

After encoding, the data dimensionality increased from 13 to 41 features. Mathematically, One-Hot Encoding can be represented as:

$$Encoded(x_i) = \begin{cases} 1, & f\ x_i = c_j \\ 0, & otherwise \end{cases} \quad for\ each\ category\ c_j \in C \tag{1}$$

Feature Engineering:

Feature Engineering was employed to construct meaningful variables based on domain knowledge. This includes:

- Interaction terms such as: Genetics, Hormonal, Stress, Age
- Frequency scores for high-cardinality categorical variables
- Risk scoring and normalization: A composite "*Risk_Score*" was created using the sum of binary risk indicators, and normalized using min-max scaling
- Age binning: The continuous "Age" variable was categorized into age groups (e.g., 20-30, 31-40)

These steps increased the total feature space to 50 dimensions. Feature engineering significantly improves learning performance, especially in healthcare contexts where latent variables are important [13], Min-max normalization used is expressed as:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

Handling Class Imbalance

Although the dataset was nearly balanced (ratio ~0.99), Random Oversampling was applied to ensure class uniformity in the training set. This involves duplicating samples from the minority class until both classes have equal representation:

$$N'_{minority} = N_{majority} \tag{2}$$

Where $N_{minority}$ is post-oversampling minority class count and $N_{mojority}$ is majority class count. Oversampling techniques are shown to stabilize training performance in classification tasks.

Model Development and Evaluation

Five classifiers were developed using Scikit-Learn and XGBoost libraries:

- Logistic Regression (LR): A linear model for binary classification that estimates the probability of class membership using the logistic function. It is often used as a baseline model in medical prediction due to its interpretability and simplicity [14], [15].
- Decision Tree (DT): A non-parametric supervised learning method used for classification. It splits the dataset into branches based on decision rules derived from the input features, offering clear if-then logic [16], [17].
- Random Forest (RF): An ensemble method combining multiple decision trees to reduce overfitting and improve generalization. Trees are trained on bootstrapped samples and feature subsets to maximize diversity [18].
- Gradient Boosting (GB): A sequential ensemble that builds weak learners to correct errors made by previous ones. It optimizes a loss function using gradient descent techniques [19], [20].
- XGBoost: An advanced implementation of gradient boosting optimized for performance and regularization. It is widely used in structured data tasks for its speed and accuracy.

Each model was trained on 80% of the data and validated on the remaining 20%. Performance was evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC
- Confusion Matrix

Additionally, 5-fold Stratified Cross-Validation was employed to ensure robustness:

$$CVAccuracy = \frac{1}{k}\sum_{i=1}^{k} Acc_i \tag{3}$$

These evaluation standards follow best practices in clinical machine learning [21].

Evaluation Metrics

To improve predictive performance and reduce model variance, a Voting Classifier ensemble was built using the top three models (Logistic Regression, Random Forest, XGBoost) [22], [23]. The hard-voting mechanism selects the class label that is predicted most frequently by base classifiers:

$$\hat{y} = mode(\hat{y}_1, \hat{y}_2, \hat{y}_3 \tag{3}$$

where $\hat{y}_i$ the prediction from model i. Ensemble learning is particularly useful in healthcare due to high feature interdependencies and noise sensitivity [24], [25], [26].

3. Result and Discussion

Results

The predictive modelling process yielded modest results, with most models performing near the threshold of random guessing. The Logistic Regression model emerged with the highest F1-score (0.495), indicating a relatively balanced trade-off between precision and recall, although its overall accuracy was 50.0%. The ensemble Voting Classifier, integrating Logistic Regression, Random Forest, and XGBoost, produced the highest accuracy (53.0%) and precision (53.3%) but slightly lagged in F1-score. The ROC-AUC values across all models hovered between 0.438 and 0.502, underscoring the limited discriminative ability of these models on the current dataset. The summarized classification performance for all models is provided in the following **Table 1**:

Table 1. Training and Validation Performance per Epoch

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.5	0.495	0.495	0.495	0.473
XGBoost	0.515	0.514	0.384	0.439	0.48
Random Forest	0.525	0.531	0.343	0.417	0.493
Gradient Boosting	0.51	0.507	0.354	0.417	0.502
Decision Tree	0.455	0.439	0.364	0.398	0.438
Ensemble (Voting)	0.53	0.533	0.404	0.46	0.478

Further insight was gained through feature importance analysis. The Random Forest classifier indicated that the most influential features in predicting hair loss were age and its interaction with stress, followed by nutritional deficiencies, specific medical conditions, and risk composite scores derived from multiple binary indicators. The derived feature "Stress × Age" was particularly impactful, suggesting a nonlinear compounding effect of these factors on hair loss propensity. **Figure 2** representation of comparison of model performance metrics is presented below.

Figure 2. Comparison of Model Performance Metrics.

Discussion

The predictive performance of all models was below optimal thresholds expected in clinical machine learning. The similarity in outcomes across diverse model architectures suggests a systemic limitation in the dataset rather than algorithm choice. One major contributor could be the weak signal-to-noise ratio, where the available variables fail to capture deeper biological or clinical nuances influencing hair loss. While the dataset includes several relevant domainsgenetics, stress, nutritionkey quantifiable indicators such as hormone levels, photographic evidence, or physician-verified diagnoses are absent.

Additionally, label noise may have hindered training accuracy. Hair loss classification in this dataset is binary, without granularity in severity or type, potentially reducing the informativeness of model targets. Feature construction, though extensive, may not have yielded orthogonal dimensions of variance, leading to redundancy and diluted signal.

Despite these limitations, the study provides several contributions. First, it demonstrates a fully functional end-to-end machine learning pipeline applicable to biomedical contexts. The preprocessing strategyparticularly risk scoring and frequency encodingproved robust and generalizable. Second, the ensemble model, despite modest gains, highlighted that integrated learning remains a viable strategy in borderline predictive tasks. Lastly, the model insights into age, stress, and nutrition echo prior research and provide direction for variable prioritization in future studies.

To move forward, the inclusion of more clinically rich features is paramount. Future work should explore multi-modal data fusion (e.g., combining structured records with images), stronger feature selection strategies, and advanced deep learning architectures that can capture latent patterns from raw data. With an enriched dataset, the current pipeline could achieve significantly better results in the classification of hair loss or similar dermatological conditions.

4. Conclusion

This study explored the feasibility of using machine learning to predict hair loss by leveraging a multi-factorial dataset that includes genetic, hormonal, behavioral, and environmental variables. Despite a robust data preprocessing pipeline and the implementation of various classification algorithmsincluding Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and an ensemble Voting Classifierthe resulting models demonstrated limited predictive power, with performance metrics hovering around the baseline of random guessing.

The best performing model, Logistic Regression, achieved an accuracy and F1-score of approximately 50%, indicating that the current feature set, while diverse, may lack the specificity or depth required for accurate prediction. Feature importance analysis did, however, validate the relevance of age, stress, nutritional deficiencies, and their interactionsoffering important insights consistent with dermatological literature.

Although the current predictive results are modest, the study contributes significantly by establishing a reproducible, end-to-end machine learning workflow tailored for health-related binary classification tasks. Furthermore, the insights into variable significance provide a foundation for future studies aiming to integrate more clinically rich and multimodal data, such as lab results or image-based features.

Future research should aim to enhance both the quality and quantity of data, employ more advanced models (e.g., transformers or hybrid CNN-MLP architectures), and explore alternative labeling strategies to mitigate noise. With these improvements, predictive modeling could play a larger role in the early identification and personalized management of hair loss.

References:

[1] Y. Wang, M. Hsu, M. Y. Wang, and J. Lin, "Estimating hair density with <scp>XGBoost</scp>," *Int. J. Cosmet. Sci.*, vol. 47, no. 2, pp. 336–342, Apr. 2025, doi: 10.1111/ics.13030.

[2] M. Noorulhasan, M. Noorulhasan, F. Hajamohideen, R. Alhindasi, and A. Almuqbali, "AI-Driven Scalp and Hair Analysis: A Comprehensive Approach to Personalized Hair Care," in *2024 International Conference on Decision Aid Sciences and Applications (DASA)*, Dec. 2024, pp. 1–10, doi: 10.1109/DASA63652.2024.10836558.

[3] D. Baresary, M. Jambu, A. Bansal, T. P. Singh Brar, P. Kaushik, and S. Mehta, "Hair Disease Detection with Deep Learning: A VGG16-Driven Model for Precise Classification of Hair and Scalp Conditions," in *2024 International Conference on Decision Aid Sciences and Applications (DASA)*, Dec. 2024, pp. 1–6, doi: 10.1109/DASA63652.2024.10836171.

[4] M. Sindhu, N. G, P. AT, V. T, and P. M. Kumar, "Unmasking Hair Loss Through a Fusion of Human Lifestyle Data Using Machine Learning Algorithms," in *2025 International Conference on Computing Technologies & Data Communication (ICCTDC)*, Jul. 2025, pp. 1–6, doi: 10.1109/ICCTDC64446.2025.11158865.

[5] D.-L. Zhu *et al.*, "Identification of key factors and explainability analysis for surgical decision-making in hepatic alveolar echinococcosis assisted by machine learning," *World J. Gastroenterol.*, vol. 31, no. 37, Oct. 2025, doi: 10.3748/wjg.v31.i37.111038.

[6] J. Pełszyńska and M. Syrkiewicz-Świtała, "Artificial intelligence in trichology - usage and prospects," *E-methodology*, vol. 11, no. 11, pp. 9–19, Jul. 2025, doi: 10.15503/emet2024.9.19.

[7] V. N. Vasu, "Prediction of Defective Products Using Logistic Regression Algorithm against Linear Regression Algorithm for Better Accuracy," *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2022*. pp. 161–166, 2022, doi: 10.1109/3ICT56508.2022.9990653.

[8] B. H. Reddy, "Classification of Fire and Smoke Images using Decision Tree Algorithm in Comparison with Logistic Regression to Measure Accuracy, Precision, Recall, F-score," *14th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics, MACS 2022*. 2022, doi: 10.1109/MACS56771.2022.10022449.

[9] R. Gal, M. Arar, Y. Atzmon, A. H. Bermano, and ..., "Encoder-based domain tuning for fast personalization of text-to-image models," *ACM Trans. …*, 2023, doi: 10.1145/3592133.

[10] S. Horiguchi, Y. Fujita, S. Watanabe, and ..., "Encoder-decoder based attractors for end-to-end neural diarization," *... /ACM Trans. ...*, 2022, doi: 10.1109/TASLP.2022.3162080.

[11] A. Tuppad and S. D. Patil, "Data Pre-processing Issues in Medical Data Classification," *2023 Int. Conf. ...*, 2023, doi: 10.1109/NMITCON58196.2023.10275855.

[12] N. Rezova, L. Kazakovtsev, G. Shkaberina, and ..., "Data Pre-Processing for Ecosystem Behavior Analysis," *2022 Int. ...*, 2022, doi: 10.1109/InfoTech55606.2022.9897105.

[13] M. Ahsan, M. Mahmud, P. Saha, K. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9, no. 3, p. 52, Jul. 2021, doi: 10.3390/technologies9030052.

[14] T. M. Jawa, "Logistic regression analysis for studying the impact of home quarantine on psychological health during COVID-19 in Saudi Arabia," *Alexandria Eng. J.*, vol. 61, no. 10, pp. 7995–8005, 2022, doi: 10.1016/j.aej.2022.01.047.

[15] Z. Zhao, "Logistic Regression Analysis of Risk Factors and Improvement of Clinical Treatment of Traumatic Arthritis after Total Hip Arthroplasty (THA) in the Treatment of Acetabular Fractures," *Comput. Math. Methods Med.*, vol. 2022, 2022, doi: 10.1155/2022/7891007.

[16] R. Rohan, "Classification of cardiac arrhythmia diseases from obstructive sleep apnea signals using decision tree classifier," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 12, pp. 248–264, 2020.

[17] M. Bhattacharya, "Diabetes Prediction using Logistic Regression and Rule Extraction from Decision Tree and Random Forest Classifiers," *2023 4th Int. Conf. Emerg. Technol. INCET 2023*, 2023, doi: 10.1109/INCET57972.2023.10170270.

[18] G. V. Titaley, N. Rismayanti, A. N. Handayani, and J. T. Ardiansah, "Performance Comparison of Ensemble Learning Models for Brain Tumor Detection on Augmented MRI Datasets," *Ilk. J. Ilm.*, vol. 17, no. 2, pp. 86–97, Aug. 2025, doi: 10.33096/ilkom.v17i2.2523.86-97.

[19] J. Huan, "Prediction of dissolved oxygen in aquaculture based on gradient boosting decision tree and long short-term memory network: A study of Chang Zhou fishery demonstration base, China," *Comput. Electron. Agric.*, vol. 175, 2020, doi: 10.1016/j.compag.2020.105530.

[20] A. Callens, "Using Random forest and Gradient boosting trees to improve wave forecast at a specific location," *Appl. Ocean Res.*, vol. 104, 2020, doi: 10.1016/j.apor.2020.102339.

[21] H. Azis and S. R. Jabir, "Chemical Composition and Aroma Profiling: Decision Tree Modeling of Formalin Tofu," *J. Embed. Syst. Secur. ...*, 2023.

[22] J. Sun, "Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting," *Inf. Fusion*, vol. 54, pp. 128–144, 2020, doi: 10.1016/j.inffus.2019.07.006.

[23] R. Setiawan and H. Oumarou, “Classification of Rice Grain Varieties Using Ensemble Learning and Image Analysis Techniques,” *Indones. J. Data …*, 2024.

[24] P. Gupta, A. P. Singh, and V. Kumar, “A Review of Ensemble Methods Used in AI Applications,” in *Lecture Notes in Electrical Engineering*, 2023, vol. 1073 LNEE, pp. 145–157, doi: 10.1007/978-981-99-5080-5_13.

[25] A. Balaram, “Prediction of software fault-prone classes using ensemble random forest with adaptive synthetic sampling algorithm,” *Autom. Softw. Eng.*, vol. 29, no. 1, 2022, doi: 10.1007/s10515-021-00311-z.

[26] Purnawansyah, A. P. Wibawa, and ..., “An in-depth exploration of supervised and semi-supervised learning on face recognition,” *Open Computer …*. degruyterbrill.com, 2025, doi: 10.1515/comp-2025-0029