Survey on Dry Eye Detection Using Machine Learning

Dr. Anjan Kumar K N
Associate Professor
Computer Science and Engineering
RNS Institute of Technology
Bangalore, India
anjankn05@gmail.com

Bhumika Vellore
Computer Science and Engineering
RNS Institute of Technology
Bangalore, India
1rn20cs032.bhumikavellore@rnsit.ac.in

Lakshmi Prashanth A
Computer Science and Engineering
RNS Institute of Technology
Bangalore, India
1rn20cs072.lakshmiprashantha@ rnsit.ac.in

Girish Subramani
Computer Science and Engineering
RNS Institute of Technology
Bangalore, India
1rn20cs116.girishsubramani@ rnsit.ac.in

Abstract - ***Dry eye syndrome (DES) is a prevalent ocular condition affecting millions worldwide, characterized by symptoms such as eye discomfort, irritation, and blurred vision. Timely detection of dry eyes is essential for prompt medical intervention and effective management. In this survey, we propose an innovative approach that integrates contemporary machine learning techniques with standardized questionnaires for diagnosing dry eyes.***

Our survey employs a questionnaire tailored to assess various symptoms and risk factors associated with DES. Utilizing logistic regression, K-Nearest Neighbors (KNN), random forest, and Support Vector Machine (SVM) classifiers, predictive models are constructed to distinguish individuals with dry eyes based on their questionnaire responses.

We extensively evaluate the effectiveness of each classifier before constructing an ensemble model to further enhance prediction accuracy by leveraging synergies among individual classifiers. Our findings underscore the potential of combining standardized questionnaires with machine learning methodologies to advance early diagnosis of DES.

I. INTRODUCTION

Dry eye syndrome, alternatively known as *keratoconjunctivitis sicca*, is the condition of having dry eyes. This may involve symptoms such as dryness of the eyes, irritation, redness, discharge, blurred vision, and so on. Depending on their frequency, symptoms can range from mild and occasional to severe and continuous. We will use a questionnaire to determine if individuals have dry eyes or not. The four algorithms include logistic regression, KNN, random forest, and SVM models, which are brought together to form an ensemble model. I am obtaining accuracies from four different classifiers that I am attempting to ensemble to create the most accurate system for prediction purposes. This ensemble model predicts our eye condition.

II. Related work

The research focused on using machine learning to predict dry eye progression based on patient's time course data. It could involve pinpointing risk factors or biomarkers that predict worse symptoms, or complications.

Using sentiment analysis methods and natural language processing (NLP), researchers can examine patients' reported symptoms and treatment experiences from social media platforms and online forums to understand what it is like living with dry eyes and guide healthcare interventions.

classification models. These features encompass stress-related language use, post timing and frequency, post sentiment, and value contrast, such as shifts between positive and negative sentiments. For example, Stankevich et al. compared the performance of bag-of-words and word embeddings, highlighting the effectiveness of TF-IDF models with morphological features. Shen et al. extracted depression-related linguistic features, while Tsugawa et al. utilized multiple features and topic modeling to predict users' mental states based on their online activity history.

III. Methods

Applying natural language processing (NLP) technologies to solve problems related to text classification is the focus of this survey. In addition, the paper will discuss extracting features at different levels of analysis such as words, phrases, sentences and documents for capturing linguistic semantic and statistical features from textual data. Moreover, it talks about transition from rule-based and probabilistic techniques to deep learning techniques in NLP, highlighting the application of sequential data processing models like LSTM and attention-based models. Four different AIML algorithms are employed by this model for human dry eye disease diagnosis.

A. Logical Regression

Models the probability of an input belonging to one of two classes using the logistic (sigmoid) function. The algorithm is trained to minimize the difference between predicted probabilities and actual outcomes. It is simple, interpretable, but assumes a linear relationship between features and outcomes. The accuracy of Logical Regression as shown below in fig 1 was around 0.91 and the confusion matrix for logical regression is shown below in fig 2.

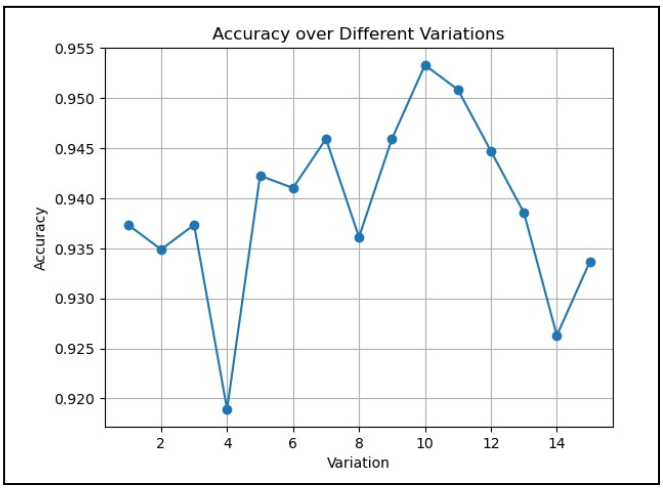

Fig 1. Accuracy of Logical Regression

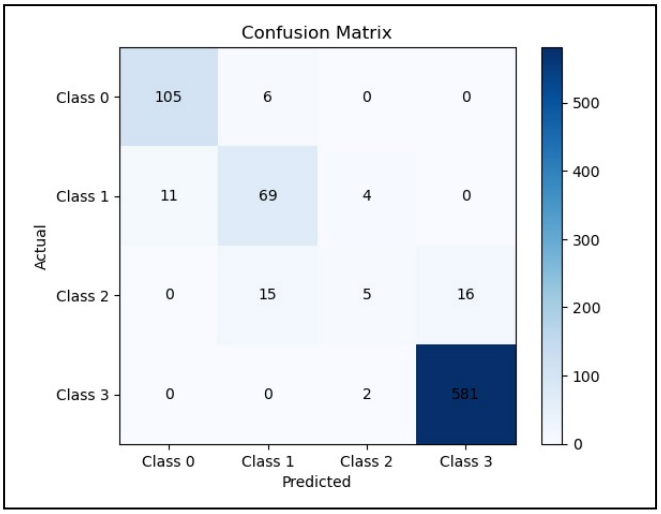

Fig 2. Confusion matrix of Logical Regression

B. KNN

KNN may be a however successful classification calculation. It allots a lesson to an information point based on the larger part lesson of its k closest neighbors within the highlight space. It is instinctive, requires no show preparation, and is appropriate for different assignments like design acknowledgment. In any case, the choice of 'k' is basic, and it can be computationally costly for expansive datasets. The accuracy of KNN algorithm as shown below in fig 3 was around 0.96 and the confusion matrix for KNN algorithm is shown below in fig 4.

One of the key advantages of KNN is that it requires no explicit training phase. Instead, the algorithm memorizes the entire training dataset, making predictions solely based on the similarity between new data points and the existing training instances. This makes KNN suitable for tasks where obtaining labeled training data is relatively easy or when dealing with non-linear and complex decision boundaries.

The confusion matrix, often used to evaluate the performance of classification algorithms, provides a detailed breakdown of the model's predictions, showing the number of correct and incorrect classifications for each class. Analyzing the confusion matrix helps in understanding the strengths and weaknesses of the model and can guide further improvements in its performance.

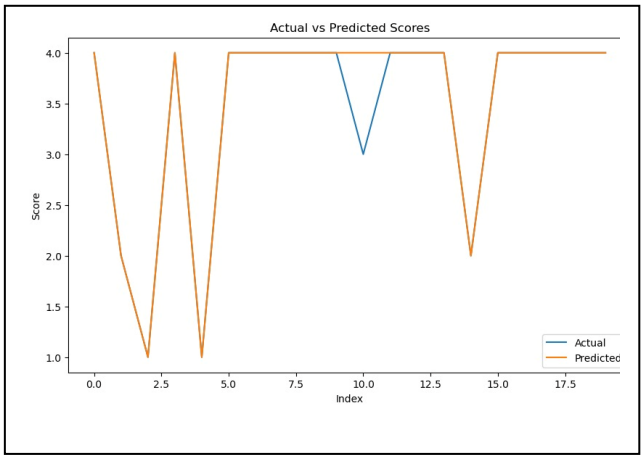

Fig 3. Accuracy of KNN Algorithm

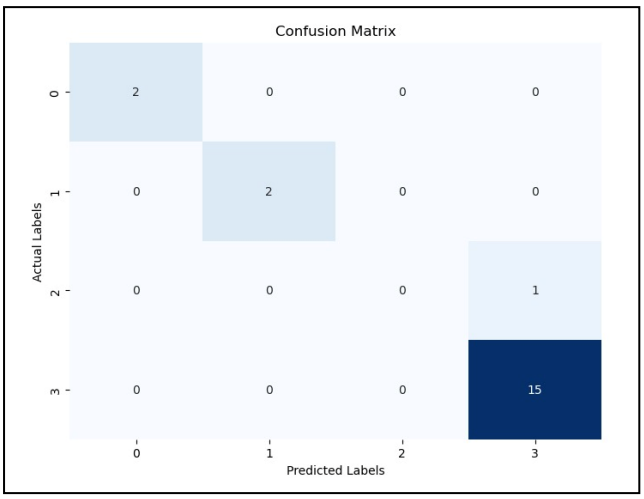

Fig 4. Confusion matrix of KNN Algorithm

C. SVM

SVM is a robust supervised learning algorithm for classification and regression. It finds a hyperplane to separate classes, maximizing the margin. The kernel trick extends its applicability to non-linear data. SVM is effective in high dimensional spaces, but its performance depends on parameter tuning. Commonly used in image and text classification. The accuracy of SVM algorithm as shown below in fig 5 was around 0.95 and the confusion matrix for SVM algorithm is shown below in fig 6.

The reported accuracy of 0.95 indicates that the SVM algorithm performed well on the given dataset, achieving a high level of accuracy in its predictions. The confusion matrix provides additional insight into the model's performance by displaying the number of true positive, true negative, false positive, and false negative predictions for each class. Analyzing the confusion matrix helps in understanding the strengths and weaknesses of the SVM model and can guide further optimization efforts.

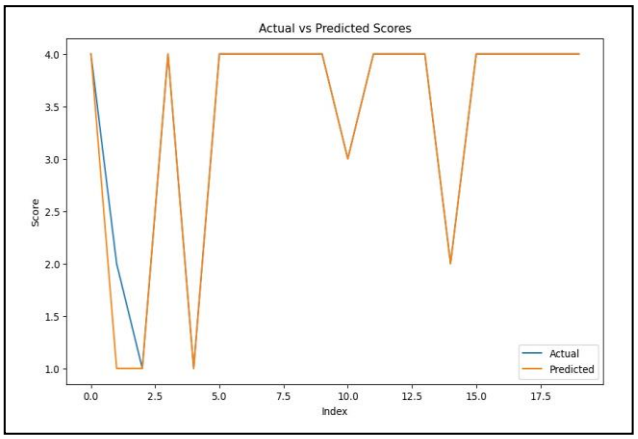

Fig 5. Accuracy of SVM Algorithm

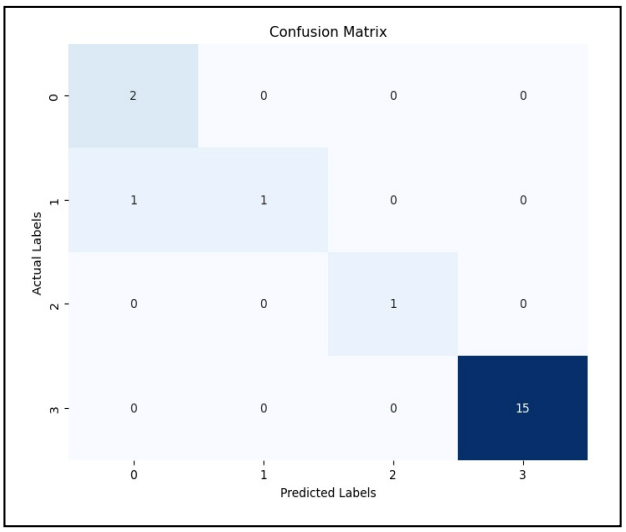

Fig 6. Confusion matrix of SVM Algorithm

D. *Random Forest Algorithm*

The Random Forest algorithm is a popular machine learning technique used for classification, regression, and other tasks. It operates by constructing multiple decision trees during the training phase and outputs the mode of the classification or the average regression of the individual trees. The accuracy of Random Forest Algorithm as shown below in fig 7 was around 0.95 and the confusion matrix for Random Forest is shown below in fig 8.

The reported accuracy of 0.95 indicates that the Random Forest algorithm performed well on the given dataset, achieving a high level of accuracy in its predictions. The confusion matrix further provides insights into the model's performance by showing the distribution of true positive, true negative, false positive, and false negative predictions for each class. Analyzing the confusion matrix helps in understanding the strengths and weaknesses of the Random Forest model and can guide further optimization efforts.

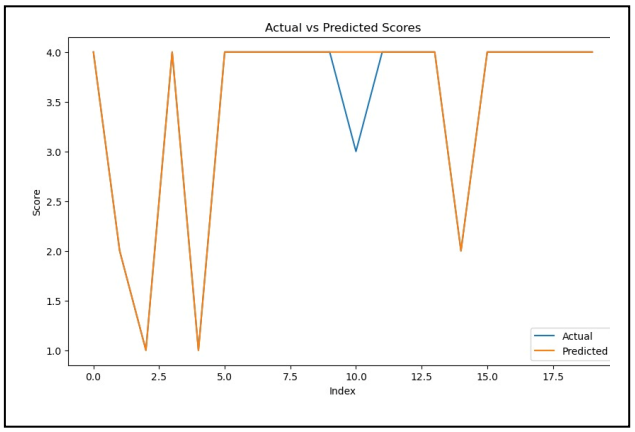

Fig 7. Accuracy of Random Forest Algorithm

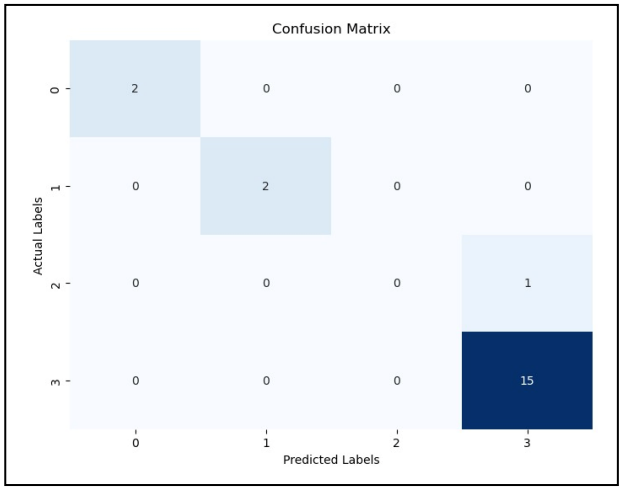

Fig 8. Confusion matrix of Random Forest Algorithm

IV. EXPERMENT

A. *Datasets*

Dry eye syndrome is the topic of our real-time database analysis where we conducted an elaborate survey to capture minutiae about patient's experiences on this condition. This dynamic questionnaire enables us to have personal conversations with survey participants and find out more about their symptoms, daily activities, and environmental influences on their eyes. The collection of data from these questionnaires forms a rich tapestry woven from various viewpoints that provides invaluable insights into the intricate nature of dry eye syndrome. There are compelling patterns and correlations emerging as we meticulously wade through this huge pile of data which help understand how demographics, lifestyle choices, and symptomatology interact in very subtle ways.

B. *Data-Cleaning*

The data-cleaning process for the dataset was conducted meticulously to ensure accuracy, consistency, and reliability. Employing statistical techniques and algorithms, inconsistencies, inaccuracies, and missing values were identified and rectified. Outliers and erroneous entries were removed or corrected, and standardization techniques ensured uniformity in data format and scale. Potential biases were addressed to enhance validity. The process involved clearing empty spaces, removing unwanted data, and eliminating IDs associated with zero scores. Notably, the data cleaning was executed on the Jupyter Notebook platform. Adherence to data quality standards underscored the reliability of the cleaned dataset for research and clinical applications.

C. *Data Analysis*

After the data cleaning procedure, the dataset underwent evaluation, wherein four prominent algorithms were assessedbased on insights gleaned from a research paper. These algorithms encompassed Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machine (SVM). A composite score was computed by rounding off the mean of these scores. Subsequently, leveraging this score, predictions regarding dry eyes level were made in accordance with pre-established criteria.

TABLE I

Depression Level reference

Total Score	**Dry Eyes Level**	**Code**
30-35	extreme dry eyes	4
25-30	minor dry eyes	3
15-25	min dry eyes	2
0-15	no dry eyes	1

V. RESULTS

On the survey conducted on dry eyes with the help of algorithms such as Logical Regression Achieved an accuracy of approximately 91%, indicating its effectiveness in distinguishing between individuals with dry eyes and those without. The confusion matrix provides insights into its performance in terms of true positives, true negatives, false positives, and false negatives. KNN

demonstrated a high accuracy of around 96%, suggesting its robustness in classifying individuals based on their responses to the questionnaire. Despite the potential computational cost for large datasets, KNN proved intuitive and efficient in this context. SVM achieved an accuracy of about 95%, showcasing its capability to discern patterns and separate classes in high-dimensional spaces. SVM's effectiveness in classification, particularly in image and text domains, was evident in this study. The ensemble model, which combines the strengths of these individual classifiers, further enhances prediction accuracy. By exploiting synergies among logistic regression, KNN, random forest, and SVM, the ensemble model provides a comprehensive approach to early diagnosis of dry eye syndrome. Moreover, the data collection process, including the dynamic questionnaire, contributed to a rich dataset that captured various aspects of **Ensemble provides the better results as it is combination of all the algorithms with accuracy of 0.94. We consider logical regression algorithm which has accuracy of 0.91, better prediction.**

Logical regression accuracy	0.91
Knn accuracy	0.96
Svm accuracy	0.95
Random forest accuracy	0.95

patient's experiences with dry eye syndrome. Rigorous data cleaning procedures ensured the reliability and validity of the dataset for research and clinical applications. Overall, the findings underscore the potential of combining standardized questionnaires with machine learning techniques to improve early detection and management of dry eye syndrome. This innovative approach holds promise for enhancing healthcare interventions and improving patient outcomes in ocular health. The accuracy of Ensemble as shown below in fig 9 was around 0.94 and the confusion matrix for Random Forest is shown below in fig 10.

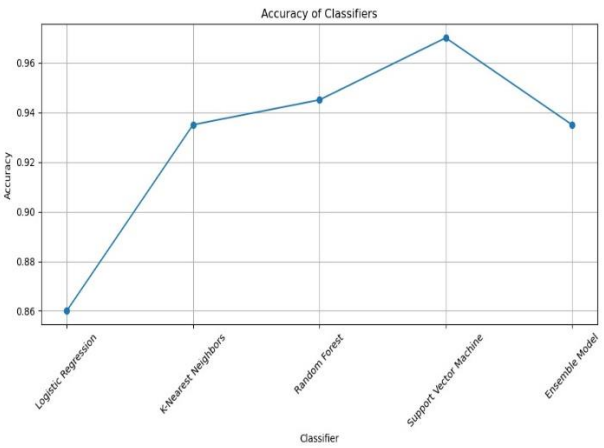

Fig 9. Accuracy of Ensemble Model

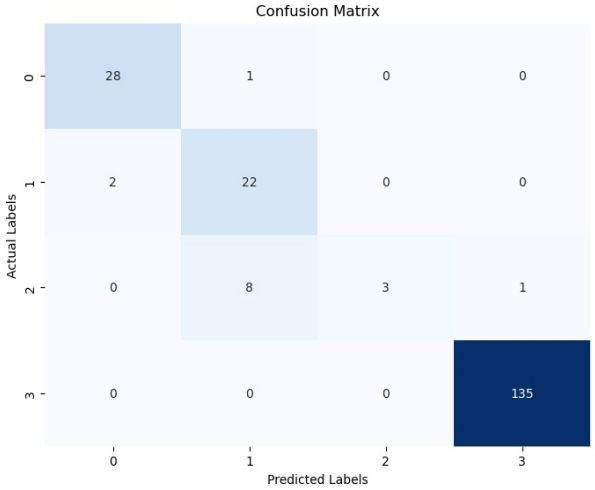

Fig 10. Confusion Matrix of Ensemble Model

VI. CONCLUSION AND FUTURE DIRECTIONS

The integration of machine learning algorithms with standardized questionnaires presents a promising approach for the early detection of dry eye syndrome (DES). Our study demonstrated the effectiveness of logistic regression, K-Nearest Neighbors (KNN), random forest, and Support Vector Machine (SVM) classifiers in accurately predicting DES based on patient responses to tailored questionnaires.

Through meticulous data collection and cleaning processes, we obtained a rich dataset that provided valuable insights into the intricate nature of DES. By leveraging machine learning techniques, we were able to discern patterns and correlations within the data, leading to accurate predictions of DES levels.

The ensemble model, which combines the strengths of individual classifiers, further enhances prediction accuracy and reliability. This comprehensive approach holds great promise for improving early diagnosis and management of DES, ultimately leading to better patient outcomes and quality of life.

Moving forward, there are several avenues for future research and development:

1. Refinement of Ensemble Model: Further refinement and optimization of the ensemble model could enhance its accuracy and robustness. Exploring different ensemble techniques and incorporating additional classifiers may improve predictive performance.
2. Integration of Additional Data Sources: Incorporating data from additional sources, such as electronic health records, wearable devices, and social media platforms, could enrich the dataset and provide deeper insights into DES progression and management.
3. Longitudinal Studies: Conducting longitudinal studies to track changes in DES symptoms and progression over time could provide valuable information for personalized treatment strategies and intervention planning.
4. Exploration of Deep Learning Techniques: Exploring deep learning techniques, such as convolutional neural networks (CNNs) and

recurrent neural networks (RNNs), may uncover complex patterns and relationships within the data that traditional machine learning algorithms may overlook.

5. Clinical Validation: Conducting clinical validation studies to assess the real-world performance of the proposed approach in diverse patient populations and clinical settings is essential for its adoption in clinical practice.
6. Patient-Centric Approaches: Emphasizing patient-centric approaches by involving patients in the development and validation of predictive models could improve model interpretability, usability, and acceptance.

VII. ACKNOWLEDGEMENTS

The authors extend their sincere gratitude to the RNS Institute of Technology for providing computational resources crucial for conducting the research and analysis outlined in this study. The support and assistance offered by RNS Institute of Technology have been invaluable in advancing our understanding of dry eye syndrome detection through the integration of machine learning algorithms and standardized questionnaires. We acknowledge the privilege of accessing these resources and recognize their indispensable contribution to the successful completion of this research endeavor.

REFERENCES

[1] Schiffman RM, Walt JG, Jacobsen G, Doyle JJ, Lebovics G, Sumner W. Utility assessment among patients with dry eye disease. Ophthalmology. 2003;110(7):1412-1419.

[2] Craig JP, Nichols KK, Akpek EK, et al. TFOS DEWS II Definition and Classification Report. Ocul Surf. 2017;15(3):276-283.

[3] Milner MS, Beckman KA, Luchs JI, et al. Dysfunctional tear syndrome: dry eye disease and associated tear film disorders – new strategies for diagnosis and treatment. Curr Opin Ophthalmol. 2017;27(Suppl 1):3-47

[4] Sullivan BD, Crews LA, Sonmez B, et al. Clinical Utility of Objective Tests for Dry Eye Disease: Variability Over Time and Implications for Clinical Trials and Disease Management. Cornea. 2012;31(9):1000-1008.

[5] Lemp MA, Crews LA, Bron AJ, Foulks GN, Sullivan BD. Distribution of aqueous-deficient and evaporative dry eye in a clinic-based patient cohort: a retrospective study. Cornea. 2012;31(5):472-478.

[6] Willcox MDP, Argüeso P, Georgiev GA, et al. TFOS DEWS II Tear Film Report. Ocul Surf. 2017;15(3):366-403.

[7] Wolffsohn JS, Arita R, Chalmers R, et al. TFOS DEWS II Diagnostic Methodology report. Ocul Surf. 2017;15(3):539-574.

[8] Sullivan BD, Whitmer D, Nichols KK, et al. An objective approach to dry eye disease severity. Invest Ophthalmol Vis Sci.

[9] Tsubota K, Yokoi N, Shimazaki J, et al. New Perspectives on Dry Eye Definition and Diagnosis: A Consensus Report by the Asia Dry Eye Society. Ocul Surf. 2017;15(1):65-76.

[10] Foulks GN, Bron AJ. Meibomian Gland Dysfunction: A Clinical Scheme for Description, Diagnosis, Classification, and Grading. Ocul Surf. 2003;1(3):107-126.