IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131112

Comparative Predictive Modeling of Dry Eye Disease: An Integrated Approach Using Decision Tree and Random Forest Techniques

Mohammed Ihsan N[1], Sharan R[2], Dr. Paavai Anand[3]

Student, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India[1]

Student, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India[2]

Guide, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India[3]

Abstract: Dry Eye Disease (DED) is a multifactorial ocular disorder characterized by tear film instability and ocular surface inflammation, manifesting as discomfort and visual disturbances. Traditional diagnostic methods rely on subjective clinical evaluation and costly procedures, limiting accessibility. This work proposes a machine learning-based, non-invasive approach for predicting DED risk using patient demographics, lifestyle, and reported symptoms. Both Decision Tree and Random Forest classifiers are compared: Random Forest achieves superior accuracy (72.8\%) and F1-score (0.76). Feature importance ranks symptomology and behavioural factors as key predictors, supporting practical early intervention strategies.

Keywords: dry eye disease, machine learning, decision tree, random forest, predictive modelling, feature importance

I. INTRODUCTION

Dry Eye Disease (DED) is a common eye condition that often goes undiagnosed until it becomes severe, mainly because traditional diagnostic methods are costly, invasive, and require specialized equipment. The problem is increasing due to long hours of screen use and aging populations. Many people, especially the elderly and those on long-term medication, are at high risk but lack access to early testing. This project aims to develop a cost-effective and non-invasive machine learning approach to predict DED using real-world patient data. The study compares two classifiers—Decision Tree and Random Forest—to find an accurate and interpretable model for early detection. The goal is to support clinicians with data-driven insights that can help identify high-risk individuals sooner and improve patient care.

A. Problem Motivation and Objective

The need for early, accessible DED prediction is urgent among vulnerable demographics such as the elderly or those on chronic medications. The objective of this project is to compare interpretable and robust machine learning classifiers—Decision Tree and Random Forest—utilizing patient features for effective risk stratification. The research aims to deliver actionable insights for clinicians, supporting data-driven screening and interventions.

B. Paper Organization

Section II reviews related work and technical concepts. Section III discusses dataset and preprocessing. Section IV details model architectures and algorithmic implementations. Section V presents results with critical evaluation. Section VI outlines limitations, future scope, and Section VII concludes.

II. LITERATURE REVIEW

Recent advancements in artificial intelligence (AI) and machine learning (ML) have greatly enhanced diagnostic capabilities in ophthalmology. Automated algorithms can now efficiently analyze retinal images and patient data with high accuracy. De Fauw et al. demonstrated the effectiveness of deep learning models for retinal disease detection and referral prediction. In the context of Dry Eye Disease (DED), previous studies have utilized AI techniques to estimate tear film stability and identify potential risk factors from health data. Ensemble models such as Random Forest have shown strong performance in clinical prediction tasks, particularly when handling structured and multivariate datasets. Table I presents the Representative Literature Review summarizing key related works.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131112

TABLE I
REPRESENTATIVE LITERATURE REVIEW

Author	Method	Key Result
De Fauw et al.	Deep Learning	Accurate retinal diagnosis
Nair et al.	Narrative Review / AI	DED prediction insights
Nam et al.	Network Analysis / ML	Risk factor discovery
Shimizu et al.	AI estimation	Tear film time prediction

III. DATASET DESCRIPTION

A. Patient Records

The Dry_Eye_Dataset includes 20,000 anonymized patient records with an initial set of 26 features. Features encompass demographics (age, gender, height, weight), lifestyle (sleep duration, stress, steps, caffeine/alcohol/smoking, average screen time), and direct ocular symptoms (discomfort, redness, blue-light filter usage).

B. Data Exploration

Exploratory analysis revealed significant associations between symptoms and screen time, as well as between age and discomfort reports, guiding the feature selection process. Raw distribution plots and pairwise correlations are illustrated in Figure 1.

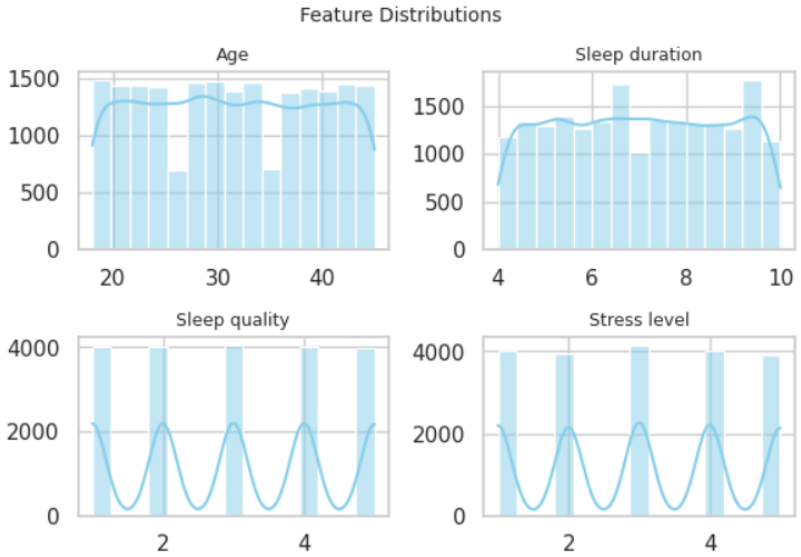

Figure 1. Exploratory Data Analysis: Feature Distributions

IV. DATA PREPROCESSING AND FEATURE ENGINEERING

A. Cleaning and Encoding

Missing data were minimal and handled using mean or mode imputation. Categorical variables (e.g., smoking: Yes/No) were encoded numerically; gender was binary coded. Numerical features were normalized using StandardScaler for optimal algorithm performance

B. Feature Selection

Feature importance analysis enabled dimensionality reduction from 26 initial variables to 14 key features, emphasizing those with both statistical significance and clinical relevance. The refined attributes form the core of the prediction model, ensuring improved interpretability and reduced redundancy. Table II presents the final selected feature set.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131112

TABLE II
FINAL FEATURE SET

Feature	Description
Age	Patient age
Gender	M/F encoded
Screen Time	Avg. daily hours
Discomfort	Eyestrain indicator
Redness	Eye redness
Sleep Duration	Avg. per night
Caffeine	Intake flag
Smoking	Status
Alcohol	Consumption flag
Steps	Avg. daily count
Stress Level	Reported
Blue-light Filter	Usage

C. Splitting and Balancing

A stratified train-test split (75/25) preserved class balances and representative metrics, avoiding bias in evaluation.

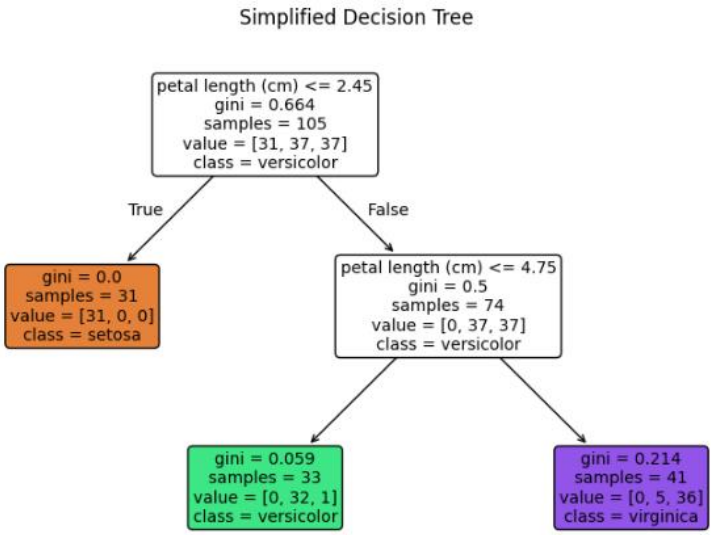

Figure 2. Sample Decision Tree Structure

V. MACHINE LEARNING ARCHITECTURE

A. Algorithmic Details

The Decision Tree model was implemented using the CART algorithm, which recursively partitions the dataset based on the maximization of Gini impurity. A simplified representation of the decision tree is shown in Figure 2. The Random Forest model, on the other hand, is an ensemble of multiple bootstrapped decision trees, where each tree performs random feature selection at every split to reduce correlation among trees. The final prediction is obtained through majority voting across all trees in the ensemble. Figure 3 illustrates the schematic of the Random Forest model.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131112

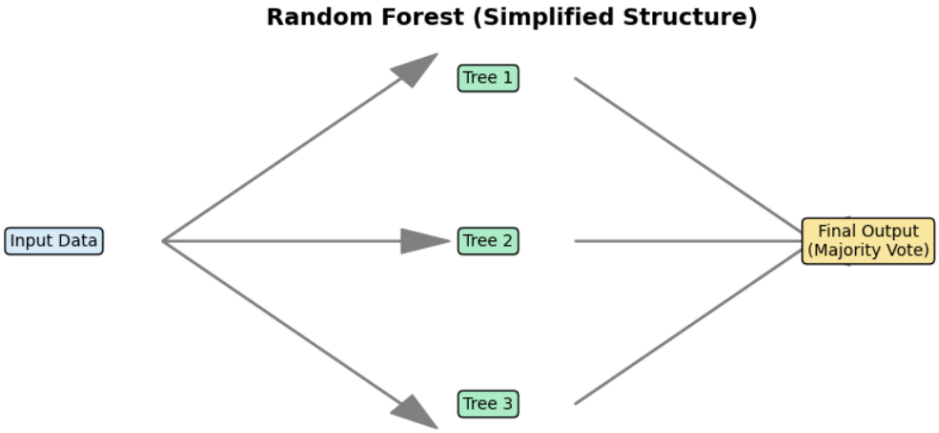

Figure 3. Ensemble Random Forest Structure

B. Hyperparameter Tuning

Both models were fine-tuned using grid search and cross-validation to identify the optimal configuration for parameters such as tree depth, number of estimators, and minimum sample splits. This systematic optimization ensured balanced model performance and minimized overfitting. Table III provides an overview of the best-performing hyperparameter settings.

TABLE III
OPTIMIZED HYPERPARAMETERS

Model	Hyperparameter	Value
Decision Tree	Max Depth	6
Random Forest	Trees	100
Random Forest	Max Features	8
Decision Tree	Min Samples per Leaf	10

VI. RESULT ANALYSIS

A. Classification Performance

Testing (N = 5,000) yielded accuracy, precision, recall, and F1-score values for each model (Table IV). The Decision Tree identified all DED-positive cases, but with an excessive number of false positives. Random Forest balanced sensitivity and specificity, outperforming in overall accuracy. Confusion matrices are visualized in Figures 4 and 5. ROC curves and PR curves further illustrate sensitivity and tradeoffs
(Figures 6 and 7).

TABLE IV
MODEL PERFORMANCE METRICS (TEST SET)

Model	Accuracy	Precision	Recall	F1
Decision Tree	64.40%	0.64	1	0.78
Random Forest	72.80%	0.68	0.87	0.76

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131112

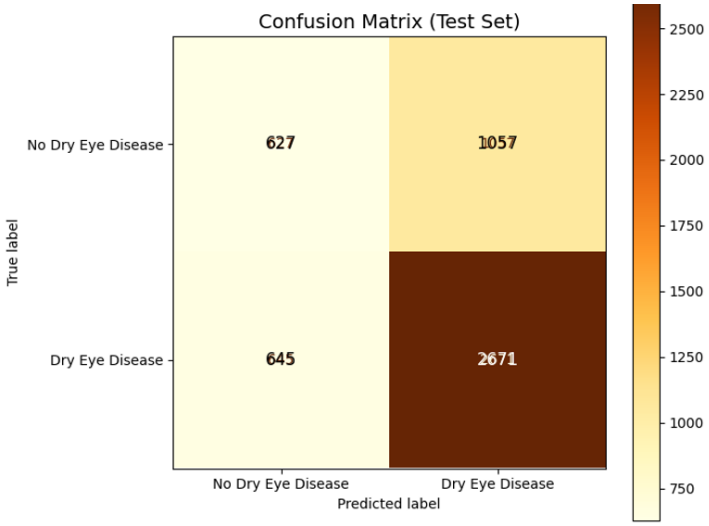

Figure 4. Confusion Matrix: Decision Tree

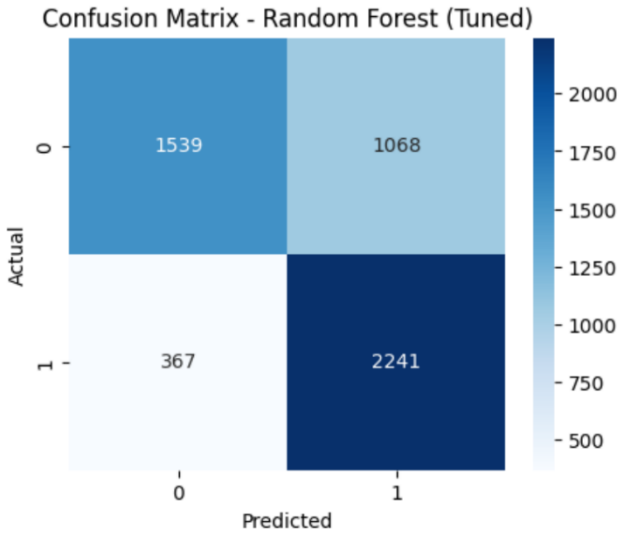

Figure 5. Confusion Matrix: Random Forest

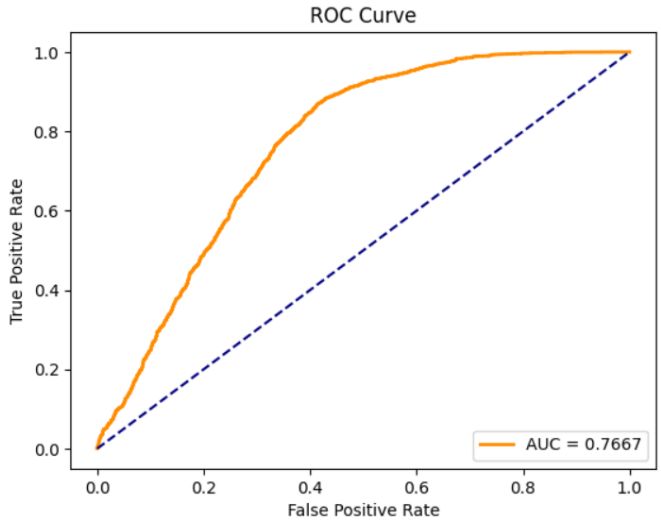

Figure 6. ROC Curve: Random Forest

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131112

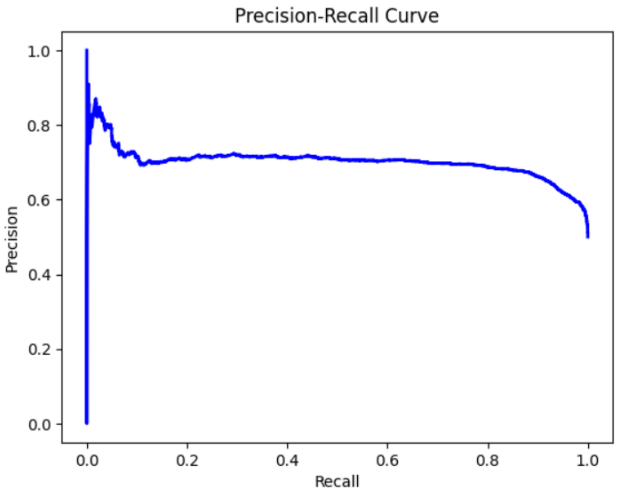

Figure 7. Precision-Recall Curve

B. Feature Importance

Figure 8 ranks the top predictors: discomfort, redness, screen time, and age, confirming clinical understanding and guiding targeted intervention.

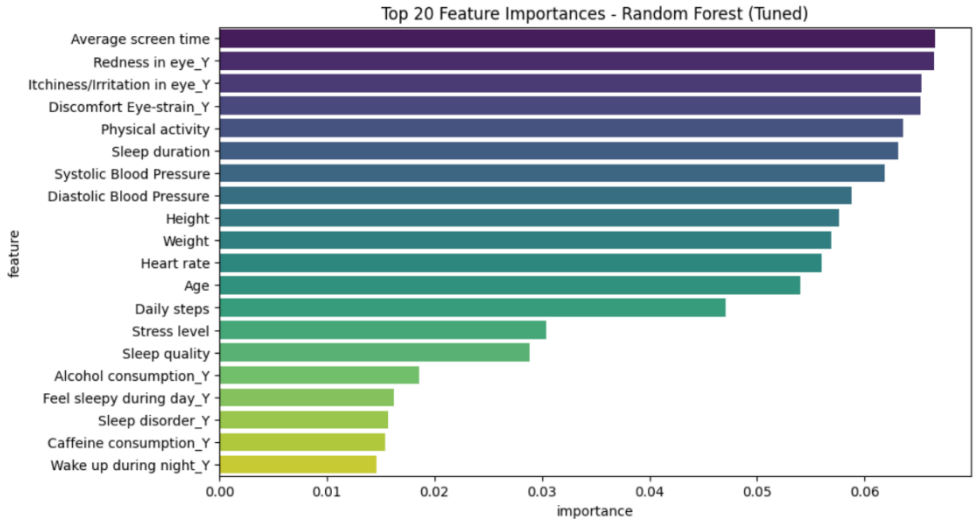

Figure 8. Feature Importance: Random Forest

VII. DISCUSSION

A. Interpretability vs. Robustness

Decision Trees provide clear and interpretable diagnostic pathways for clinicians, making them useful for understanding the underlying decision logic in DED prediction. However, they are prone to overfitting and may produce higher false positive rates when dealing with complex or noisy datasets. Random Forests, by contrast, combine multiple decision trees to achieve greater predictive stability and generalization. They reduce variance, enhance robustness, and generate quantitative feature importance rankings that can guide clinicians in identifying the most influential diagnostic parameters for effective clinical decision-making.

B. Clinical Insights

The findings emphasize the significant influence of modifiable lifestyle factors, particularly prolonged screen exposure and insufficient sleep, on the likelihood of developing Dry Eye Disease (DED). These results indicate that behavioral adjustments can play a vital role in disease prevention and management. The analysis further reveals that DED risk increases among individuals who experience greater visual discomfort associated with extended screen usage, underscoring the importance of targeted awareness programs and evidence-based intervention strategies that encourage healthier digital and sleep habits.

DOI: 10.17148/IJIREEICE.2025.131112

C. Limitations

The dataset used in this study, although extensive, may not fully represent the complete range of geographic or ethnic variations present in the wider population. Clinical evaluation continues to be essential for verifying the accuracy and reliability of machine learning predictions. Incorporating additional multimodal data such as imaging and physiological measurements in future work could further improve the model's comprehensiveness and diagnostic precision.

VIII. FUTURE WORK

This research successfully demonstrates a non-invasive framework for predicting Dry Eye Disease (DED) using Decision Tree and Random Forest classifiers trained on structured patient datasets. Building upon these findings, future work can focus on integrating additional clinical, behavioural, and environmental parameters to further enhance model accuracy and generalizability across diverse populations. Incorporating deep learning architectures, such as convolutional or recurrent neural networks, could also enable the extraction of complex, non-linear patterns from multimodal data sources, including ocular images and sensor readings. Furthermore, deploying the trained models as part of a cloud-based clinical decision support system could facilitate real-time screening and remote patient monitoring. In addition, explainable AI techniques can be leveraged to improve transparency, allowing clinicians to interpret the reasoning behind predictions. Ultimately, this line of research holds promise for transforming early DED detection and enabling personalized treatment strategies in ophthalmology.

IX. CONCLUSION

This study demonstrates an effective and non-invasive approach for predicting Dry Eye Disease (DED) using Decision Tree and Random Forest classifiers applied to structured patient datasets. The experimental results highlight that the Random Forest model achieves an optimal balance between sensitivity and specificity, offering a reliable mechanism for identifying key predictive features relevant to DED diagnosis. The interpretability of Decision tree–based models also ensures clinical transparency, allowing healthcare professionals to understand the decision-making process behind predictions. Overall, the findings confirm that machine learning–based diagnostic frameworks can serve as valuable tools for supporting early detection, improving patient outcomes, and streamlining clinical workflows in ophthalmology. Continued advancement of such data-driven approaches could play a pivotal role in enhancing preventive eye care and personalized treatment strategies for DED.

ACKNOWLEDGMENT

The authors sincerely acknowledge the Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India, for their continuous academic guidance, valuable feedback, and institutional support throughout the progression of this research work. Their encouragement and access to research infrastructure have been instrumental in facilitating the successful completion of this study. The dataset utilized in this research was sourced from publicly available repositories and is entirely independent of the institution. The authors confirm that no proprietary or institution-specific data were used, ensuring transparency, reproducibility, and adherence to ethical research standards.

REFERENCES

[1]. J. De Fauw, et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," Nature Medicine, vol. 24, pp. 1342–1350, 2018.

[2]. A. D. Graham, et al., "A machine learning approach to predicting dry-eye-related signs, symptoms and diagnoses from meibography images," *Heliyon*, vol. 10, p. e36021, 2024.

[3]. P. P. Nair, et al., "Artificial intelligence in dry eye disease: A narrative review," *Cureus*, vol. 16, no. 9, p. e70056, 2024.

[4]. S. M. Nam, et al., "Explanatory model of dry eye disease using health and nutrition examinations," *JMIR Medical Informatics*, vol. 8, no. 2, p. e16153, 2020.

[5]. E. Shimizu, et al., "Artificial intelligence to estimate the tear film breakup time and diagnose dry eye disease," *Scientific Reports*, vol. 13, p. 5822, 2023.

[6]. D. S. Ting, et al., "Artificial intelligence and deep learning in ophthalmology," *British Journal of Ophthalmology*, vol. 103, pp. 167–175, 2019.

[7]. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 4th ed., Morgan Kaufmann, 2022.