International Journal of Technology in Education and Science

www.ijtes.net

An Explainable AI-Based Decision Support System for Teaching and Classifying Hair Loss Types

Trisnani Widowati [1] *, Ade Novi Nurul Ihsani [2], Anik Maghfiroh [3] , Clarita Aprilliani [4] , Septian Eko Prasetyo [5]

[1] Beauty Education Study Program, Faculty of Engineering, Universitas Negeri Semarang, Indonesia, 0000-0001-6156-0245
[2] Beauty Education Study Program, Faculty of Engineering, Universitas Negeri Semarang, Indonesia, 0000-0002-3283-8524
[3] Beauty Education Study Program, Faculty of Engineering, Universitas Negeri Semarang, Indonesia, 0000-0003-3940-1488
[4] Beauty professional, Indonesia, 0009-0002-1486-9842
[5] Department of Electrical Engineering, Faculty of Engineering, Universitas Negeri Semarang, Indonesia, 0009-0003-5342-8510
* Corresponding author: Trisnani Widowati (niwid@mail.unnes.ac.id)

Article Info

Article History

Received:
27 February 2026

Revised:
17 April 2026

Accepted:
3 May 2026

Published:
11 May 2026

Keywords

Decision support system
Explainable artificial intelligence
Hair loss classification
Leakage-resistant pipeline
Multi-class classification

Abstract

Hair loss is a multifactorial condition that requires accurate classification to support reliable and personalized decision-making. However, many existing machine learning approaches suffer from data leakage and limited interpretability, reducing their robustness and practical applicability in decision support systems. This study proposes a leakage-resistant machine learning framework for multi-class hair loss classification, integrating explainable artificial intelligence to enhance transparency and reliability. The framework employs a unified preprocessing pipeline within nested cross-validation to prevent information leakage, while SMOTEENN is used to address class imbalance. Several algorithms, including Decision Tree, Random Forest, K-Nearest Neighbors, Logistic Regression, and Extreme Gradient Boosting, are evaluated within this pipeline. Experimental results indicate that Extreme Gradient Boosting achieves the best performance, with an accuracy of 0.8300, F1-score of 0.7908, and AUC of 0.9305 in nested cross-validation. Evaluation on a holdout dataset demonstrates stable generalization, achieving an accuracy of 0.8250, F1-score of 0.7517, and AUC of 0.9182. Furthermore, the integration of explainable artificial intelligence enables interpretable predictions that can be utilized in a rule-based decision support system. Overall, the proposed framework provides a robust, transparent, and leakage-resistant solution for reliable machine learning-based classification in practical applications.

Citation: Widowati, T., Ihsani, A. N. N., Maghfiroh, A., Aprilliani, & C. Prasetyo, S. E. (2026). An explainable AI-based decision support system for teaching and classifying hair loss types. *International Journal of Technology in Education and Science (IJTES), 10*(3), 641-663. https://doi.org/10.46328/ijtes.8226

ISSN: 2651-5369 / © International Journal of Technology in Education and Science (IJTES).

Introduction

Hair loss is a prevalent dermatological condition affecting individuals across various age groups and genders, often leading to psychological distress and decreased quality of life (Aukerman & Jafferany, 2023; Maloh et al., 2023). Clinically, hair loss may arise from multiple underlying factors, including genetic predisposition, hormonal imbalance, autoimmune disorders, nutritional deficiencies, and environmental influences (Fatani et al., 2023). These conditions manifest in various forms such as androgenetic alopecia, alopecia areata, telogen effluvium, and other scalp-related disorders, each requiring different diagnostic and treatment approaches (Parikh et al., 2024). Therefore, accurate classification of hair loss types is essential to support effective diagnosis and personalized intervention strategies.

Several studies have investigated computational approaches for hair loss classification using structured clinical datasets. Hair loss classification has been widely addressed through data-driven methods that utilize patient records, clinical attributes, and demographic information. Previous research has applied machine learning techniques such as decision trees, support vector machines, and ensemble learning methods to identify conditions including androgenetic alopecia, alopecia areata, and telogen effluvium (Arif Hidayat & Sutedi, 2025; Kapoor & Mishra, 2018; Siami & Azis, 2025; Sirish Kumar et al., 2025; Yoraeni & Rakhmah, 2025). However, most existing studies are limited to binary classification or a small number of disease categories, and often rely on either handcrafted feature engineering or conventional machine learning models without systematically addressing critical machine learning pipeline issues such as data leakage, class imbalance, and model interpretability. These limitations highlight the need for more comprehensive and reliable frameworks for multi-class hair loss classification based on structured clinical data.

In recent years, the rapid advancement of artificial intelligence, particularly machine learning and deep learning, has significantly transformed the landscape of medical diagnostics (Jeong et al., 2023). Automated classification systems have been widely developed to assist clinicians in identifying dermatological conditions using both image-based and structured clinical datasets (Aboulmira et al., 2024). In the field of medical imaging, convolutional neural networks (CNNs) have become a dominant approach due to their strong capability in automatic feature extraction and high performance in image recognition tasks.

Recent developments show that CNN-based architectures are increasingly applied in dermatological and trichological analysis, including scalp and hair disorder classification, due to their ability to learn complex visual patterns from clinical images (Mienye et al., 2025; Teoh, 2023). In addition, transfer learning approaches using pretrained deep learning models such as ResNet and EfficientNet have been widely adopted to improve model generalization, particularly when training data are limited or imbalanced. These methods have demonstrated strong performance in dermatological image classification tasks and are commonly reported to enhance diagnostic accuracy across multiple skin disease categories in recent studies (Sarhan et al., 2025; Venkatesh et al., 2026).

In addition to deep learning, traditional machine learning approaches remain relevant, particularly for structured datasets involving clinical attributes. Algorithms such as Random Forest, Support Vector Machine, and Gradient

Boosting have shown competitive performance with lower computational complexity (Liu et al., 2020; Omar et al., 2024; Taha, 2025). For instance, comparative experiments have shown that classifiers such as Random Forest, Support Vector Machine (SVM), and ensemble boosting methods can achieve competitive performance in dermatological diagnosis tasks, especially when combined with effective feature selection strategies (Almustafa, 2025). Similarly, SVM-based and tree-based ensemble models are frequently reported as strong baseline approaches due to their robustness and relatively low computational cost compared to deep neural networks (Akilandasowmya et al., 2024). In parallel, imbalanced business classification using multiple machine learning models and statistical robustness tests, finding that the F1-score provides the most stable and balanced evaluation across conditions, while MCC offers complementary insights and accuracy/precision show limited robustness (Sujon et al., 2025).

Despite these advancements, several critical challenges persist in the current body of research. First, several studies highlight that data leakage remains a critical issue in machine learning pipelines, particularly when preprocessing steps such as normalization, feature selection, and resampling are applied prior to train-test splitting. This practice allows information from the test set to influence the training process, leading to overly optimistic performance estimates and poor generalization in real-world applications (Apicella et al., 2025; Ichwani et al., 2026; Sasse et al., 2025). Second, class imbalance remains a prevalent issue in medical datasets, particularly in dermatological and hair loss classification tasks, where minority classes are underrepresented and may lead to biased model predictions. Although resampling techniques such as SMOTE and its variants are widely adopted, recent studies highlight that improper integration within the machine learning pipeline can introduce synthetic noise, class overlap, and overgeneralization, ultimately reducing model reliability and generalization performance (Hairani et al., 2024; Saad Hussein et al., 2019; Salmi et al., 2024). Third, the lack of interpretability in many machine learning models remains a significant barrier to their adoption in clinical and decision support systems, where transparency, trust, and accountability are critical requirements. Recent studies emphasize that black-box models are often difficult to justify in medical contexts, leading to increased interest in explainable AI techniques such as SHAP and LIME to enhance model transparency and clinical trust (Abbas et al., 2025; Chaddad et al., 2023).

To overcome these limitations, recent research trends emphasize the integration of robust evaluation strategies and explainable artificial intelligence. Nested cross-validation has been increasingly adopted as a reliable approach to prevent overfitting and eliminate data leakage by separating model selection and evaluation processes (Wainer & Cawley, 2021). Moreover, hybrid resampling techniques such as SMOTEENN, which combine oversampling with noise filtering, have demonstrated improved classification performance in imbalanced datasets, with reported increases in F1-score of up to 5–10% compared to standalone SMOTE (Han & Joe, 2024). In parallel, explainable artificial intelligence techniques, including SHAP (SHapley Additive exPlanations), have been utilized to provide feature-level interpretability, enabling users to understand the contribution of each input variable to the model's prediction (Nohara et al., 2022). Recent studies have shown that integrating SHAP with machine learning models significantly enhances user trust and supports decision-making processes in healthcare applications (Agrawal et al., 2025; Hur et al., 2025).

However, existing studies on hair loss classification typically address challenges such as class imbalance, data leakage, and model interpretability in isolation. Most approaches either focus on improving predictive performance through resampling techniques or enhancing transparency using explainable AI, without embedding these components into a unified and leakage-resistant learning pipeline. Consequently, a systematic framework that integrates strict data separation, imbalance-aware learning, and post-hoc explainability within a single coherent pipeline remains underdeveloped (Apicella et al., 2025; Hairani et al., 2024; Salmi et al., 2024; Sasse et al., 2025). As a result, there remains a significant research gap in developing an integrated, leakage-resistant machine learning pipeline that simultaneously ensures robustness, fairness, and interpretability for multi-class hair loss classification. Furthermore, the integration of machine learning outputs into actionable clinical decision support systems remains underexplored, limiting the translational impact of existing approaches.

To overcome these limitations, this study proposes a comprehensive machine learning framework for multi-class hair loss classification. The proposed pipeline consists of structured preprocessing with strict data separation, imbalance handling using SMOTEENN, model optimization through nested cross-validation, and explainable AI modules based on SHAP for interpretability. This integrated design ensures that each stage of the machine learning process is systematically optimized while preventing information leakage, mitigating class imbalance bias, and enhancing model transparency.

Therefore, the objective of this study is to develop a leakage-resistant and explainable machine learning framework for multi-class hair loss classification to support reliable clinical decision-making. The proposed method integrates robust preprocessing, hybrid resampling, nested cross-validation, and SHAP-based interpretability within a unified pipeline. The expected outcome is a generalizable, transparent, and high-performing classification system capable of supporting intelligent decision-making in hair loss diagnosis.

Method

The overall research workflow is illustrated in Figure 1. The proposed framework is designed as a leakage-resistant machine learning pipeline for multi-class hair loss classification, integrating preprocessing, imbalance handling, robust evaluation, and explainable AI for decision support. The process begins with data acquisition, followed by controlled preprocessing to prevent data leakage. Model development is conducted using nested cross-validation, with SMOTEENN applied only to training folds. Multiple algorithms are evaluated, and the best model is selected based on cross-validation performance, then validated on a holdout test set. Finally, explainable AI is incorporated to generate interpretable insights, which are used to support rule-based recommendations.

Data Acquisition

The dataset used in this study was obtained from Kaggle and consists of 400 longitudinal daily records with 14 attributes, including 13 input features and one target variable representing hair loss severity. The features capture multiple dimensions of daily life, including lifestyle habits, psychological factors, cognitive workload, and personal care variables, with libido included as a proxy for hormonal conditions. The target variable is categorized

into four levels: Few, Medium, Many, and A lot. Hair loss measurements are based on self-reported observations, where the number of fallen hair strands is categorized into discrete severity levels, conceptually aligned with clinical screening approaches such as the hair pull test. Although not derived from controlled clinical settings, the dataset reflects realistic daily behavioral patterns related to hair loss. Initial inspection reveals several data quality challenges, including missing values (notably in school_assessment and dandruff), class imbalance with dominance of Few and Medium classes, heterogeneous data types (numerical, nominal, and ordinal), and mild outliers in certain numerical features. These characteristics represent common issues in real-world datasets and require appropriate preprocessing strategies to ensure reliable model performance.

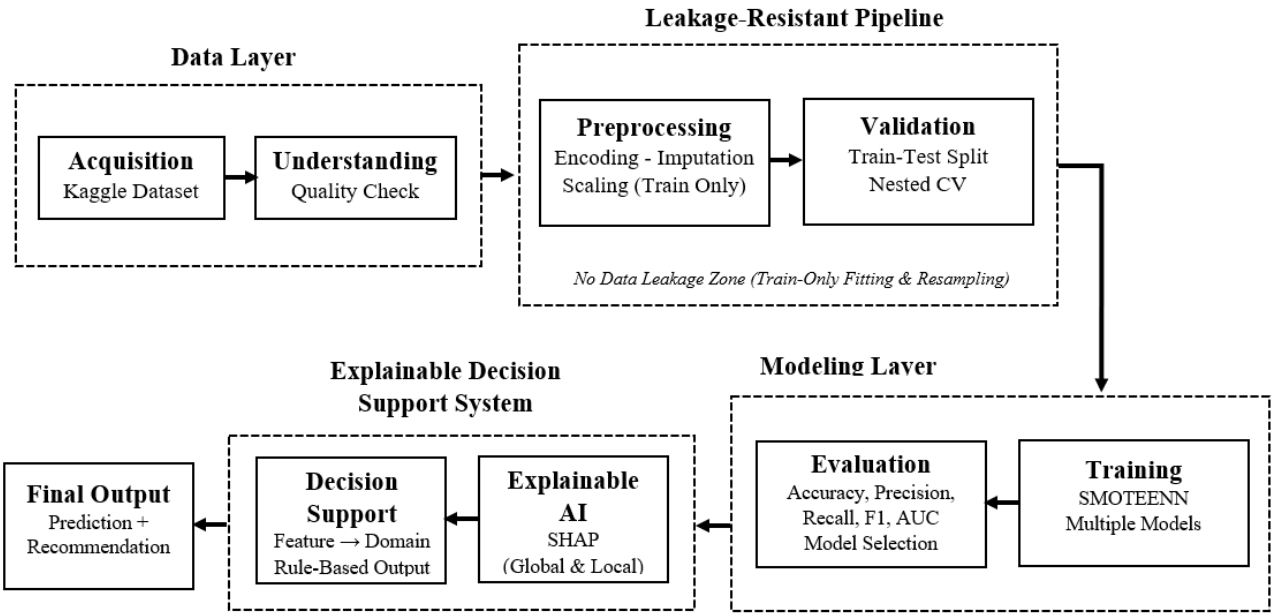

Figure 1. Research Workflow of Proposed System

Furthermore, the dataset consists of heterogeneous data types, including numerical, nominal, and ordinal variables, requiring appropriate encoding strategies during preprocessing. Several numerical features also exhibit moderate variability and mild outliers, particularly in variables such as stay_up_late, indicating the need for normalization techniques. These characteristics reflect common challenges in real-world datasets, including missing data, imbalance, and noise, which must be carefully handled to ensure robust machine learning performance (Lusito et al., 2024). To provide a structured overview, the key data quality aspects are summarized in Table 1.

The distribution of the target variable further confirms the presence of class imbalance within the dataset. Based on the observed data, the majority classes are Few (169 instances) and Medium (167 instances), which together account for approximately 84% of the total records. In contrast, the minority classes Many (42 instances) and A lot (22 instances) are significantly underrepresented, comprising only about 16% of the dataset. This skewed distribution reflects real-world conditions where severe hair loss cases occur less frequently compared to mild or moderate cases. However, such imbalance can lead to biased model learning, where the classifier tends to favor majority classes and underperform on minority classes. Therefore, this study incorporates resampling techniques,

specifically SMOTEENN, within a leakage-resistant pipeline to address this imbalance and ensure that the model can learn discriminative patterns across all hair loss severity levels.

Table 1. Data Quality Summary

Aspect	Observation
Total Records	400 data
Total Features	14 (13 input, 1 target)
Missing Values	High in *school_assessment* and *dandruff* (>70%), low in *hair_grease*
Class Distribution	Imbalanced (Few & Medium dominate; Many & A lot are minority classes)
Data Types	Mixed (numerical, ordinal, nominal)
Outliers	Mild outliers detected in numerical features (e.g., *stay_up_late*)
Data Characteristics	Real-world dataset with noise, imbalance, and incomplete values

Data Preprocessing

Data preprocessing is a critical stage designed to improve data quality while preventing data leakage. All preprocessing steps are implemented within a unified pipeline and applied exclusively to the training data in each cross-validation fold, ensuring no information from validation or test sets is introduced during model training (Sasse et al., 2025). The process includes data cleaning and transformation, where non-informative features (e.g., timestamps) are removed, and both target and categorical variables are encoded into numerical representations. All features are standardized into a consistent numerical format, with invalid entries handled as missing values to ensure data uniformity.

Handling missing values is performed using median imputation, which is robust to skewed distributions and outliers commonly found in real-world datasets (Dhanka et al., 2026). The imputation process is defined by Equation (1).

$$x_i^{imp} = \begin{cases} x_i & if\ x_i \neq NaN \\ \tilde{x} & if\ x_i = NaN \end{cases}$$

where $\tilde{x}$ denotes the median value of the corresponding feature computed solely from the training data. This approach ensures that missing values are estimated without introducing bias from unseen data.

Following imputation, feature scaling is applied using the RobustScaler technique. Unlike standard normalization methods, RobustScaler utilizes the median and interquartile range (IQR), making it more resilient to outliers(Vinutha et al., 2018). The transformation is defined by Equation.

$$x' = \frac{x - Q_2}{Q_3 - Q_1}$$

where Q_2 represents the median, and Q_1 and Q_3 denote the first and third quartiles, respectively. This scaling ensures that features with different ranges are normalized while minimizing the influence of extreme values.
To formalize the preprocessing stage, the transformation process can be represented as a composition of functions using Formula (3).

$$f_{preprocess}(X) = f_{scale}\left(f_{impute}\left(f_{encode}(X)\right)\right)$$

This formulation ensures that each transformation step is applied sequentially and consistently within the training pipeline. Importantly, all preprocessing parameters (e.g., median, quartiles, encoding mappings) are learned exclusively from the training data in each cross-validation fold. Overall, this preprocessing strategy effectively addresses key data challenges, including missing values, heterogeneous data types, and outliers, while maintaining a leakage-resistant design. This ensures that the subsequent modeling stage operates on clean, normalized, and reliable data, ultimately improving the robustness and generalization capability of the proposed framework.

Data Partitioning and Nested Cross-Validation

To ensure robust and unbiased evaluation, this study employs a stratified nested cross-validation framework combined with an independent holdout test set (Wainer & Cawley, 2021). The dataset is first split into training and testing subsets (80:20) using stratification, with the test set reserved exclusively for final evaluation. Within the training data, a nested stratified K-fold scheme is applied, consisting of an outer loop (K = 5) for performance estimation and an inner loop (K = 3) for hyperparameter tuning via grid search. This structure prevents optimistic bias by separating model selection from evaluation. Stratification is maintained in all folds to preserve class distribution, and a fixed random seed (random_state = 42) is used to ensure reproducibility. Final performance is computed as the average across outer folds.

The overall performance is computed as the average score across all outer cross-validation folds using Formula.

$$\dot{Performance} = \frac{1}{K}\sum_{i=1}^{K} Score_{outer}^{(i)}$$

where K denotes the number of folds and $Score_{outer}^{(i)}$ represents the evaluation metric obtained from the i-th outer fold.

Multiple evaluation metrics are employed, including accuracy, precision, recall, F1-score (macro-averaged), and area under the curve (AUC) using a one-vs-rest strategy for multi-class classification. Importantly, all preprocessing and resampling operations, including imputation, scaling, and class imbalance handling, are embedded within the machine learning pipeline and executed independently within each fold of the nested cross-validation process. This ensures strict separation between training and validation data, effectively preventing data leakage.

After nested cross-validation, the best-performing model configuration is selected based on the average outer-fold performance and retrained on the full training dataset. The final model is then evaluated on the independent holdout test set to provide an unbiased estimate of generalization performance on unseen data. This two-stage evaluation strategy ensures both reliable model selection and realistic performance assessment.

Imbalance Handling using SMOTEENN

Class imbalance represents a critical challenge in multi-class classification problems, particularly in real-world datasets where minority classes are often underrepresented (Han & Joe, 2024; Saad Hussein et al., 2019). In this

study, the imbalance observed in hair loss severity levels is addressed using a hybrid resampling technique, namely SMOTEENN (Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors). This method integrates both over-sampling and under-sampling strategies to simultaneously enhance minority class representation and reduce noise in the dataset.

The SMOTE component operates by generating synthetic samples for minority classes through interpolation between a given instance and its nearest neighbors in the feature space. Formally, a synthetic sample can be expressed as shown in Equation.

$$x_{new}^{\cdot} = x_i + \lambda(x_{nn} - x_i), \ \lambda \in [0,1]$$

where x_i is a minority class instance and x_{nn} is one of its nearest neighbors. This mechanism enables the model to learn more generalized decision boundaries by expanding the minority class region rather than simply duplicating existing samples.

Following the over-sampling phase, the Edited Nearest Neighbors (ENN) method is applied as a noise filtering mechanism. ENN removes samples whose class labels differ from the majority class among their k-nearest neighbors, thereby eliminating borderline and potentially mislabeled instances. This cleaning process improves class separability and reduces the risk of overfitting caused by noisy data points (Han & Joe, 2024).

The combined SMOTEENN approach can be interpreted as a two-stage transformation, as formally expressed in Equation.

$$D' = f_{ENN}\left(f_{SMOTE}(D)\right)$$

where D is the original dataset and D' is the resampled dataset. This sequential process ensures that synthetic minority samples are first generated and then refined through noise reduction.

A key contribution of this study lies in the strict integration of SMOTEENN within the machine learning pipeline, where resampling is applied only to training data in each fold of stratified cross-validation. This design prevents data leakage and avoids overly optimistic performance estimates. Beyond balancing class distribution, SMOTEENN also improves data quality by removing noisy instances, leading to better generalization. As a result, the model achieves more reliable performance, particularly in macro-averaged metrics such as precision, recall, and F1-score, which are sensitive to minority classes.

Machine Learning Models

To evaluate the effectiveness of the proposed framework for multi-class hair loss classification, a diverse set of machine learning algorithms is employed. The selected models represent different learning paradigms, including tree-based methods, ensemble learning, linear models, and instance-based approaches, enabling a comprehensive comparative analysis. The models considered in this study include Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Logistic Regression (LR), and K-Nearest Neighbors (KNN).

Decision Tree is utilized as a baseline interpretable model, while Random Forest is included as an ensemble

method to improve predictive performance through variance reduction. XGBoost is incorporated as an advanced boosting technique with built-in regularization, known for its strong performance on structured data and its ability to capture complex patterns. Logistic Regression serves as a linear benchmark model, whereas K-Nearest Neighbors represents a non-parametric approach based on distance metrics (Bartz-Beielstein & Zaefferer, 2023; Pensa et al., 2025). This combination of models enables a balanced evaluation across different learning characteristics and supports a robust comparison of classification performance.

All models are implemented within a unified pipeline to ensure consistent preprocessing and experimental conditions. The imbalance handling strategy described in Imbalance Handling using SMOTEENN Section is applied uniformly across all models, while model evaluation follows the validation protocol outlined in Data Partitioning and Nested Cross-Validation Section. This design ensures that performance comparisons are fair, reproducible, and not influenced by inconsistencies in data processing. The use of multiple models allows for a comprehensive assessment of different learning mechanisms in capturing complex relationships within the dataset. This comparative approach facilitates the identification of the most suitable model for integration into the proposed decision support framework.

Explainable Artificial Intelligence

To enhance the interpretability and transparency of the proposed machine learning framework, this study integrates explainable artificial intelligence (XAI) techniques using SHAP (SHapley Additive exPlanations). SHAP is a theoretically grounded approach based on cooperative game theory that assigns each feature an importance value representing its contribution to the model's prediction (Li et al., 2024). The SHAP value for a feature is defined by the following expression in Equation.

$$\dot{\phi}_{\iota} = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

where ϕ_i represents the contribution of feature i, F denotes the set of all features and and $f(\cdot)$ is the prediction function of the model. This formulation ensures a fair attribution of feature contributions by considering all possible feature subsets.

SHAP is applied to the best-performing model to provide both global and local interpretability, where global insights identify the most influential features across all samples and local explanations reveal how individual features contribute to specific predictions. To ensure efficiency, TreeSHAP is employed for exact and scalable computation, particularly for tree-based models. Overall, this integration enhances model transparency while uncovering key factors influencing hair loss classification, supporting more explainable and informed decision-making.

Decision Support Integration

To extend the framework beyond predictive modeling, this study integrates a rule-based decision support system (DSS) that transforms model outputs into actionable recommendations, bridging the gap between data-driven

predictions and practical decision-making. The DSS leverages two main components: the predicted class label, which indicates the severity level of hair loss, and SHAP-based feature contributions, which explain the factors influencing the prediction. By combining these outputs, the system constructs interpretable decision rules that generate domain-relevant and explainable recommendations.

Formally, let y denote the predicted class and $\phi = \{\phi_1, \phi_2, \dots, \phi_3\}$ represent the SHAP values for all features. A subset of dominant features $F^* \subseteq F$ is identified based on a predefined importance threshold τ, as defined in Equation.

$$F^* = \{ f_i \in F \mid \mid \phi_i \mid > \tau \}$$

Dominant features are mapped into a knowledge-driven rule base, where each feature corresponds to a specific intervention strategy. This rule base is developed through a hybrid approach that combines domain knowledge from the literature, empirical evidence from the dataset, and insights derived from SHAP analysis. From a domain perspective, key factors such as genetic predisposition, hormonal imbalance, stress, sleep patterns, and lifestyle are incorporated based on established findings in dermatological research. These are further reinforced by SHAP-based feature importance, ensuring that the rules are not only theoretically grounded but also aligned with actual model behavior, resulting in more adaptive and data-driven recommendations.

The decision-making process is represented by the mapping function given in Equation.

$$R = g(y, F^*)$$

where R represents the set of recommended actions. This formulation ensures that recommendations are not solely dependent on the predicted class, but also on the contributing factors that drive the prediction.

The rule-based system groups recommendations into domains such as medical consultation, lifestyle modification, and hair care. Clinical actions are prioritized when genetic and hormonal factors dominate, while behavioral interventions are emphasized for lifestyle-related factors like stress and sleep. This integration enhances interpretability, practical relevance, and trust by linking predictions to clear causes and actionable recommendations, enabling a complete pipeline from prediction to decision support.

$$S_M = \sum\nolimits_{f_i \in f^*,\ \delta(f_i)=Medical} |\phi_i|$$

$$S_L = \sum\nolimits_{f_i \in f^*,\ \delta(f_i)=Li} |\phi_i|$$

$$S_H = \sum\nolimits_{f_i \in f^*,\ \delta(f_i)=HairCare} |\phi_i|$$

To operationalize the decision support mechanism, the selected dominant feature set F^* is further grouped into three predefined domain categories: medical factors, lifestyle factors, and hair care factors. Each feature $f_i \in F^*$ is assigned to a specific domain based on a predefined domain mapping function $\delta(f_i)$, which is defined according to domain knowledge and validated feature interpretation from SHAP analysis. The total contribution of each domain is computed through the following set of Equations (10–12). The dominant factor category is determined using an argmax function, as defined in Equation.

$$D = \arg\max(S_M, S_L, S_H)$$

where D represents the dominant decision domain. The final recommendation is then generated based on the

following rule-based mapping:

If $D = Medical$, then $R = Medical\ Consultation$

If $D = Lifestyle$, then $R = Behavioral\ Consultation$

If $D = HairCare$, then $R = Hair\ Treatment\ and\ Care$

This domain aggregation mechanism ensures that the decision support system is not driven by isolated feature contributions, but instead by structured interpretability across clinically meaningful factor groups.

Feature-to-Domain Mapping Strategy

To enable a structured and interpretable decision-making process within the proposed decision support system (DSS), a feature-to-domain mapping strategy is introduced. This strategy serves as an intermediate layer that connects SHAP-based feature importance values with higher-level decision domains used in the rule-based inference engine. Given that the dataset consists of self-reported behavioral and physiological proxy indicators, the feature grouping is designed to reflect three major domains that represent different aspects influencing hair loss conditions, namely Medical (physiological-related) factors, Lifestyle factors, and Hair Care factors. Each feature is assigned to a specific domain based on domain knowledge interpretation and its functional relevance within the context of hair loss behavior analysis.

This mapping allows individual SHAP values to be aggregated at the domain level, enabling the system to identify the dominant contributing factor group for each prediction instance. The resulting dominant domain D is then used as the basis for generating personalized recommendations through a deterministic rule-based mapping mechanism. Formally, each feature f_i is assigned to a domain using a mapping function $\delta(f_i)$, where $\delta(\cdot)$ defines the relationship between input features and their corresponding conceptual categories. This structured grouping ensures that model explanations are transformed into interpretable decision components that align with human-understandable factors.

The complete feature-to-domain mapping is presented in Table 2.

Table 2. Feature-to-Domain Mapping for Decision Support System

Domain	Features	Supporting Reference
Medical Factors	libido, stress_level	(Boghosian et al., 2026; Chien Yin et al., 2021).
Lifestyle Factors	stay_up_late, coffee_consumed, brain_work_duration, school_assessment, pressure_level	(Boghosian et al., 2026; Ly et al., 2025)
Hair Care Factors	hair_grease, hair_washing, shampoo_brand, dandruff, swimming	(Borda & Wikramanayake, 2015; Gavazzoni Dias, 2015)

This structured mapping enables the aggregation of SHAP values at the domain level, allowing the decision support system to determine the dominant factor group for each prediction instance. The resulting dominant

domain is subsequently used in a rule-based inference mechanism to generate personalized recommendations in the form of Medical Consultation, Behavioral Intervention, or Hair Treatment and Care. Overall, this mapping strategy enhances the interpretability, transparency, and reproducibility of the proposed system by explicitly defining how low-level feature contributions are translated into high-level decision categories within the DSS framework.

Model Evaluation

The performance of the proposed framework is evaluated using standard classification metrics derived from the confusion matrix, including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC). These metrics provide a comprehensive assessment of model performance, particularly in the presence of class imbalance.

A confusion matrix summarizes the classification outcomes by comparing predicted labels with actual labels. It consists of four fundamental components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP represents correctly predicted positive instances, TN denotes correctly predicted negative instances, FP indicates incorrectly predicted positive instances, and FN represents incorrectly predicted negative instances.

Based on these components, the evaluation metrics are calculated using Equation (14-17).

$$\dot{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\dot{Precision} = \frac{TP}{TP + FP}$$

$$\dot{Recall} = \frac{TP}{TP + FN}$$

$$\dot{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

In this study, all metrics are computed using macro-averaging to ensure equal contribution from each class, which is particularly important in imbalanced multi-class classification scenarios. Macro-averaging calculates the metric independently for each class and then computes the unweighted mean, preventing dominance by majority classes. In addition, the Area Under the Receiver Operating Characteristic Curve (AUC) is employed to evaluate the model's discriminative capability across all classes. Given the multi-class nature of the problem, AUC is computed using the One-vs-Rest (OvR) strategy, where each class is evaluated against all others. A higher AUC value indicates better class separability and model robustness.

To ensure reliable and unbiased performance estimation, the evaluation process incorporates both nested cross-validation and a final holdout test set. Nested cross-validation provides a statistically robust estimate by separating model selection from performance evaluation, while the holdout dataset offers an independent assessment of generalization performance on unseen data. The combination of multiple evaluation and rigorous validation ensures that the reported results accurately reflect the effectiveness, robustness, and generalizability of the proposed framework.

Results

The experimental setup in this study was designed to ensure fair, robust, and leakage-resistant evaluation across all machine learning models. All experiments were conducted using a unified pipeline integrating preprocessing, resampling, and model training within a nested cross-validation framework. Hyperparameter tuning was performed using grid search in the inner loop, while model performance was evaluated in the outer loop. Additionally, a final holdout dataset was used to assess generalization performance. The evaluation metrics included accuracy, precision, recall, F1-score, and area under the curve (AUC), providing a comprehensive of classification performance.

Performance Evaluation using Nested Cross-Validation

Table 3 presents the performance comparison of all evaluated models using stratified nested cross-validation, where the reported values represent the average across folds to ensure robust generalization estimates. Among the evaluated models, Extreme Gradient Boosting (XGBoost) achieves the best overall performance, with an accuracy of 0.8300, precision of 0.7846, recall of 0.8254, F1-score of 0.7908, and AUC of 0.9305. These results indicate that XGBoost provides the most balanced trade-off across evaluation metrics, particularly in maintaining strong macro-averaged performance under class imbalance conditions.

Table 3. Performance Comparison Using Nested Cross-validation

Model	Accuracy	Precision	Recall	F1-Score	AUC
Decision Tree	0.8150	0.7784	0.8255	0.7825	0.8793
Random Forest	0.8125	0.7849	0.8159	0.7736	0.9555
KNN	0.8125	0.7581	0.8153	0.7635	0.8902
Logistic Regression	0.7975	0.7497	0.8108	0.7517	0.9225
XGBoost	0.8300	0.7846	0.8254	0.7908	0.9305

The superior performance of XGBoost can be attributed to its boosting-based ensemble mechanism, which iteratively refines errors from previous learners, enabling it to effectively capture complex and non-linear relationships in the data. This leads to a more balanced performance across evaluation metrics, particularly reflected in its highest F1-score. In comparison, Random Forest achieves the highest AUC, indicating strong class separability, but its lower F1-score suggests less balanced performance across classes. Decision Tree and KNN also demonstrate competitive results but tend to be less stable, with Decision Tree being more prone to overfitting and KNN sensitive to data distribution. Logistic Regression, as a linear model, shows the lowest performance, highlighting its limitations in modeling non-linear patterns.

Overall, the consistent performance across models is strongly influenced by the integration of a leakage-resistant preprocessing pipeline combined with SMOTEENN resampling. By ensuring that preprocessing and resampling are confined within each training fold, the framework effectively prevents data leakage and produces more reliable evaluation results. This design enhances generalization capability and improves performance on minority classes,

as reflected in the macro-averaged evaluation metrics.

Generalization Performance on Holdout Dataset

To further assess the generalization capability of the proposed framework, the best-performing model identified during cross-validation, namely XGBoost, was evaluated on an independent holdout dataset that was strictly excluded from the model development process. The evaluation results are summarized in Table 4.

Table 4. Final evaluation on holdout dataset

Metric	Value
Accuracy	0.8250
Precision	0.7569
Recall	0.7629
F1-Score	0.7517
AUC	0.9182

The results demonstrate that the model maintains stable predictive performance on unseen data, with a moderate decrease compared to cross-validation results (F1-score: 0.7908 → 0.7517). This gap remains within a reasonable range, indicating that the evaluation during cross-validation is reliable and that the model does not suffer from significant overfitting. The consistency between these results confirms the effectiveness of the leakage-resistant pipeline, where preprocessing and resampling are strictly confined within each training fold.

From a generalization perspective, this stability suggests that the model successfully captures underlying data patterns rather than memorizing training instances, despite challenges such as class imbalance and missing values. The AUC value of 0.9182 further indicates strong discriminative capability across classes, even in a multi-class setting. However, slightly lower precision and recall compared to accuracy highlight the ongoing difficulty in handling minority classes, reinforcing the importance of macro-averaged metrics for balanced evaluation.

Overall, the holdout evaluation confirms that the proposed framework achieves a good balance between predictive performance and generalization capability, supporting its applicability in real-world hair loss classification and decision support systems.

Impact of Leakage-Resistant Pipeline

A central contribution of this study is the implementation of a leakage-resistant machine learning pipeline, which fundamentally alters how data preprocessing and model evaluation are conducted. In conventional machine learning workflows, preprocessing steps such as imputation, encoding, and scaling are often performed prior to data partitioning. While computationally convenient, such practices introduce a critical methodological flaw known as data leakage, where information from validation or test data inadvertently influences the training process. This results in overly optimistic performance estimates and undermines the validity of the model.

In contrast, the proposed framework strictly encapsulates all preprocessing operations within a unified pipeline that is executed independently within each training fold of cross-validation. Consequently, all transformation parameters—including imputation statistics (e.g., median values), scaling factors (e.g., interquartile range), and encoding mappings—are learned exclusively from the training subset and subsequently applied to the corresponding validation subset. This design enforces a strict separation between training and evaluation phases, ensuring that no information from unseen data is incorporated during model learning.

The impact of this design is evident in the consistency observed between cross-validation and holdout evaluation results. Specifically, the relatively small performance gap between the cross-validated F1-score (0.7908) and the holdout F1-score (0.7517) indicates that the model maintains stable predictive behavior across different data partitions. From a statistical perspective, this suggests that the model is not overfitting to specific folds or exploiting unintended data correlations introduced during preprocessing.

Furthermore, the leakage-resistant design enhances the credibility and reproducibility of the experimental results. By ensuring that each fold operates under an identical and isolated preprocessing regime, the evaluation process more accurately reflects real-world deployment scenarios, where future data are entirely unseen during model training. This is particularly important in healthcare-related applications such as hair loss analysis, where unreliable performance estimates could lead to misleading recommendations.

It is also worth noting that the robustness of the proposed pipeline becomes increasingly critical in the presence of dataset challenges, including class imbalance, missing values, and heterogeneous feature types. In such conditions, improper preprocessing can amplify bias and artificially inflate performance metrics. The integration of preprocessing, resampling (SMOTEENN), and model training within a single pipeline mitigates these risks by maintaining strict data boundaries throughout the learning process. Overall, the leakage-resistant pipeline not only improves the methodological rigor of the study but also ensures that the reported performance metrics are realistic, unbiased, and generalizable. This design represents a crucial step toward reliable machine learning deployment, particularly in decision support systems where trust, transparency, and robustness are essential.

Explainable Artificial Intelligence Analysis

To enhance interpretability, SHAP (SHapley Additive exPlanations) was employed to quantify feature contributions at both global and local levels, providing a principled framework for decomposing model predictions into individual feature attributions. At the global level, the results indicate that genetic factors and hormonal imbalance are the most influential predictors, consistently associated with higher severity classes (“Many” and “A lot”), suggesting that the model captures clinically relevant patterns aligned with biologically driven hair loss mechanisms. In contrast, lifestyle-related variables such as stress level and sleep behavior exhibit moderate and more heterogeneous effects, while nutritional factors show context-dependent contributions across samples, reflecting the multifactorial nature of hair loss progression.

At the local level, SHAP enables instance-specific explanations that support fine-grained interpretation of

individual predictions, particularly in cases where similar clinical profiles result in different outcomes due to feature interactions. This capability enhances the transparency of the model by identifying the dominant factors influencing each prediction, thereby supporting personalized decision-making. Overall, the alignment between SHAP-derived feature importance and established dermatological knowledge strengthens the credibility of the proposed model, demonstrating that it is not only statistically robust but also clinically interpretable. Consequently, SHAP acts as a bridge between the black-box nature of machine learning models and domain-explainable reasoning, supporting trustworthy and transparent decision-making within the proposed clinical decision support framework.

Integration with Decision Support System

The integration of the proposed machine learning model with a rule-based decision support system (DSS) enables the transformation of predictive outputs and SHAP-based explanations into structured and interpretable personalized recommendations. Unlike conventional classification systems that rely solely on label prediction, the proposed DSS incorporates explainability outputs to derive decision logic that is both deterministic and domain-aligned. Based on the feature-to-domain mapping strategy defined in the Method section, SHAP values are aggregated into three primary domains, namely medical (physiological-related) factors, lifestyle factors, and hair care factors. This aggregation process ensures that feature-level contributions are systematically translated into higher-level decision representations that reflect the dominant explanatory patterns for each prediction instance.

The rule-based decision engine assigns each case to the dominant domain by selecting the highest aggregated SHAP contribution among the predefined categories. When medical-related factors dominate, the system generates a recommendation for medical consultation as an indication of the need for professional health evaluation. When lifestyle factors exhibit the highest contribution, the system produces behavioral intervention recommendations focused on improving stress management, sleep patterns, and daily habits. Meanwhile, when hair care factors are dominant, the system recommends hair treatment and care interventions, including scalp maintenance and optimization of hair care routines. This deterministic mapping ensures that the decision generation process is consistent, transparent, and fully traceable from model explanation to final recommendation.

The final output of the DSS represents an integrated decision that combines the predicted class label with the dominant explanatory domain, ensuring alignment between classification outcomes and explanation-driven reasoning. The results indicate that the proposed system maintains consistent mapping behavior, where identical SHAP-based explanatory patterns produce identical recommendation categories. This enhances the interpretability, reliability, and practical usability of the system, making it suitable for personalized decision support applications in hair loss analysis.

Overall, the integration demonstrates that the combination of machine learning, SHAP-based explainability, and rule-based reasoning forms a coherent and actionable decision support framework.

Comparison with Previous Studies

Table 5 presents a comparative analysis between the proposed framework and recent studies in machine learning-based classification systems using structured tabular data. The comparison considers dataset size, model type, evaluation strategy, and predictive performance in terms of accuracy and F1-score. Unlike several prior studies that primarily rely on standard evaluation pipelines, the proposed framework integrates a leakage-resistant evaluation strategy, explainable artificial intelligence, and a rule-based decision support layer, aiming to improve both methodological rigor and practical interpretability.

Table 5. Comparison with Previous Studies

Study	Dataset	Method	Best Accuracy	Best F1-Score
(Siami & Azis, 2025)	999 Record	ML Algorithm	50%	0.495
(Yoraeni & Rakhmah, 2025)	199 Record	Naïve Bayes	55%	0.553
(Arif Hidayat & Sutedi, 2025)	999 Record	Naïve Bayes	76%	0.78
(Kapoor & Mishra, 2018)	100 Record	Regression	91%	-
(Sirish Kumar et al., 2025)	2000 Record	ML Algorithm	100%	-
Proposed Method	400 Record	ML + XAI	83%	0.79

Although several prior studies report higher accuracy values, including results exceeding 90% and even reaching 100%, such figures should be interpreted cautiously due to methodological and experimental differences across studies. In particular, extremely high performance is often associated with limited dataset complexity, simplified classification boundaries, or evaluation settings that may not adequately address data leakage risks and generalization constraints. Therefore, direct comparison of accuracy across heterogeneous experimental designs is not strictly valid.

In contrast, the proposed framework prioritizes robust and unbiased performance estimation rather than optimistic accuracy reporting. The use of nested cross-validation combined with SMOTEENN ensures that model evaluation is conducted under leakage-resistant and class-imbalance-aware conditions, leading to more realistic generalization performance. Furthermore, the integration of SHAP-based explainability and a rule-based decision support layer positions the framework beyond conventional classification approaches, emphasizing interpretability and actionable decision-making in addition to predictive capability.

Discussion

The experimental results demonstrate that the proposed framework achieves a balanced trade-off between predictive performance, robustness, and interpretability for structured tabular-based hair loss classification. While prior studies, particularly those based on deep learning and image-based datasets, may report higher accuracy, such results are not directly comparable due to fundamental differences in data modality, feature representation, and problem formulation. The proposed framework is explicitly designed for tabular clinical-like data, where

interpretability and decision traceability are prioritized over marginal gains in predictive performance.

When compared with conventional machine learning approaches such as Naïve Bayes and ensemble-based models, the proposed framework demonstrates competitive predictive performance while offering additional methodological advantages. The adoption of nested cross-validation reduces the risk of overfitting and data leakage, while SMOTEENN improves robustness under imbalanced class distributions. These design choices contribute to more stable and reliable generalization behavior, particularly in scenarios where data distribution is heterogeneous and limited.

Beyond performance, the integration of SHAP-based explainability provides clinically meaningful insight into feature contributions at both global and local levels. The results indicate that physiological and hormonal-related factors are consistently among the most influential predictors, while lifestyle and hair-care-related variables contribute in a more context-dependent manner. This enhances transparency by allowing model predictions to be interpreted rather than treated as black-box outputs, which is critical in healthcare-oriented decision support systems.

Furthermore, the proposed rule-based decision support system extends the utility of the model by translating probabilistic predictions into structured and actionable recommendations. This represents a shift from conventional classification-focused studies toward decision-oriented machine learning systems. However, the rule-based mechanism introduces inherent limitations, as predefined decision logic may not fully capture complex nonlinear interactions that could be learned through adaptive or end-to-end learning-based optimization strategies. Despite these limitations, the proposed framework offers a more realistic and deployable solution for healthcare decision support applications, where interpretability, reliability, and accountability are essential requirements. Future work should explore hybrid approaches that integrate rule-based reasoning with adaptive learning mechanisms, as well as validation on larger, more diverse, and potentially multimodal datasets to further enhance generalizability and clinical applicability.

Conclusion

This study proposes a leakage-resistant machine learning-based framework for multi-class hair loss classification, integrated with explainable artificial intelligence and a decision support mechanism. The framework is implemented as a unified pipeline that incorporates preprocessing, imbalance handling using SMOTEENN, and nested cross-validation to ensure robust and unbiased evaluation. Experimental results show that XGBoost achieves the best performance, with an accuracy of 0.8300 and AUC of 0.9305 in nested cross-validation, while maintaining stable generalization on the holdout dataset with an accuracy of 0.8250 and AUC of 0.9182. These findings confirm the effectiveness of the proposed leakage-resistant approach in producing reliable and generalizable predictions.

Beyond predictive performance, the integration of explainable artificial intelligence provides insights into feature contributions, enhancing model transparency and interpretability. This is further extended through a rule-based

decision support layer that translates model outputs into actionable recommendations, increasing the practical relevance of the system. Compared to conventional approaches that focus primarily on accuracy, the proposed framework emphasizes robustness, interpretability, and system-level integration, which are essential for real-world deployment.

Despite these contributions, several limitations remain. The use of structured tabular data restricts the exploration of visual features that could be captured through image-based approaches. In addition, the rule-based decision support system requires further validation to ensure its effectiveness in real-world or clinical contexts. Future work may focus on integrating multimodal data, such as combining clinical attributes with scalp images, and developing more adaptive recommendation mechanisms using advanced knowledge-based or learning-based approaches. Overall, this study contributes a comprehensive and reliable framework that advances the development of interpretable and leakage-resistant machine learning systems for multi-class classification, particularly in the domain of hair loss analysis and decision support.

Statements and Declarations

Acknowledgments/Notes: The authors would like to express their sincere gratitude to Universitas Negeri Semarang (UNNES) for providing academic support and research facilities that enabled the completion of this study. The authors also acknowledge Kaggle and the dataset contributor (Luke X) for providing access to the hair loss dataset used in this research.

Supplementary Materials: The dataset used in this study is publicly available from Kaggle and can be accessed at: https://www.kaggle.com/datasets/lukexun/luke-hair-loss-dataset. This dataset, namely the Luke Hair Loss Dataset, contains structured attributes related to hair loss conditions and was utilized as the primary source for model development and evaluation.

Author Contributions: Conceptualization, T.W. and S.E.P.; methodology, T.W. and S.E.P.; software, S.E.P.; validation, T.W., A.M., and A.N.N.I.; formal analysis, S.E.P.; investigation, C.A. and A.N.N.I.; data curation, A.M. and C.A.; writing—original draft preparation, S.E.P.; writing—review and editing, T.W., A.N.N.I., and A.M.; supervision, T.W.; project administration, T.W. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Data Availability: Not applicable.

Ethics Approval: Not applicable.

Informed Consent: Not applicable.

Conflicts of Interest: Not applicable.

References

Abbas, Q., Jeong, W., & Lee, S. W. (2025). Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges. *Healthcare*, *13*(17), 2154. https://doi.org/10.3390/healthcare13172154

Aboulmira, A., Hrimech, H., & Lachgar, M. (2024). Skin Diseases Classification with Machine Learning and Deep Learning Techniques: A Systematic Review. *International Journal of Advanced Computer Science and Applications*, *15*(10). https://doi.org/10.14569/IJACSA.2024.01510118

Agrawal, R., Gupta, T., Gupta, S., Chauhan, S., Patel, P., & Hamdare, S. (2025). Fostering trust and interpretability: integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency. *Diagnostic Pathology*, *20*(1), 105. https://doi.org/10.1186/s13000-025-01686-3

Akilandasowmya, G., Nirmaladevi, G., Suganthi, SU., & Aishwariya, A. (2024). Skin cancer diagnosis: Leveraging deep hidden features and ensemble classifiers for early detection and classification. *Biomedical Signal Processing and Control*, *88*, 105306. https://doi.org/10.1016/j.bspc.2023.105306

Almustafa, K. M. (2025). Predictive modeling and optimization in dermatology: Machine learning for skin disease classification. *Computers in Biology and Medicine*, *189*, 109946. https://doi.org/10.1016/j.compbiomed.2025.109946

Apicella, A., Isgrò, F., & Prevete, R. (2025). Don't push the button! Exploring data leakage risks in machine learning and transfer learning. *Artificial Intelligence Review*, *58*(11), 339. https://doi.org/10.1007/s10462-025-11326-3

Arif Hidayat, & Sutedi, S. (2025). Klasifikasi Rambut Rontok Menggunakan Metode Naive Bayes. *SATESI: Jurnal Sains Teknologi Dan Sistem Informasi*, *5*(2), 183–189. https://doi.org/10.54259/satesi.v5i2.5612

Aukerman, E. L., & Jafferany, M. (2023). The psychological consequences of androgenetic alopecia: A systematic review. *Journal of Cosmetic Dermatology*, *22*(1), 89–95. https://doi.org/10.1111/jocd.14983

Bartz-Beielstein, T., & Zaefferer, M. (2023). Models. In *Hyperparameter Tuning for Machine and Deep Learning with R* (pp. 27–69). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-5170-1_3

Boghosian, T., Mendez, H., Sayegh, M., Rabionet, A., Beer, J., & Tosti, A. (2026). The Intersection of Sleep and Hair Loss: A Systematic Review. *Dermatology and Therapy*, *16*(2), 937–952. https://doi.org/10.1007/s13555-025-01641-6

Borda, L. J., & Wikramanayake, T. C. (2015). Seborrheic Dermatitis and Dandruff: A Comprehensive Review. *Journal of Clinical and Investigative Dermatology*, *3*(2). https://doi.org/10.13188/2373-1044.1000019

Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of Explainable AI Techniques in Healthcare. *Sensors*, *23*(2), 634. https://doi.org/10.3390/s23020634

Chien Yin, G. O., Siong-See, J. L., & Wang, E. C. E. (2021). Telogen Effluvium – a review of the science and current obstacles. *Journal of Dermatological Science*, *101*(3), 156–163. https://doi.org/10.1016/j.jdermsci.2021.01.007

Dhanka, S., Kumar, A., Maini, S., Kumar, N., Singh, J., Khan, M., Abbas, M., & Ksibi, A. (2026). Padding

interpolation, median imputation, RobustScalar, and particle swarm optimization with heterogeneous classifiers: a robust combination for effective heart disease diagnosis. *Frontiers in Medicine*, *12*. https://doi.org/10.3389/fmed.2025.1721740

Fatani, M. I. A., Alkhalifah, A., Alruwaili, A. F. S., Alharbi, A. H. S., Alharithy, R., Khardaly, A. M., Almudaiheem, H. Y., Al-Jedai, A., & Eshmawi, M. T. Y. (2023). Diagnosis and Management of Alopecia Areata: A Saudi Expert Consensus Statement (2023). *Dermatology and Therapy*, *13*(10), 2129–2151. https://doi.org/10.1007/s13555-023-00991-3

Gavazzoni Dias, M. F. (2015). Hair cosmetics: An overview. *International Journal of Trichology*, *7*(1), 2. https://doi.org/10.4103/0974-7753.153450

Hairani, H., Widiyaningtyas, T., & Dwi Prasetya, D. (2024). Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies. *JOIV : International Journal on Informatics Visualization*, *8*(3), 1310. https://doi.org/10.62527/joiv.8.3.2283

Han, Y., & Joe, I. (2024). Enhancing Machine Learning Models Through PCA, SMOTE-ENN, and Stochastic Weighted Averaging. *Applied Sciences*, *14*(21), 9772. https://doi.org/10.3390/app14219772

Hur, S., Lee, Y., Park, J., Jeon, Y. J., Cho, J. H., Cho, D., Lim, D., Hwang, W., Cha, W. C., & Yoo, J. (2025). Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour. *Npj Digital Medicine*, *8*(1), 578. https://doi.org/10.1038/s41746-025-01958-8

Ichwani, A., Kesuma, R. I., Setiawan, A., Wicaksono, I. E., & Hanifah, R. (2026). Preventing Data Leakage in Classification via Integrated Machine Learning Pipelines: Preprocessing, Feature Transformation, and Hyperparameter Tuning. *Jurnal Teknik Informatika (Jutif)*, *7*(1), 391–410. https://doi.org/10.52436/1.jutif.2026.7.1.5490

Jeong, K., Mallard, A. R., Coombe, L., & Ward, J. (2023). Artificial intelligence and prediction of cardiometabolic disease: Systematic review of model performance and potential benefits in indigenous populations. *Artificial Intelligence in Medicine*, *139*, 102534. https://doi.org/10.1016/j.artmed.2023.102534

Kapoor, I., & Mishra, A. (2018). Automated Classification Method for Early Diagnosis of Alopecia Using Machine Learning. *Procedia Computer Science*, *132*, 437–443. https://doi.org/10.1016/j.procs.2018.05.157

Li, M., Sun, H., Huang, Y., & Chen, H. (2024). Shapley value: from cooperative game to explainable artificial intelligence. *Autonomous Intelligent Systems*, *4*(1), 2. https://doi.org/10.1007/s43684-023-00060-8

Liu, S., Du, H., & Feng, M. (2020). Robust Predictive Models in Clinical Data—Random Forest and Support Vector Machines. In *Leveraging Data Science for Global Health* (pp. 219–228). Springer International Publishing. https://doi.org/10.1007/978-3-030-47994-7_13

Lusito, S., Pugnana, A., & Guidotti, R. (2024). Solving imbalanced learning with outlier detection and features reduction. *Machine Learning*, *113*(8), 5273–5330. https://doi.org/10.1007/s10994-023-06448-0

Ly, N., Paiewonsky, B., Fruechte, S., Goldfarb, N., Hordinsky, M. K., Bakker, C., Sadick, N., Arruda, S., & Farah, R. S. (2025). Caffeine Supplementation and Hair: A Systematic Review. *Journal of Drugs in Dermatology*, *24*(11), 1070–1074. https://doi.org/10.36849/JDD.8902

Maloh, J., Engel, T., Natarelli, N., Nong, Y., Zufall, A., & Sivamani, R. K. (2023). Systematic Review of Psychological Interventions for Quality of Life, Mental Health, and Hair Growth in Alopecia Areata and

Scarring Alopecia. *Journal of Clinical Medicine*, *12*(3), 964. https://doi.org/10.3390/jcm12030964

Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M., & Ilono, P. (2025). Deep Convolutional Neural Networks in Medical Image Analysis: A Review. *Information*, *16*(3), 195. https://doi.org/10.3390/info16030195

Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022). Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, *214*, 106584. https://doi.org/10.1016/j.cmpb.2021.106584

Omar, E. D., Mat, H., Abd Karim, A. Z., Sanaudi, R., Ibrahim, F., Omar, M. A., Ismail, M. Z. H., Jayaraj, V., & Goh, B. L. (2024). Comparative Analysis of Logistic Regression, Gradient Boosted Trees, SVM, and Random Forest Algorithms for Prediction of Acute Kidney Injury Requiring Dialysis After Cardiac Surgery. *International Journal of Nephrology and Renovascular Disease*, *Volume 17*, 197–204. https://doi.org/10.2147/IJNRD.S461028

Parikh, A. K., Tan, I. J., Wolfe, S. M., & Cohen, B. A. (2024). Advances in Topical Therapies for Clinically Relevant and Prevalent Forms of Alopecia. *Life*, *14*(12), 1577. https://doi.org/10.3390/life14121577

Pensa, R. G., Crombach, A., Peignier, S., & Rigotti, C. (2025). Explaining Random Forest and XGBoost with Shallow Decision Trees by Co-clustering Feature Importance. *Machine Learning*, *114*(12), 287. https://doi.org/10.1007/s10994-025-06932-9

Saad Hussein, A., Li, T., Yohannese, C. W., & Bashir, K. (2019). A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE. *International Journal of Computational Intelligence Systems*, *12*(2), 1412. https://doi.org/10.2991/ijcis.d.191114.002

Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review*, *57*(10), 273. https://doi.org/10.1007/s10462-024-10884-2

Sarhan, A. M., Ali, H. A., Yasser, S., Gobara, M., Kandil, A. A., Sherif, G., & Moustafa, E. (2025). Achieving high-accuracy skin cancer classification with deep learning optimized by ant colony algorithm. *Journal of Electrical Systems and Information Technology*, *12*(1), 49. https://doi.org/10.1186/s43067-025-00243-8

Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S. B., Götz, M., Hamdan, S., Komeyer, V., Kulkarni, A., Lahnakoski, J. M., Love, B. C., Raimondo, F., & Patil, K. R. (2025). Overview of leakage scenarios in supervised machine learning. *Journal of Big Data*, *12*(1), 135. https://doi.org/10.1186/s40537-025-01193-8

Siami, M. I., & Azis, H. (2025). Predicting Hair Loss with Machine Learning: A Multi-Factor Analysis. *International Journal of Artificial Intelligence in Medical Issues*, *3*(1), 60–68. https://doi.org/10.56705/ijaimi.v3i1.360

Sirish Kumar, M., Reddy, P. L. K., Dinesh Reddy, G., Kumar, A. S., & Nagendra, P. (2025). Predictive Modeling of Hair Fall using Random Forest Algorithms. *Proceedings of the 4th International Conference on Information Technology, Civil Innovation, Science, and Management, ICITSM 2025, 28-29 April 2025, Tiruchengode, Tamil Nadu, India, Part II*. https://doi.org/10.4108/eai.28-4-2025.2358120

Sujon, K. M., Hassan, R., Choi, K., & Samad, M. A. (2025). Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *Journal of Big Data*, *12*(1), 268. https://doi.org/10.1186/s40537-025-01313-4

Taha, K. (2025). Machine learning in biomedical and health big data: a comprehensive survey with empirical and experimental insights. *Journal of Big Data*, *12*(1), 61. https://doi.org/10.1186/s40537-025-01108-7

Teoh, T. T. (2023). *Convolutional Neural Networks for Medical Applications*. Springer Nature Singapore. https://doi.org/10.1007/978-981-19-8814-1

Venkatesh, J., Vijayalakshmi, R. A. K., Krishnan, D., & Partheeban, P. (2026). EfficientNet-based soft computing techniques for dermatological condition detection. *Discover Artificial Intelligence*. https://doi.org/10.1007/s44163-026-01147-w

Vinutha, H. P., Poornima, B., & Sagar, B. M. (2018). *Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset* (pp. 511–518). https://doi.org/10.1007/978-981-10-7563-6_53

Wainer, J., & Cawley, G. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, *182*, 115222. https://doi.org/10.1016/j.eswa.2021.115222

Yoraeni, A., & Rakhmah, S. N. (2025). Penerapan Algoritma Naive Bayes untuk Prediksi Kerontokan Rambut. *Jurnal Bumigora Information Technology (BITe)*, *7*(1), 63–70. https://doi.org/10.30812/bite.v7i1.5201