Unmasking Hair Loss Through a Fusion of Human Lifestyle Data Using Machine Learning Algorithms

M Sindhu Nisanth G

*Assistant Professor Department of Computer*

*Department of Computer Applications*

*Applications Kongu Engineering College*

*Kongu Engineering College*  Perundurai, Erode, India

Perundurai, Erode, India [nisanthg1010@gmail.com](mailto:nisanthg1010@gmail.com%20)

[sindhuitsri@gmail.com](mailto:sindhuitsri@gmail.com)

Preteka AT Varshini T Dr.P.Mahesh Kumar,

*Department of Computer Department of Computer Assistant Professor,*

*Applications Applications Department of*

*Kongu Engineering College* *Kongu Engineering College* Mathematics

Perundurai, Erode, India Perundurai, Erode, India Kongu Arts and

pretekaa@gmail.com [varshinithirunavukkarasu206@gmail.com](mailto:varshinithirunavukkarasu206@gmail.com) Science College,

Erode.

[mathsmaheshkumar@gmail.com](mailto:mathsmaheshkumar@gmail.com)

***Abstract*—** **Hair loss is a prevalent problem that is impacted by genetics, lifestyle, and medical factors. Prediction accuracy is decreased by traditional studies' frequent reliance on small datasets. To determine the main causes of hair loss, this study combines real-time data with machine learning algorithms like Random Forest, Decision Tree, XGBoost, Extra Trees, and Ensemble approaches. To improve feature selection and model performance, we use Generative Adversarial Networks (GANs) for augmentation and SMOTE for data balance. Our method facilitates early detection, increases predictive accuracy, and permits tailored therapy suggestions. This study offers practical advice that enables people to properly manage the health of their hair.**

**Keywords — Hair Loss prediction,Data Augmentation,Random Forest, Decision Tree, XGBoost, Extra Trees, Ensemble Methods, Machine Learning**.

INTRODUCTION

Hair loss is one of the prominent aspects for an ordinary person. People endure more struggles without taking care of their hair health in their daily routine. Periodic hair fall creates significant impact on down self-esteem, social criticism, feeling awkward, unappealing and aging problems faced by those who get affected. Various factors fall on hair loss issues such as stress, hormonal imbalances, poor nutrition, genetics and lifestyle habits contribute to this condition. Hair loss falls into two categories: temporary part caused by factors like stress and illness and permanent part is about resulting from genetics, medical conditions or aging. A hair loss prediction model uses machine learning techniques to analyze individual traits, lifestyle factors, and health records to estimate the likelihood and severity of hair loss. This approach helps to forecast hair loss patterns based on relevant features, enabling early detection and personalized care recommendations [3]. Early detection of hair loss is essential to minimize its impact and maintain hair health by finding future paths for growing datasets, combining various data sources, and developing understandable tools for better hair care management [7]. Few studies have been done on predicting hair loss from personal characteristics, lifestyle choices, and medical problems. The integration of diverse datasets and methodologies in the XGBoost and Random Forest approaches, to analyze complex hair density data by leveraging its robust feature-handling capabilities and strengthens the framework for data-driven solutions [3]. Techniques like AdaBoost and regression models facilitated non-invasive predictions of hair health, opening avenues for tailored hair care strategies [8]. The purpose of this study is to improve prediction accuracy by investigating model selection and evaluation using a variety of machine learning approaches. By focusing on predicting and analyzing hair loss using a dataset real-time data collected from individuals. The analysis of the 940 data set utilized various smoothing techniques to improve the accuracy of the results. Based on important contributing factors, the suggested method focuses on accurately classifying the causes of hair loss and making improvements in prediction accuracy.

II. LITERATURE REVIEW

The increase in falls brought on by distractions like cell phone use and traffic hazards has raised serious concerns among the public, especially for elderly people and soldiers in unfamiliar settings. Using a smart phone-based three-axis gyroscope, a 40,000-row dataset was acquired from Kaggle, and records fall when walking, running, standing, and sitting. For fall prediction, several machine learning techniques were evaluated, such as Bayesian Regularization and Scaled Conjugate Gradient. As a performance metric, cross entropy showed that Bayesian Regularization had the best accuracy. This technology can lessen the chance of injury and avoid falls. It provides insightful information for keeping an eye on and safeguarding those who are at risk. [1]

Alopecia Areata, Tinea Capitis, Telogen Effluvium, Scarring Alopecia, Trichotillomania, Folliculitis, Head Lice, and Psoriasis are among the hair diseases that can be categorized using the machine learning model presented in this study. With accuracy values ranging from 88.04% to 88.89%, the model exhibits encouraging precision, recall, and F1-scores across several hair disease classifications. It achieved an overall accuracy of 88.53% after being trained on a dataset of 8550 photos. The model's capacity to sustain a fair balance between precision and recall is demonstrated by the balanced F1-scores (88.02% to 88.99%). According to the study, the model's macro and weighted average F1-scores of 88.53% and 88.54%, respectively, demonstrate its high classification accuracy for hair illnesses. The model is a useful tool for automated dermatology diagnosis because of its consistent performance across all hair conditions. [2]

For both men and women, hair is frequently regarded as a sign of beauty, and many of us worry about its condition. Numerous factors can contribute to premature hair loss, which is a problem that more and more women face each year. Stress, dandruff, allergies, and heredity are the main causes of hair loss. By identifying the many contributing factors and developing a dataset with machine learning algorithms to assess prediction accuracy, this study seeks to help those who are impacted by hair loss. SVM, Logistic Regression, Naive Bayes, Decision Trees, Random Forest, K-Nearest Neighbors, and XGBoost were among the algorithms whose performance we compared. With an accuracy of 89%, the experimental findings showed that XGBoost performed better than the others. [3]

Using imaging features taken from binary hair form masks, this work suggests a machine learning (ML) framework to categorize hair types in dermoscopy pictures into four groups. The structural similarity index measure (SSIM) is used to evaluate the difference between the original and pre-processed pictures to determine the ideal kernel size for the black-hat hair removal technique. The SVM classifier performed the best, achieving 80% accuracy and 79.8% AUC. For picture filtering without causing a noticeable loss of texture in the lesion, a kernel size of 20 by 20 is advised. In a clinical setting, this study seeks to maximize the kernel size for dermatoscopy hair removal while preserving good picture quality. It also functions as an automated pre-processing phase for deep learning in the classification of skin diseases. [4]

In this paper, a machine learning-based framework for detecting face forms and selecting the best haircut or hairstyle is presented. It emphasizes how crucial face shape and hair length are when choosing a hairstyle, which has a big influence on a woman's appearance and mental health. The suggested model incorporates user input and expert information to classify face shapes and suggest haircuts using Naïve Bayes classification. After processing user-uploaded photos, the model recognizes facial landmarks and provides personalized suggestions. The hairstyle recommendation model obtained 83% accuracy, while the face shape classification model, trained on 5,000 photos using Python, achieved 91% accuracy. [5]

Because of the demands and strain of their jobs, IT workers are especially prone to stress, which is a major contributor to hair loss. KNN and other machine learning (ML) algorithms have demonstrated efficacy in predicting stress levels through the analysis of psychological and physiological variables. Early detection, individualized care, and excellent stress prediction accuracy (85%–95%) are among the benefits. Dependency on data quality, possible biases, and difficulties with big datasets are some of the limitations. Prediction accuracy can be increased by using sophisticated strategies like ensemble methods and feature engineering. This method offers practical advice for managing stress and avoiding stress-related hair loss. [6]

This study uses a large dataset that includes personal characteristics and lifestyle factors to investigate how well different machine learning algorithms predict hair health. AdaBoost, random forest, and logistic regression were tested; random forest had an outstanding accuracy of 94.6%. The results demonstrate how machine learning can be used to make accurate, non-invasive, and customized forecasts about hair health. Targeted care strategies were made possible by the identification of important elements like stress, nutrition, and genetics. The study highlights the potential of machine learning in precisely predicting hair loss and makes recommendations for future paths for growing datasets, combining various data sources, and developing understandable tools for better hair care management. [7]

This review evaluates natural phytochemicals as treatments for hair loss (alopecia) and examines the variables that impact hair health, including genetics, illness, lifestyle, and chemical exposure. According to scientific research, they can promote hair development through biological processes and lessen the need for harmful synthetic treatments. Benefits include efficacy, safety, and historical records proving its application. Limitations include the need for additional study and absence of scientific evidence that is suitable for market.[8]

To promote hair development and lessen the greasy and sticky feel of direct application, this study creates tea seed oil-loaded nanostructured lipid carriers (NLC). Tween 80, Vari soft 442, and their combination were used to develop the formulations (NLC-T, NLC-V, and NLC-C), which demonstrated consistent pH levels and stable particle sizes (130–430 nm) throughout a 90-day period. When compared to intact tea seed oil, the NLC-C formulation demonstrated greater nontoxicity and hair growth stimulating properties on human follicle dermal papilla cells. A high entrapment effectiveness of 96.26% was also shown by NLC-C. NLC-C improved serum diffusion efficiency by 81.4% while reducing the greasy feel, however NLC-T and NLC-V showed substantial cytotoxicity. The requirement for additional clinical testing to confirm long-term effects is one of the limitations. [9]

About 2% of people have alopecia areata (AA), a chronic autoimmune disease that causes bald patches and abrupt hair loss. Conventional diagnosis is based on visual inspection, which is frequently inaccurate. To improve AA diagnosis, this study suggests a machine learning-based framework that distinguishes between hair that is healthy and hair that exhibits AA symptoms. A dataset comprising more than 500 AA pictures and 1000 photographs of healthy hair was gathered and preprocessed using methods such as data augmentation, segmentation, and image enhancement. The algorithms that were compared were SVM, KNN, Random Forest, Gaussian Naive Bayes, and CNN; their respective accuracies were 85%, 78%, 88%, 80%, and 92%. By improving dermatology diagnostic precision, this method makes AA forecasts more trustworthy. [10]

This study suggests a Convolutional Neural Network (CNN) model based on Deep Learning for identifying hair disorders brought on by chemical treatments. A Kaggle dataset on hair diseases was used to train the model, while SGD and Adam were used for optimization. With an accuracy of 95% at epoch 85, the Adam optimizer performed best, whereas SGD's accuracy was 89% at epoch 50. The model's benefits include high accuracy and early hair illness identification, which may help avoid health problems and hair loss. The reliance on the dataset and optimization techniques are among the drawbacks. The model performs better than the current methods, demonstrating its potential for real-world application in the identification of hair diseases. [11]

Using a Cat Swarm-based Convolutional Neural System (CS-CNS), this study suggests an optimization-based Convolutional Neural Network (CNN) for efficient hair and scalp recognition. Adaptive Wiener Filter (AWF) is used to preprocess hair follicle pictures, Hexagonal Scale Invariant Feature Transform (H-SIFT) is used to extract features, and Cellular Automation-based Rough Set Theory (CA-RST) is used to improve segmentation. The Cat Swarm optimization technique is used in the classification phase to optimize the hair follicle status prediction. The accuracy, precision, recall, F1-score, and error rate were used to assess the model's performance. This method offers a high degree of accuracy and dependability in determining the extent of hair loss. Nevertheless, processing complexity and the requirement for high-quality datasets are drawbacks. [12]

Using a Convolutional Neural Network (CNN) model to categorize hair photos with and without scalp disorders, this study investigates the application of Deep Learning (DL) techniques for early scalp disease diagnosis. DL improves classification efficiency by doing away with the requirement for human feature extraction and data reconstruction. The model outperformed earlier research with a high accuracy of 96.63% using common performance criteria like precision and recall. The Local Interpretable Model-Agnostic Explanations (LIME) technique was used to enhance the interpretability of the model. It was found that although the high accuracy is noticeable it does not always correspond with the practicality or real-world applicability of the model. [13]

A hybrid CNN-Random Forest model for classifying hair strength is presented in this work, with an accuracy of 76.77%. The model's strong precision, recall, and F1-scores demonstrate its ability to classify hair into different strength levels. Two convolutional layers, pooling, flattening, and Random Forest classification make up the architecture. Benefits include automated, precise hair quality assessment with possible uses in hair care management, cosmetics, and healthcare. Limitations include the necessity for parameter modification and increasing the dataset to increase generalization. This model builds the groundwork for next developments and demonstrates how machine learning can completely transform hair care. [14]

The application of XGBoost for automated hair density estimation in dermatology is investigated in this work. 895 scalp images are used in the procedure; 745 are used for training, and 150 are used for testing in order to assess the accuracy of the model. With an accuracy of 89.5% on the training set and 95.3% on the test set, XGBoost fared better than previous approaches, outperforming past research. Higher accuracy and less work required in hair density analysis are among the benefits. The reliance on image quality and the requirement for a variety of datasets are among the drawbacks. This method provides a clinical hair analysis tool that is more efficient and objective. [15]

III.WORKFLOW

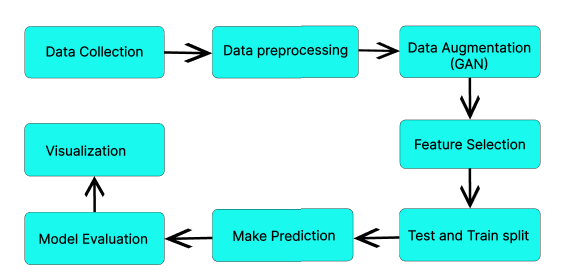


Fig.1. Workflow

*Explanation for workflow:*

*Data Collection:* Identifies most relevant features in the data to ensure meaningful analysis that entails using a Google Form to grab real-time data from individuals on the several parameters influencing hair loss. Lifestyle, surroundings, and physiological characteristics—such as age, gender, stress levels, dietary inadequacies, hair care routines, and family history of hair loss—are all stated based on situations illustrating their usual experiences and routines.

*Data Preprocessing:* One crucial stage in data preparation is to look for null values in real-world data. Since there were no missing values in this instance, the dataset is entire and prepared for additional analysis. Encoding - Categorical features were converted into numerical representations using label encoding, making it suitable for machine learning models.

*Data Augmentation:* A dataset of 1000 real data points was initially collected. To enhance dataset diversity and volume, 800 synthetic data points were generated using a Generative Adversarial Network (GAN). By integrating both real and synthetic data, this approach resulted in an expanded dataset of 1800 data points for further research. The discriminator in the GAN framework plays a crucial role by distinguishing between real and synthetic data, thereby enhancing the generator's ability to create more realistic samples that closely resemble real-world data.

*Feature Selection:* Feature selection was performed using Mutual Information Classification, which identifies the most relevant features for hair loss prediction by evaluating and ranking them based on their mutual information scores. Aims to reduce noise, improve model performance, and concentrates on the most important aspects to increase the accuracy of hair loss prediction.

*Test and Train Split:* The combined dataset—which includes the best features chosen based on mutual information scores—is divided into 80% for training and 20% for testing to guarantee efficient evaluation of models and performance assessment.

*Make Prediction:* A trained machine learning model gets utilized input features to estimate the possibility or severity of hair loss. After analyzing this data, the model develops an outlook that shows a person's likelihood or classification of hair loss.

*Model Evaluation:* Metrics including accuracy, precision, recall, and F1 score are used in the hair loss prediction project to assess the model's performance. By striking a balance between precision and recall, these measures guarantee that the model correctly detects and forecasts cases of hair loss for a thorough evaluation.

*Visualization:* Visualization plays a significant part in the analysis of results upon metrics that involve accuracy, precision, and other performance indicators. Additionally, it aids in uncovering the root factors influencing each algorithm’s performance, offering deeper insights and stimulating a more comprehensive understanding of the outcomes.

IV. METHODOLOGY

*XGBoost Classifier:* XGBoost is utilized to predict hair loss by training on real-world data that comprises a variety of contributing factors. This method enables the identification of the most significant predictors of hair loss from the dataset. Applies gradient boosting enhanced prediction accuracy and effectively handles both classification and regression tasks. The model is assessed using metrics such as accuracy, classification reports, referring to Fig.3.

*Random Forest Classifier:* The datasets are employed to train the Random Forest model, which takes advantage of its capacity to generate multiple decision trees and aggregate their results to increase accuracy and decrease overfitting. Test data is used to make predictions, and metrics are used to assess the model's performance. This aids in assessing how well the model categorizes the causes and severity of hair loss. Furthermore, the model's feature importance analysis provides information about the factors that have the biggest effects on hair loss, as shown in Fig.5, directing future research and possible preventative actions.

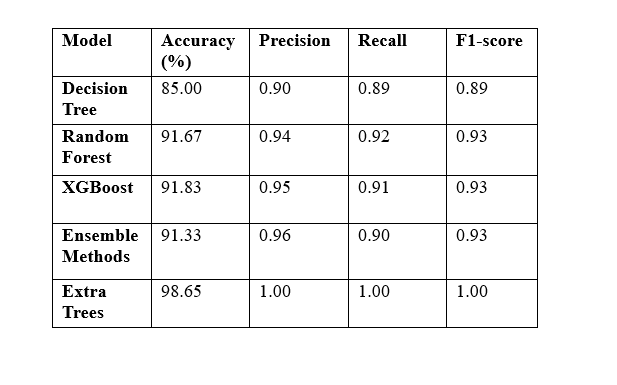
*Decision Tree Classifier:* A supervised learning technique called the Decision Tree algorithm divides data into branches in order to generate predictions based on feature values. Selected features from both synthetic and real data are used to train the Decision Tree Classifier for the hair loss prediction project. Following training, test data is utilized to make predictions, and accuracy scores are used to assess the model's performance. As shown in Fig. 4, this approach offers easy interpretability and enables a thorough comprehension of the ways in which different factors contribute to hair loss.

*Extra Trees Classifier:* In order to improve predictive accuracy, the Extra Trees (Extremely Randomized Trees) algorithm is an ensemble learning technique that constructs numerous decision trees using randomly chosen features and data splits. Selected features from both synthetic and real data are used to train the Extra Trees Classifier for the hair loss prediction project. Following training, predictions are made using test data, and accuracy scores are used to evaluate the model's performance. As shown in Fig. 7, this method increases feature importance analysis and prediction stability, offering a deeper understanding of the causes of hair loss.

*Ensemble Methods***:** Several machine learning models are used in ensemble learning to improve prediction accuracy and lessen overfitting. The predictions of classifiers like Random Forest, Extra Trees, and XGBoost are combined in the hair loss prediction project using ensemble approaches like bagging and boosting.Ensemble approaches increase robustness and generalization by combining various models. Accuracy scores are used to gauge how well the trained ensemble model performs when tested on test data. As shown in Fig. 6, our method guarantees a more accurate prediction framework and offers thorough insights into the main causes of hair loss.

V. RESULT AND CONCLUSION

The Extra Trees model stands achieved the highest accuracy of 98.65% accuracy, demonstrating perfect scores in both precision and recall, making it the most effective model for real-time prediction. XGBoost and Random Forest also performed well with accuracies of 91.83% and 91.67%, respectively. The results validate the robustness of ensemble methods and their effectiveness in identifying primary causes of hair loss such as stress and poor sleep. Clear visual comparisons in Fig. 3 to Fig. 8 support these findings.

Table1. Output values

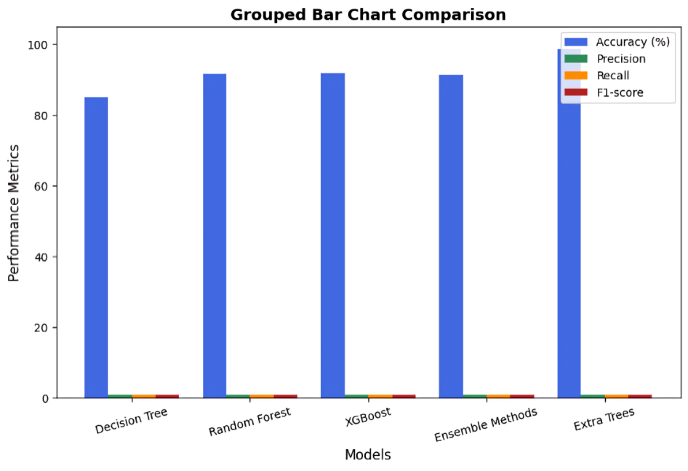
*Diagrammatic representation of outputs*

Fig.2. Result Analysis

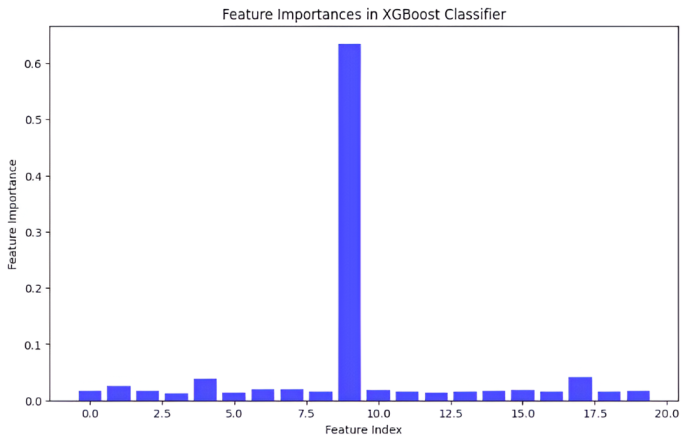


Fig.3. XGBoost Analysis

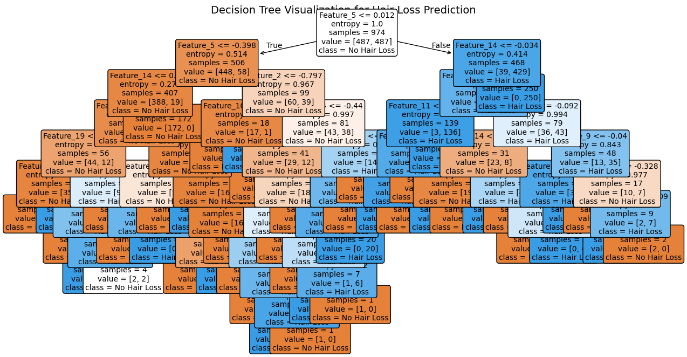


Fig.4. Decision tree Analysis

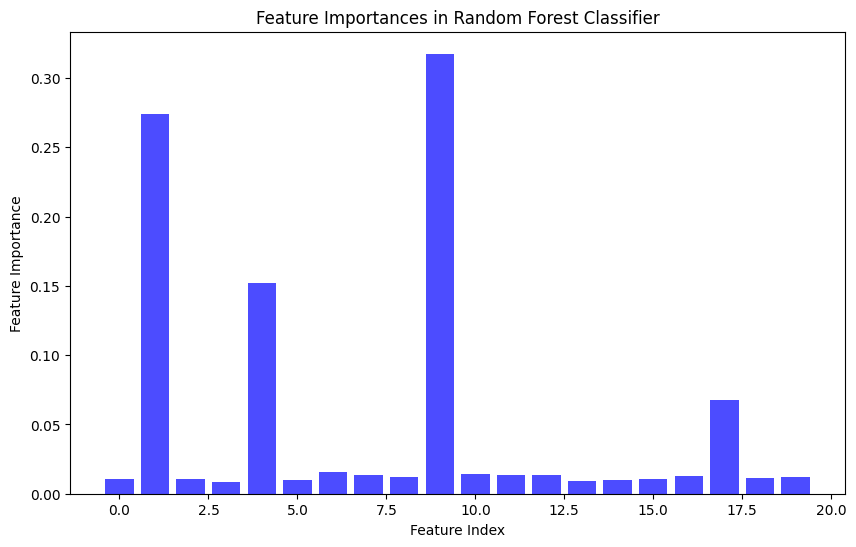
****

Fig.5. Random Forest Analysis

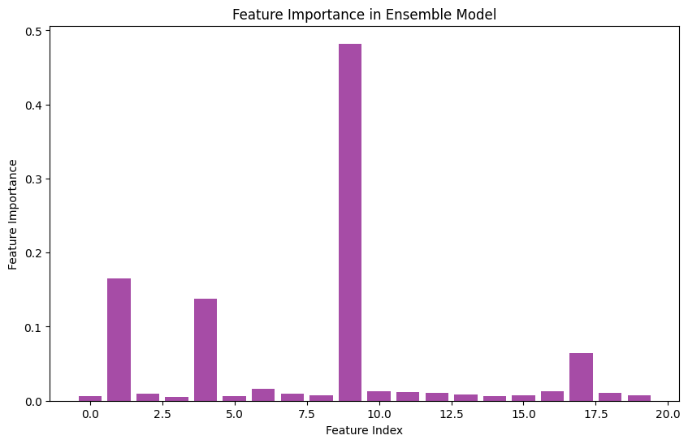
****

Fig.6. Ensemble methods

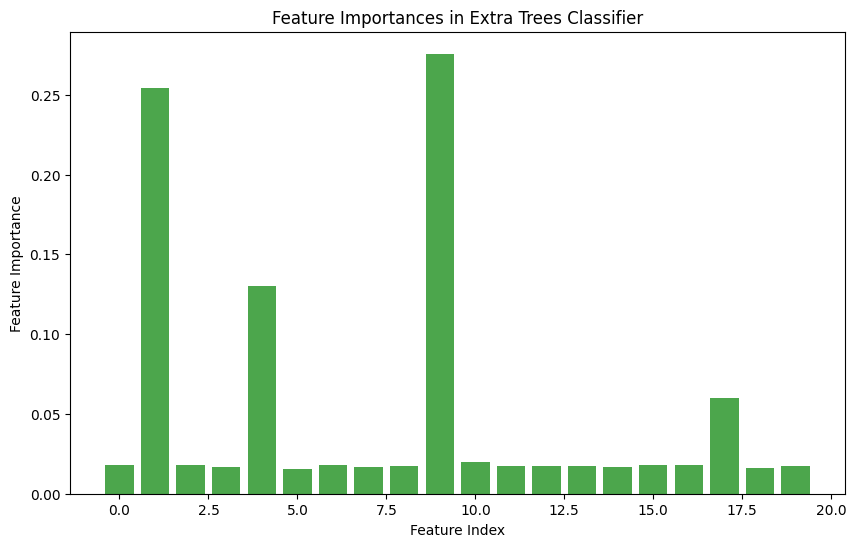


Fig.7. Extra Tress Analysis

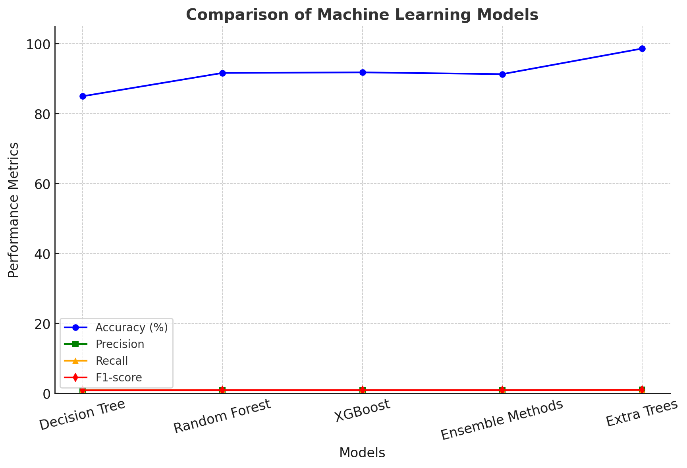


Fig.8. Overall Analysis

VI. FUTURE SCOPES

To provide individualized treatment suggestions, future developments in hair loss prediction will concentrate on combining wearable IoT devices with AI-driven real-time monitoring systems. The accuracy and generalizability of the model will be enhanced by adding multi-modal inputs to the dataset, such as genetic markers, lifestyle information, and scalp imaging. To increase openness and confidence in AI-driven diagnosis, explainable AI (XAI) will be essential.

Additionally, secure data sharing across various healthcare institutions will be made possible by privacy-preserving strategies like federated learning. To get over data constraints and increase prediction robustness, future research should also investigate the incorporation of generative models (GANs). Finally, scalable cloud-based platforms will facilitate international cooperation and promote telemedicine and real-time diagnostics for increased accessibility.

REFERENCES

[1] A. Vinora, P. U. Maheswari, R. N. Deborah, E. Ajitha, A. Srinivasan, and G. Sivakarthi, “An Effective Approach Using Machine Learning for Fall Prediction,” in *2023 9th International Conference on Smart Computing and Communications (ICSCC)*, Kochi, Kerala, India: IEEE, Aug. 2023, pp. 315–320. doi: 10.1109/ICSCC59169.2023.10334969.

[2] D. Banerjee, V. Kukreja, D. Bordoloi, and A. Choudhary, “Enhanced Hair Disease Classification Using Deep Learning,” in *2024 11th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India: IEEE, Mar. 2024, pp. 58–63. doi: 10.1109/SPIN60856.2024.10511333.

[3] F. Khatun, M. R. Ajmain, S. A. Khushbu, N. J. Ria, and S. R. H. Noori, “Survey-based Machine learning approaches to diagnosis of hair fall disorder in Bangladeshi Community,” in *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India: IEEE, Oct. 2022, pp. 1–6. doi: 10.1109/ICCCNT54827.2022.9984226.

[4] G. S. Ioannidis *et al.*, “A Machine Learning Framework for Hair Type Categorization to Optimize the Hair Removal Algorithm in Dermoscopy Images,” in *2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology*, Malta: IEEE, Dec. 2023, pp. 121–122. doi: 10.1109/IEEECONF58974.2023.10404510.

[5] H. Weerasinghe and D. Vidanagama, “Machine Learning Approach for Hairstyle Recommendation,” in *2020 5th International Conference on Information Technology Research (ICITR)*, Moratuwa, Sri Lanka: IEEE, Dec. 2020, pp. 1–4. doi: 10.1109/ICITR51448.2020.9310868.

[6] L. Srinivasan, A. Jeevika, R. K. Navina and S. Priyadharshini, “An Enhanced Stress based Hairfall Detection and Prevention using KNN and Machine Learning Techniques,” in 2023 7th International Conference on Trends in Electronics and Informatics Tirunelveli, India: IEEE, Apr. 2023, pp. 11101115.doi: 10.1109/ICOEI56765.2023.10125720.[7] P. Duraisamy, V. Vidhya, T. Preetha, and V. Niranjani, “Implementing Machine Learning Techniques for Hair Health Prediction in Follicle Futures,” in *2024 Second International Conference on Data Science and Information System (ICDSIS)*, Hassan, India: IEEE, May 2024, pp. 1–6. doi: 10.1109/ICDSIS61070.2024.10594069.

[8] P. Kesika, B. S. Sivamaruthi, S. Thangaleela, M. Bharathi, and C. Chaiyasut, “Role and Mechanisms of Phytochemicals in Hair Growth and Health,” Pharmaceuticals, vol. 16, no. 2, p. 206, Jan. 2023, doi: 10.3390/ph16020206.

[9] P. Riangjanapatee *et al.*, “Development of Tea Seed Oil Nanostructured Lipid Carriers and In Vitro Studies on Their Applications in Inducing Human Hair Growth,” *Pharmaceutics*, vol. 14, no. 5, p. 984, May 2022, doi: 10.3390/pharmaceutics14050984.

[10] S. Aditya, S. Sidhu, and M. Kanchana, “Prediction of Alopecia Areata using Machine Learning Techniques,” in *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*, Hassan, India: IEEE, Jul. 2022, pp. 1–6. doi: 10.1109/ICDSIS55133.2022.9915804.

[11] S. Srivastav, K. Guleria, and S. Sharma, “Deep Learning-based Convolutional Neural Network Model for Hair Diseases Detection,” in *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, Bangalore, India: IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/SMARTGENCON60755.2023.10442545.

[12] V. Khan and K. Subramaniam, “An Effective

Optimization-Based Convolutional Neural Network for Effective Hair and Scalp Detection System,” in *2023 1st International Conference on Cognitive Computing and Engineering Education (ICCCEE)*, Pune, India: IEEE, Apr. 2023, pp. 1–8. doi: 10.1109/ICCCEE55951.2023.10424503.

[13] W. Y. Chow, W. W. Heng, N. A. Abdul-Kadir, and H. Nadaraj, “Explainable Machine Learning on Classification of Healthy and Unhealthy Hair,” in *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, Bandung, Indonesia: IEEE, Dec. 2022, pp. 162–167. doi: 10.1109/ICICyTA57421.2022.10037969.

[14] Yashu, V. Kukreja, P. Srivastava, and A. Garg, “An Empirical Study on Hair Strength Classification using CNN and Random Forest Ensembles,” in *2024 2nd International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India: IEEE, Mar. 2024, pp. 209–214. doi: 10.1109/ICDT61202.2024.10489441.

[15] Y. Wang, M. Hsu, M. Y. Wang, and J. Lin, “Estimating hair density with XGBoost,” *Int. J. Cosmet. Sci.*, p. ics.13030, Nov. 2024, doi: 10.1111/ics.13030.