Predicting Digital Eye Strain Severity Using Machine Learning: A Hybrid Web-Based System

Lakshidha P R
RA2312704010043
Department of Computer Science
SRM Institute of Science and Technology
lp1822@srmist.edu.in

Rajyalakshmi Ambati
RA2312704010046
Department of Computer Science
SRM Institute of Science and Technology
ra9172@srmist.edu.in

Vamshi Polkampally
RA2312704010047
Department of Computer Science
SRM Institute of Science and Technology
vp8171@srmist.edu.in

Guide: Dr. Yasotha B
Assistant Professor
Department of Data Science and Business Systems
SRM Institute of Science and Technology

***Abstract*— The explosion in the use of digital devices has made Digital Eye Strain (DES), also referred to as Computer Vision Syndrome (CVS), one of the most common workplace health issues of the 21st century. A recent 2023 meta-analysis, pooling 103 studies and more than 66,000 participants, to establish the global incidence of CVS at 69.0%, with the COVID-19 pandemic further exacerbating the problem to 74% by imposing work-from-home and online learning protocols. Yet existing solutions mainly focus on passive screen time monitoring, subjective diagnosis through questionnaires and no predictive system for at-risk individuals. This paper introduces a data-driven model that predicts the severity of DES using nine behavioral and environmental factors with machine learning. We created a dataset of 2000 user profiles with features such as screen time, blink rate, frequency of breaks, screen brightness, room light, distance from the screen, sleep duration, age and use of corrective lenses. Linear Regression predicted continuous Eye Strain Index (ESI) with R^2 of 0.8846 and Root Mean Square Error (RMSE) of 4.73. Logistic Regression achieved 87.00% classification accuracy and Random Forest 83.25% for classification into risk levels (Low / Medium / High). K-Means (k=3) clustering identified three user profiles. Screen time (19.83%) and frequency of breaks (21.92%) are the most important features. Findings are presented in a Streamlit interactive dashboard for real-time prediction and recommendations. The system is a feasible preventive approach for DES.**

Index Terms: Computer Vision Syndrome; Digital Eye Strain; Machine Learning; Logistic Regression; Random Forest; K-Means Clustering; Streamlit; Preventive Healthcare; Screen Ergonomics; Blink Rate.

I. Introduction

The inclusion of digital devices in nearly all aspects of human endeavour - work, study, health, communication and leisure has ushered in a new realm of screen addiction. Laptops, mobile phones, tablets and e-readers are embedded into every aspect of our lives, with a world average daily screen time of over six hours for employed adults, and increasing. This revolution, which has delivered unsurpassed productivity and connectivity, has also brought a visual health epidemic: Digital Eye Strain (DES) defined in the 2023 Tear Film and Ocular Surface Society (TFOS) Lifestyles Report as "the development or exacerbation of recurrent ocular symptoms and/or signs related specifically to digital device screen viewing" [1].

DES, previously known as Computer Vision Syndrome (CVS) by the American Optometric Association, is a complex syndrome of ocular and non-ocular symptoms resulting from the distinctive visual requirements of digital displays. Digital screens use pixels with reduced contrast and flicker to display text, requiring the ciliary muscle to continuously accommodate. This is coupled with the known decline in blink frequency while viewing screens (from a physiological baseline of 14-16 blinks per minute, to 4-6 blinks per minute), leading to tear film instability, ocular surface drying and dry eye symptoms [2]. The symptoms fluctuate according to screen brightness, ambient light levels and distance [3].

The prevalence of DES is high. In a seminal 2023 meta-analysis of 103 cross-sectional studies (with 66,577 participants), Ccami-Bernal et al. reported a global pooled prevalence of CVS at 69.0% (higher in women and in Africa and Asia) [4]. A pivotal moment was the COVID-19 pandemic: a 2024 systematic review by León-Figueroa et al. found an aggregate rate of 74% in 18 studies conducted during the pandemic, with the increase linked to the mass shift to remote work and online education [5]. A recent review of the literature from 2014 to 2024 supported these findings, confirming CVS as a condition that affects most of the world's digital population, and advocated for evidence-based digital prevention strategies [6].

The symptoms of DES include eye strain, dryness, burning, blurred vision, headaches, double vision, photophobia (light sensitivity) and neck and shoulder pain [7]. These symptoms affect academic achievement, work productivity and well-being. Importantly, patients are often not aware of the connections between their screen use and their symptoms, an opportunity for data-driven educational tools.

The technological ecosystem is not yet well positioned to take on this major health issue. The majority of existing digital wellness tools merely log screen time, without calculating a strain estimate or providing feedback to support behavior change. The current clinical approach is largely based on established questionnaires (Computer Vision Syndrome Questionnaire (CVS-Q) and Computer Vision Symptom Scale (CVSS17), which are retrospective and subjective [1]. Lacking a publicly available, predictive, data-driven system that anyone can use in real time to assess their individual risk of DES depending on their usage, there is no tool available to an individual to gauge their risk of DES in real time, based

on their particular usage patterns.

This paper aims to fill this gap by proposing a machine learning system to predict DES severity. The system is built on a multi-model approach using Linear Regression for predicting strain index on a continuous scale, Logistic Regression and Random Forest classification for predicting strain index on a discrete scale (risk levels), and K-Means clustering for clustering the screen users based on usage patterns, using a 2,000-record data set of screen-usage parameters. The outcomes are presented via a real-time, interactive Streamlit dashboard offering personalised recommendations for prevention

.

II. Background and Motivation

A. Physiological and Behavioral Basis of Digital Eye Strain

Digital Eye Strain (DES) is caused by both physiological and behavioural effects of extended use of digital devices. Physiologically, the continuous accommodation required for screen viewing leads to continuous ciliary muscle contraction, causing eye strain and fatigue. Compared to reading text on paper, digital screens present information in the form of pixelated structures, with a low degree of contrast and variability in luminance, which places an additional load on visual processing [8].

Another major contributor to DES is the decrease in blink frequency when viewing a screen. While the average blink rate is around 14-16 blinks per minute, this decreases to 4-6 blinks per minute while using digital displays [2], [3]. This can cause a disruption in the tear film, resulting in dryness of the ocular surface, irritation and dry eye-like symptoms. Factors like excessive screen luminance, inappropriate lighting and close viewing distance also contribute to visual discomfort [2].

Behavioural factors are also important in the development of DES. Extended viewing times, insufficient breaks, and suboptimal posture all contribute to overall visual fatigue. The combination of physiological and behavioural factors leads to a complex and multifaceted issue that cannot be solved via single-point interventions.

B. Limitations of Existing Systems and Need for Predictive Approaches

Although the incidence of DES is increasing, the current technological and clinical responses are largely reactive. Existing digital wellness systems primarily track screen time, providing little understanding of the impact of individual lifestyles on eye strain. They fail to cater for the multiple factors involved or offer recommendations tailored to individuals' unique needs. Currently, DES diagnosis is based on subjective questionnaires like the Computer Vision Syndrome Questionnaire (CVS-Q) which evaluate symptoms retrospectively [1]. These tools are useful for diagnostic purposes, but lack real-time predictive capabilities and preventative measures. This means that users only realise the issue after the symptoms have already set in, making it more difficult to correct. Additionally, the current research in the field of ophthalmology is mainly focused on detection of disease using medical imaging like retinal imaging and Optical Coherence Tomography (OCT) [15]. These methods have high performance but lack practicality and accessibility.

C. Motivation for a Data-Driven Machine Learning Framework

The shortcomings of existing methods suggest a shift from diagnostic to predictive approaches based on data. Machine learning offers an effective approach to capturing interactions between multiple behavioral and environmental variables that can be used to predict the severity of DES from user-specific data. Computationally, the interactions between factors like viewing time, blinking and break-taking are nonlinear and complex. While conventional statistical techniques may overlook these complexities, machine learning approaches can learn complex relationships and patterns in the data [21]. This makes them well- suited for predicting behavioral health outcomes like DES. The aim of this study is to build an integrated system that incorporates regression, classification and clustering methods to gain a holistic view of eye strain. By predicting continuous strain levels, categorizing users into risk groups, and and clustering users based on their patterns, this system provides a holistic approach to DES. Moreover, the incorporation of these models into a interactive dashboard increases the ease of use, enabling users to get real-time feedback and suggestions. This is in line with the growing focus on prevention in health care, where smart systems enable people to make better choices and develop healthy digital lifestyles [20].

III. System Architecture

The system to predict the severity of Digital Eye Strain (DES) is a well-organised, layered architecture that combines data processing, machine learning algorithms and human interaction into a seamless workflow. It is structured with modularity and scalability in mind to allow accurate predictions while being useful to the user. It has a linear structure, starting with data collection and culminating with user feedback via an interactive dashboard.

TABLE I
SYSTEM COMPONENTS AND PRIMARY FUNCTIONS

Layer	Components	Primary Function
Data Acquisition	User input interface, dataset (2000 records)	Collection of behavioral and environmental features
Preprocessing	StandardScaler, train-test split, PCA	Data normalization and preparation
Analytics	Linear Regression, Logistic Regression, Random Forest, K-Means	Prediction, classification, and clustering
Prediction & Recommendation	ESI mapping, risk categorization, rule-based suggestions	Risk interpretation and personalized recommendations
Engagement	Streamlit dashboard	Visualization, user interaction, and result delivery

A. Data Acquisition Layer

The data acquisition layer acquires variables of interest to DES prediction. Currently, a structured data set of 2,000 observations with nine features based on clinically established risk factors [2], [3], [8] is used in the system. These features are screen time, blink rate, background light, screen brightness, distance from screen, sleep time, number of breaks, age and use of spectacles.

The system allows static and real-time data input. Currently, data are input via a user interface but the architecture allows future extensions to include automated data input such as sensor measurements or dynamic data collection systems. This allows for adaptability in both lab and field settings for data collection.usage.

B. Data Preprocessing Layer

The data transformation layer converts the raw input data to a format that can be used for training the model and making predictions. The dataset is not sparse, with no missing data, so preprocessing mainly consists of feature scaling and data splitting. Continuous features are scaled using z-score normalization to equalise the weight of features with different scales. To partition the data into training and testing data with a 80:20 ratio, stratified sampling method is used for classification problems to maintain class distribution. To visualise the clusters in 2D, we use Principal Component Analysis (PCA) to map the high dimensional feature space onto a plane.

C. Analytics Layer

The analytics layer represents the computational engine, and is responsible for the multi-model machine learning prediction of DES. We use Linear Regression to predict the continuous Eye Strain Index (ESI), which is a measure of strain. To classify users into Low, Medium and High risk categories, the Logistic Regression and Random Forest algorithms are used. Logistic Regression provides probabilistic outputs, and is suitable for linearly separable features, while Random Forest captures complex interactions and increases model stability through ensemble techniques. To complement the supervised learning approach, K-Means clustering is used to determine underlying patterns in the dataset. The Elbow Method is used for identifying the number of clusters, allowing us to cluster users into groups with similar usage patterns. The Random Forest model's feature importance also enhances interpretability by measuring the impact of each feature on prediction results.

D. Prediction and Recommendation Module

The prediction module combines results of regression, classification and clustering models to provide a holistic prediction of user risk. The continuous ESI score is discretised into risk levels for better understanding. The predicted risk and feature importance analysis are used to develop personalised recommendations to reduce eye strain. These recommendations are tailored to changeable factors like screen time, breaks, and screen distance, and are consistent with clinical practice guidelines.

E. Engagement Layer

The engagement layer is implemented through a Streamlit-based dashboard providing an interface for user interaction and results visualization. The users define their daily screen usage habits, after which the system performs live predictions and shows outputs such as the Eye Strain Index (ESI) score, risk classification, and relevant visualizations. Designing the interface in such a way allows presenting the results clearly to the non-technical audience.

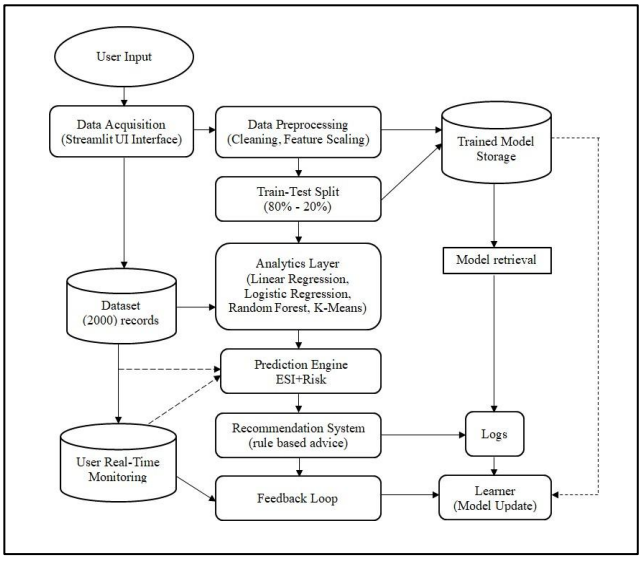

Fig 1.1 System Architecture Diagram

IV. Machine Learning Methodology|

1. Preprocessing Pipeline

It was verified that all nine features had complete data with no missing values to fill in. The continuous features underwent standardization based on the z-score normalization (using StandardScaler from scikit-learn). This step is necessary when using distance-sensitive algorithms (K-Means, Logistic Regression) and regularized models. In the case of using StandardScaler, all the features get scaled to zero mean and unit variance. Features with different units should be comparable when being learned by the algorithm [21].

To create the train-test split (train – 80%, test – 20%) for training our models and evaluating their performance, stratified random sampling technique with the specified random seed (42) was used. Stratified sampling technique was chosen due to the significant class imbalance (Low – 18.9%, Medium – 71.6%, and High – 9.5%) among three classes.

2. Linear Regression — Continuous ESI Prediction

Linear Regression was chosen as a baseline regression model for predicting continuous Eye Strain Index. Based on the assumptions of linear relations between the nine input features and output variable, Linear Regression computes coefficients by minimizing the ordinary least square function. As simple as it sounds,

Linear Regression provides a good baseline and is well-fitted for use in healthcare applications due to transparency [20]. For performance assessment, R^2 (coefficient of determination) and RMSE metrics were calculated. R^2 metric shows the proportion of ESI variance explained by the model, whereas RMSE represents average error in the original unit of ESI measurement (0–100).

3. Logistic Regression — Multinomial Classification

For binary classification problem, Logistic Regression was generalized to the multinomial setting. L2 penalty was used (with C = 1.0) to avoid overfitting and maximum of 1,000 iterations set as the criterion for early stopping. The result of Logistic Regression is probabilities of belonging to one of the three classes, which adds another advantage to its application as a probabilistic classifier in user-facing apps [21]. To measure the quality of predictions, overall accuracy as well as precision, recall, and F1-score for each of three classes was computed.

4. Random Forest — Ensemble Classification and Feature Importance

The Random Forest model, as developed by Breiman [18], builds up a forest of decision trees employing the bootstrapping procedure. In this approach, each tree is constructed based on a random selection of subsets of training data and evaluated using a random set of features per split. The resulting prediction is made through the vote of all the trees. This method decreases variance and is shown to be robust towards overfitting, making it especially appropriate for heterogeneous datasets and complex feature interactions [10], [11]. For this experiment, the model included 100 trees (n_estimators = 100). The feature importance was computed as the average mean decrease of node impurities (Gini importance) across all trees.

5. K-Means Clustering — Behavioral Segmentation

K-Means clustering was performed on the fully normalized dataset of features in order to discover archetypal users' behavioral patterns irrespective of labels related to the values of ESI. The algorithm iteratively assigns every observation to the closest centroid and refines them until convergence [21]. The optimal value of k (k = 3) was found using the Elbow Method, which involves plotting the within-cluster sum of squares by k and finding the inflection point of diminishing returns [21]. Principal Component Analysis (PCA)[19] was applied as a post-clustering visualization step, reducing the nine-dimensional feature space to two principal components to enable 2D cluster scatter plot representation while preserving maximal variance. The silhouette score was computed as a quantitative measure of cluster cohesion and separation.

V. MULTI-MODEL INFERENCE AND FALLBACK HEURISTICS

The main innovation offered by the current work concerns the implementation of the comprehensive machine learning pipeline for predicting the level of Digital Eye Strain (DES) based on behavior-driven data. While extensive research exists in the literature dedicated to clinical and machine learning-related aspects of Computer Vision Syndrome and applications of machine learning methods to ophthalmology, one should note that there still remains an open niche related to real-time, user-focused predictive inference frameworks integrating different analytical perspectives. The current study closes that niche.

A. Multi-Model Machine Learning Framework

One of the important innovations in the current work is the introduction of the multi-model machine learning framework. The system includes the Linear Regression for estimating the continuous ESI values and Logistic Regression & Random Forest for classification as well as K-means clustering for behavioral analysis.

B. Use of Non-Invasive Behavioral and Environmental Features

The proposed system uses non-invasive behavioural, physiological and environmental characteristics (i.e., screen time, blink frequency, break frequency, ambient light level, distance of view) as opposed to the more typical clinical imaging techniques like fundus photography or optical coherence tomography (OCT). Consequently, this system is both non-invasive and easy to deploy, allowing for use by the general public without special medical equipment or clinical oversight.

C. Preventive and Real-Time Prediction Approach

One of the greatest differences in this work is the use of a multi-model machine learning algorithm as a method for predicting eye strain index (ESI). In contrast to most predictive systems that rely on a single prediction method, the proposed system combines linear regression for continuous prediction of ESI, logistic regression and random forest for categorical predictions and k-means clustering to identify behaviours. This multi-faceted modelling strategy enables the proposed system to quantify severity of results and qualitatively describe user behaviours, providing a well-rounded picture of daily eye strain (DES) risk.

preventive health focus, not exclusively a diagnostic focus. Most currently available tools and studies are limited to determining conditions only after symptoms appear; therefore, this system is attempting to model the likelihood and severity of eye strain conditions that occur prior to developing into a clinically significant issue. By identifying potential risks associated with eye strain prior to developing into clinically significant conditions, an individual can initiate preventive care to mitigate long-term visual discomfort and lose productivity due to visual discomfort.

D. Deployment Through Interactive Dashboard

Another major contribution of this work is the incorporation of machine learning models as an interactive Streamlit-based dashboard. By implementing a deployment layer for the machine learning model, theoretical research will become practically usable by individuals as an application with the ability to provide real-time predictions for individual users and provide them with personalized recommendations. For example, the user will be able to input how much time he/she spends on screens on a daily basis, and the system will provide the user with a severity score based on his/her input data and offer the user suggestions on how to

decrease severity of eye strain. Furthermore, this research provides interpretability through feature importance analysis, thereby recognizing and validating that key behavioral characteristics, such as the frequency of breaks and the amount of time spent in front of a screen, significantly contribute to the severity of eye strain. This will not only validate the currently available clinical guidelines, but also foster user trust by providing transparency and understandability of the basis for how the model reached its conclusion.

VI. Implementation Challenges and Ethical Considerations

1. Data Privacy and Security

The protection of user data, especially information regarding habits along with other sensitive health information is a primary concern of deploying a DES predictive system. The dataset in this study is structured and anonymised; however, when deployed in the real world, the system is designed for regular use; therefore, it will capture information about a user on an ongoing basis (for example: sleeping patterns; screen-time; and potentially even physiological measures, for example: blink-rate). If this data were accessed without authorization or leaks were to occur, it would open the possibility that personal health information could be used incorrectly.

To combat this concern, it is recommended that strong encryption methods, such as AES-256, be put in place for all data at rest and during transmission. Compliance with legislation on data protection, specifically, the Digital Personal Data Protection Act (DPDP) of India is also needed. Techniques, like federated learning, can be used for privacy preserving purposes so that user data does not leave the user device but model changes can be shared. Therefore, privacy policies that are clear and easy to understand as well as obtaining consent from users are important actions required to trust an AI application will securely and ethically handle their data and comply with recommendations for AI in healthcare [20].

2. Algorithmic Bias and Fairness

Machine learning models rely heavily on the datasets used for training and any difficulty with or imbalance in the training dataset introduces possible bias into the trained machine learning model. For instance, data used to develop predictive models for DES can comprise differences related to age, sex, job type, and devices used. Discriminating models with respect to these differences will produce disparate model performance between these groups of users. For instance, if a predictive model used only student data to train the model, the model may not perform very well in predicting DES for older working adults. Discriminating model performance due to bias could lead to inaccurate predictions or the misclassification of high-risk individuals, resulting in less effective preventive recommendations. Therefore, it is important to ensure the diversity of datasets during the data collection process. Datasets collected should include a wide range of demographic and behavioral characteristics, suitable for use with many different types of users. Additionally, introducing retrospective updating of the currently trained model with respect to previously trained datasets and using explainable models can further enhance fairness. These considerations are also consistent with the concepts of transparency and accountability in medical AI systems [20], [21].

3. Digital Divide and Accessibility

The User access to digital infrastructures is a key aspect of how well the proposed system is going to work, presenting key challenges related specifically to the digital divide. People in low-income, rural, or low-resource settings might not have adequate access to high-performance devices or a stable, high-speed internet connection to be able to utilize a Streamlit-based dashboard, or the digital literacy needed to use such a tool. This will exclude segments of the population who are already vulnerable due to limited access to health care. Therefore, lightweight and offline-compatible versions of the predicted system should be created in order to make predictions without needing to be connected.

Furthermore, simplified user interfaces that support multiple languages along with visual cues will increase the usability of the proposed system for users who are not technically-oriented. The ability to integrate seamlessly with mobile device platforms, due to the fact that most mobile devices (especially smartphones) are readily available compared to desktops, will further increase the level of accessibility offered by the proposed system. Additionally, community-based distribution via educational institutions or workplaces has the potential to reduce accessibility barriers for many potential users. By addressing the aforementioned issues, the proposed system is in alignment with the principles of inclusive digital health, as emphasised in previous studies on devices worn on and outside the body and public health studies [16].

4. Model Reliability and Generalization

A primary challenge will be ensuring that trained models adequately generalize to real-world situations that extend beyond the artificial datasets used during training. Although current models display good performance statistics (R^2 = 0.8846, 87% accuracy), the actual behaviour of users in real life has much more variation and noise than users during testing. For example, issues such as inconsistent data entry, variability in external conditions, unmeasured confounding factors, etc., all have an impact on the accuracy of predictions being provided. A variety of ways by which predictive models can be improved include using cross-validation, regularisation, and Ensembling Methods. Furthermore, integrating real-time sensor data (e.g., detecting whether a person is blinking using a Web camera), will lessen the reliance upon self-reported measures thereby increasing the robustness and thus the reliability of predictions made by these systems. These processes align with significant aspects surrounding Predictive Healthcare Modelling [10, 14].

5. Interpretability and User Trust

In the context of health-related applications it is therefore

important that users are able to interpret the results from these models so that the users have confidence in the model predictions and will adopt health career practices consistent with the recommendations provided by the prediction system. When users are required to change their behaviour based upon what is suggested by the model, they may do so with some level of scepticism, and usage of the suggested behaviours could be compromised. This is especially relevant to users using DES (Digital Eye Strain) when they are able to receive classification results indicating that they are classified as being at a high risk for developing DES because of excessive screen use or inadequate breaks between uses of electronic devices. Therefore, in order to improve interpretability with regards to how well a prediction can be understood, descriptive feature importance analysis as taken from the Random Forest model are provided to users upon request showing users exactly how much of an influence each factor has when making the prediction that they will develop DES.Visual explanations, such as bar charts and personalized recommendations, can further improve user understanding. Providing confidence scores alongside predictions can help users gauge the reliability of outputs. This approach aligns with the emphasis on explainable AI in healthcare, where transparency is essential for clinical acceptance and user engagement [20].

TABLE II

IMPLEMENTATION CHALLENGES, ASSOCIATED RISKS, AND MITIGATION STRATEGIES

Challenge	**Associated Risk**	**Mitigation Strategy**
Data Privacy	Data breach, misuse of user information	Encryption, DPDP compliance, secure access
Algorithmic Bias	Unfair predictions across users	Balanced data, bias checks, retraining
Digital Divide	Limited access for some users	Offline mode, simple UI, mobile support
Model Reliability	Inaccurate real-world predictions	Real-data validation, monitoring, updates
Interpretability	Low user trust in results	Explainable outputs, feature insights

To successfully transfer from a design prototype to an operational application, it is necessary to resolve these administration issues. This will require the use of adequate privacy protection techniques, fairness-based modelling, accessibility to the model, proper verification and validating of the model output, and the ability to understand the model outputs. Addressing these issues ensures the DES model will function effectively from a technical perspective, as well as from an ethical standpoint.

VI. RELATED WORKS

A. Epidemiology and Clinical Understanding of DES

The past decade has brought increase clinical insight into dry eye syndrome (DES). Rosenfield [8] published one of the most cited reviews of computer vision syndrome, where he documented the accommodative, vergent and ocular surface mechanisms underlying symptom onset of computer users, as reported by 50% or more of computer users. The review highlighted the physiological stressor of work done at a proximity to visual distance and endorsed many behavioural changes including the 20-20-20 rule (20 minutes staring at a distance of 20 feet for 20 seconds) and ergonomic modification to address near work stressser's. Kaur et al. [2] provided a contemporary comprehensive systematic review of DES, including pre-COVID prevalence of 5-65%, and an increase during the COVID pandemic to 80-94% of the total population of DES. Symptoms of DES were described in the review about what comprehensive evidence pertaining to risks, literature including screen time, viewing distance, ambient light and uncorrected refractive error. Pavel et al.[7] classified DES as a contemporary ophthalmic condition of pathological process, describing ocular and musculoskeletal components of aetiology and a framework for managing patients digitally. Wolffsohn et al. [3] in the TFOS Lifestyle Report described a comprehensive literature review assessing the effects of digital environment upon the ocular surface. They showed that prolonged exposure to screens decreases blink rate, disturbs the stability of the tear film and worsens dry eye symptoms.Importantly, they noted that blue-light blocking interventions showed limited efficacy, and that behavioral modifications — particularly increasing break frequency and optimizing viewing distance — represent the most evidence-supported management approach. Talens-Estarelles et al. [9] conducted a controlled study on the 20-20-20 rule using bespoke software that monitored breaks via webcam in 29 symptomatic computer users over two weeks. Their findings demonstrated reductions in DES and dry-eye symptoms as assessed by the CVS-Q and Ocular Surface Disease Index (OSDI), lending empirical support to break frequency as a modifiable and clinically meaningful variable — consistent with its emergence as the strongest predictor in the present study's feature importance analysis.

B. Machine Learning in Ophthalmic Prediction

Rapid advances in the use of machine learning (ML) in the field of ophthalmology have resulted in several possible applications of ML - the analysis of retinal images, the detection of glaucoma, and the prediction of myopia progression. Al Marouf and colleagues [10] built an ML framework for predicting five of the most common eye diseases from symptom data, utilizing nine classifiers: Decision Tree, Random Forest, Naive Bayes, AdaBoost, and Logistic Regression with ranker-based feature selection. The results showed that ensemble methods, and specifically Random Forests, achieved consistently higher accuracy, as well as balanced precision and recall, than did the individual

classifiers; this finding supports the ensemble approach taken in the present study. A second example of the

for use in eye health classification tasks is Santos et al. [11], who developed an artificial intelligence-based system, the Optometry Random Forest, to automate the classification process for optometric diagnostic categories. The epoch-trained model achieved a classification accuracy of 97.17% and a specificity of 100%, confirming the ability of Random Forest architecture to successfully classify eye health. Further supporting the flexibility of tree-based prediction methods in health prediction applications is the autoRegressive integrated moving average (ARIMA) component used by Santos et al. for trend forecasting. Finally, Li et al. [12] used Random Forest to predict myopia progression over a period of five years in 2,740 primary school children and found that the six variables collectively contributed 77% of the total predictive weight and achieved a predictive accuracy of greater than 80%.In particular, their application of feature importance ranking to detect clinical predictors resembles the current study, where feature importance was both used to interpret the behavior of models and as a guide for implementing behavioral intervention strategies. Zhao et al[13] have expanded on this technique by applying it to predictive modeling of myopia occurrence based on 15 years of refractive longitudinal data. The authors compared Random Forest and XGBoost algorithms with classical regression models and concluded that Machine Learning algorithms demonstrated higher efficiency in nonlinear health predictions tasks. Chen et al. [14] have successfully implemented Random Forest Regression in predicting peak intraocular pressure within 24 hours of glaucoma patients and reported the result in the form of 5.248 and 2.291 MSE and RMSE respectively — the same output that can be provided by the regression part of this study. It is evident from their results that ML algorithms for continuous health parameter prediction have practical applications within ophthalmology. Elkholy and Marzouk [15] used CNN to predict eye diseases from OCT retinal images, thus proving the potential of deep learning approaches for this task. Although the methods of this paper and the image analysis in Elkholy and Marzouk are fundamentally different, they pursue the same goal of automatizing ophthalmologic diagnosis to facilitate faster and cheaper identification of eye diseases.

C. Clustering and Behavioral Segmentation in Digital

K-means is commonly used in wearable and digital health technologies studies for behavioral segmentation. According to Sabry et al. [16], who reviewed ML applications in wearable health monitoring, clustering techniques, especially K-means, are frequently employed in order to detect meaningful segments in unlabeled health data, which makes it possible to implement personalized intervention strategies for patients not having a labeled sample. Hijazi et al[17] used K-means clustering with the Elbow method in order to analyze heart rate variability patterns within wearable-based COVID-19 monitoring system. Thus, K-means proved to provide clinically meaningful segments in unlabeled health data despite the fact that Silhouette score values were relatively low. Similarly, Shahabi et al. [22] have used K-means clustering as an element of their digital health pipeline

significant potential of the Random Forest architecture

together with supervised classifiers.

VI. FUTURE DIRECTIONS

Although the proposed system shows high performance and reliability for DES severity prediction, there are still several areas for further improvement and real-life application of this solution. In the first place, as it has been mentioned before, real-time data acquisition techniques can be added to the solution in order to increase the efficiency and accuracy of its work. At the moment, the user input behavioral parameters are used, whereas computer vision techniques may allow for monitoring screen distance and blink rate automatically by means of web-camera-based approach.

Secondly, as it concerns the validation of the predictive model, the use of the clinical datasets from real life would be beneficial since the present dataset is based on clinical distribution. Thus, the deployment of the model on the data acquired from real users would help to make the model more universal and effective. Moreover, to address the class imbalance problem, techniques like Synthetic Minority Oversampling and cost-sensitive learning should be considered.

Moreover, the system can be transferred to mobile application format to increase user-friendly character of the platform. Integrating this system with some physiological data available from wearable devices such as smart watches can create more advanced digital health environment. In addition, it is worth considering advanced machine learning approaches such as gradient boosting or even deep learning to build more effici

VIII. SDG ALIGNMENT AND BROADER IMPACT

Regarding SDG 3, the system will contribute directly to improving digital eye strain (DES). It is crucial since, according to recent research, the problem is becoming increasingly common, especially among students and employees spending many hours in front of screens. DES affects up to 74% of those people. In general, such a percentage shows the extent of the problem, as it means that DES is one of the significant health burdens today.

By contrast with conventional ways of assessing the presence of DES (relying upon retrospective reporting using questionnaires), this project will help in the prevention of eye strain. Using the algorithms for calculating the Eye Strain Index (ESI), the system will allow identifying risky behavior and preventing it at an early stage in order not to develop more serious disorders associated with long-term screen exposure. Thus, the solution will be able to reduce the rate of diseases related to DES in the context of goal 3 of the SDGs.

Identifying modifiable risk factors (especially, break frequency, screen time) will help formulate specific recommendations to improve behavior and reduce strain. According to studies, DES can be managed successfully by applying clinically validated behavioral changes. For instance, the frequency of breaks should

meet the requirements of the 20-20-20 rule and the screen viewing distance should be optimized. As a result, the implementation of recommendations into the solution will help prevent health problems.

In addition to the above-discussed health benefits, the development will contribute to SDG 3 in the sense of accessibility. Unlike the conventional approach which requires specialized medical devices, the project offers an easy-to-use, non-intrusive application to assess eye strain. It means that the system is highly convenient for broad distribution across regions characterized by limited access to ophthalmology services. Therefore, the tool contributes to the achievement of another goal of SDG 3.

Last but not least, the project incorporates the explainability of its algorithms, including Logistic Regression and Random Forest, which are important in terms of user experience. The feature increases the trust in predictions and allows interpreting outcomes in the healthcare field. Users can understand why particular results of calculations are received and what measures should be taken to avoid risks.Beyond its primary alignment with SDG 3 (Good Health and Well-being), the proposed system also contributes to broader sustainable development objectives. By promoting healthier digital habits among students and professionals, it supports SDG 4 (Quality Education) through improved concentration and learning efficiency. Additionally, the integration of machine learning for preventive healthcare reflects SDG 9 (Industry, Innovation, and Infrastructure), while its accessible, low-cost design helps reduce disparities in health awareness, aligning with SDG 10 (Reduced Inequalities). Furthermore, by encouraging responsible screen usage patterns, the system indirectly advances SDG 12 (Responsible Consumption and Production).

VI. CONCLUSION

In this paper, we propose an ML-based tool for DES severity prediction using nine behavioral and environmental features of the user. With a dataset containing 2,000 records, we used Linear Regression and received R^2=0.8846, RMSE=4.73 for predicting the continuous ESI value; Logistic Regression provided an accuracy of 87.00%, while Random Forest showed 83.25% accuracy for classifying users based on three severity levels. Using K-Means clustering algorithm, we detected three distinct user behavioral categories. Break frequency (21.92%), along with screen time (19.83%), was found to be the most significant predictors for severity classification and agrees with known clinical prevention recommendations.

A custom-built Streamlit dashboard converts model outputs into understandable metrics available for non-clinical users, thus serving as an example of practical implementation of DES preventive measures. Taking into account the increase in worldwide digital devices use (up to 69-74% of population suffering from DES among those who use electronic screens [4], [5]), tools allowing to evaluate personal severity level and adjust behavior accordingly can become a helpful addition to clinical measures.

Future research should involve SMOTE method for solving the problem of class imbalance, collecting and validating the tool on real-world prospective data, implementing computer vision-based real-time blink detection using webcam, developing a mobile application, and exploring novel ensemble and deep learning techniques.

VII. REFERENCES

[1] J. S. Wolffsohn et al., "TFOS lifestyle: Impact of the digital environment on the ocular surface," Ocular Surface, vol. 28, pp. 213–252, 2023.

[2] K. Kaur et al., "Digital eye strain — A comprehensive review," Ophthalmology and Therapy, vol. 11, no. 5, pp. 1655–1680, 2022.

[3] J. S. Wolffsohn et al., "TFOS DEWS II diagnostic methodology report," Ocular Surface, vol. 15, no. 3, pp. 539–574, 2017.

[4] F. Ccami-Bernal et al., "Prevalence of computer vision syndrome: A systematic review and meta-analysis," Journal of Optometry, vol. 17, no. 1, p. 100482, 2024.

[5] D. A. León-Figueroa et al., "Prevalence of computer vision syndrome during the COVID-19 pandemic: A systematic review and meta-analysis," BMC Public Health, vol. 24, no. 1, p. 640, 2024.

[6] A. Akiki et al., "Computer vision syndrome: A comprehensive literature review," Taylor & Francis Online, 2025. doi: 10.1080/20565623.2025.2476923.

[7] I. A. Pavel et al., "Computer vision syndrome: An ophthalmic pathology of the modern era," Medicina, vol. 59, no. 2, p. 412, 2023.

[8] M. Rosenfield, "Computer vision syndrome: A review of ocular causes and potential treatments," Ophthalmic and Physiological Optics, vol. 36, no. 5, pp. 502–515,

2016.

[9] C. Talens-Estarelles et al., "The effects of breaks on digital eye strain, dry eye and binocular vision: Testing the 20-20-20 rule," Contact Lens and Anterior Eye, vol. 46, no. 2, p. 101744, 2023.

[10] A. Al Marouf and M. M. Mottalib, "An efficient approach to predict eye diseases from symptoms using machine learning and ranker-based feature selection methods," Bioengineering, vol. 10, no. 1, p. 25, 2023.

[11] L. F. F. M. Santos et al., "Artificial intelligence-driven diagnostics in eye care: A random forest approach

[12] J. S. Wolffsohn et al., "TFOS lifestyle: Impact of the digital environment on the ocular surface," Ocular Surface, vol. 28, pp. 213–252, 2023.

[13] K. Kaur et al., "Digital eye strain — A comprehensive review," Ophthalmology and Therapy, vol. 11, no. 5, pp. 1655–1680, 2022.

[14] J. S. Wolffsohn et al., "TFOS DEWS II diagnostic methodology report," Ocular Surface, vol. 15, no. 3, pp. 539–574, 2017.

[15] F. Ccami-Bernal et al., "Prevalence of computer vision syndrome: A systematic review and meta-analysis," Journal of Optometry, vol. 17, no. 1, p. 100482, 2024.

[16] D. A. León-Figueroa et al., "Prevalence of computer vision syndrome during the COVID-19 pandemic: A systematic review and meta-analysis," BMC Public Health, vol. 24, no. 1, p. 640, 2024.

[17] A. Akiki et al., "Computer vision syndrome: A comprehensive literature review," Taylor & Francis Online, 2025. doi: 10.1080/20565623.2025.2476923.

[18] I. A. Pavel et al., "Computer vision syndrome: An ophthalmic pathology of the modern era," Medicina, vol. 59, no. 2, p. 412, 2023.

[19]M. Rosenfield, "Computer vision syndrome: A review of ocular causes and potential treatments," Ophthalmic and Physiological Optics, vol. 36, no. 5, pp. 502–515,

2016.

[20] C. Talens-Estarelles et al., "The effects of breaks on digital eye strain, dry eye and binocular vision: Testing the 20-20-20 rule," Contact Lens and Anterior Eye, vol. 46, no. 2, p. 101744, 2023.

[21] A. Al Marouf and M. M. Mottalib, "An efficient approach to predict eye diseases from symptoms using machine learning and ranker-based feature selection methods," Bioengineering, vol. 10, no. 1, p. 25, 2023.

[22] L. F. F. M. Santos et al., "Artificial intelligence-driven diagnostics in eye care: A random forest approach.