



OPEN GAINSeq: glaucoma pre-symptomatic detection using machine learning models driven by next-generation sequencing data

Muhammad Iqbal^{1,4}, Arshad Iqbal^{1,4}✉, Humaira Ayub², Maqbool Khan¹, Naveed Ahmad³, Yasir Javed³ & Mohammed Ali Alshara³

Congenital glaucoma, a complex and diverse condition, presents considerable difficulties in its identification and categorization. This research used Next Generation Sequencing (NGS) whole-exome data to create a categorization framework using machine learning methods. This study specifically investigated the effectiveness of decision tree, random forests, and support vector classification (SVC) algorithms in distinguishing different glaucoma genotypes. Proposed methodology used a range of genomic characteristics, such as percentage variation, PhyloP scores, and Grantham scores, to comprehensively understand the genetic pathways that contribute to the illness. This investigation showed that Decision Tree and Random Forest algorithms consistently performed better than earlier techniques in identifying congenital glaucoma subtypes. These algorithms demonstrated outstanding accuracy and resilience. The findings highlight the capacity of machine learning methods to reveal complex patterns in NGS data, therefore improving the proposed comprehension of the causes of congenital glaucoma. Moreover, the knowledge obtained from this research shows potential for enhancing the accuracy of diagnoses and developing tailored treatment approaches for afflicted people.

Keywords AI, Deep learning, Machine learning, Genome, DNA, Glaucoma

Glaucoma is an aggregation of eye diseases that stems from neurological impairment of the optic nerve affecting vision in either one or both eyes¹. The optic nerve gradually deteriorates and develops a blind spot in the vision. This damage emanates visual field defects leading to stone blindness if left untreated. With this severity, by 2010, 60.5 million individuals were affected by glaucoma², and projected to affect over 111 million around the globe by 2040³. This rapidly increasing disease primarily affects the elderly population, therefore, effective healthcare strategies require early prediction for the prevalence of this disease. This congenital disease is developed through hereditary, environmental, and several other factors⁴. Glaucoma development is influenced by several factors including specific genes that significantly contribute, to causing this congenital disease, worldwide and within individual populations⁵.

Genetic analysis of glaucoma is a challenging task due to genetic heterogeneity, population-specific variations, complex inheritance patterns, limited genetic studies, phenotype variability, and technological limitations⁶. The challenge with the illness arises from whether it is heterogenic, whereby many genes cause the same medical type or whether one gene result in diverse clinical features⁷. Still, clinical genetic genealogy and association studies have proven to be quite helpful in categorizing pathogenic mutations and 'low frequency' and 'high frequency' polymorphisms genetic changes⁸. These features have been discovered to happen at higher charges in suffered individuals in contrast to control inhabitants of non-affected, wholesome people⁹. Genetic predisposition, the likelihood of developing the congenital disease is a primary contributing factor to glaucoma. However, environmental variables, i.e., high intraocular pressure, also contribute to developing the disease¹⁰. Emergent glaucoma and its subtypes are far from guaranteed for every country and race¹¹.

¹Sino-Pak Center for Artificial Intelligence (SPCAI), School of Computing Sciences, Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology (PAF-IAST), Haripur 22620, Pakistan. ²Department of Biological and Health Sciences, Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology (PAF-IAST), Haripur 22620, Pakistan. ³College of Computer and Information Sciences, Prince Sultan University, 11586 Riyadh, Saudi Arabia. ⁴Muhammad Iqbal and Arshad Iqbal contributed equally to this work. ✉email: arshad.iqbal@spcai.paf-iast.edu.pk

Linkage with in addition 80 genes by the genome-wide association studies (GWAS) to primary open-angle glaucoma (POAG), are the widespread type^{12,13}. It shows that these genes cause differences in the health status of optic nerve cells, drainage, and aqueous humor formation. The former analyses revealed that POAG is polygenic and characterized by multilayered inheritance together with numerous gene and ecological factor interactions. Nevertheless, identifying these genetic predispositions are useful in creating early-screening measures and/or individual-specific treatment modalities to prevent blindness¹⁴. Gene mutations are not the special cause of inherited eye problems. However, some forms of alterations recognized as epigenetic factors that interfere with Gene expression except for affecting the DNA coding itself, may in principle be involved¹⁵. Additionally, the interaction of genes and environment influences the severity and progression of the disease. For instance, glaucoma may only develop in people with the characteristic of a particular POAG susceptibility gene in case of long-term high intraocular pressure caused by other diseases¹⁶.

Conventionally, diagnosing congenital glaucoma relies on vigilant parental analysis, well-thought healthcare expert analysis, and comprehensive clinical examinations¹⁷. These well-established methods led to significant success. However, this approach often involves uncertainty as critical time lost in this process may lead to the onset of blindness. Therefore, novel methods are required to diagnose this congenital disease at the outset which can assist in the pre-symptomatic treatment of this disease. The convergence of Next-Generation Sequencing (NGS) and Artificial Intelligence (AI) technologies changes the glaucoma diagnostic methods that can assist in identifying it at the gene level¹⁸. It paves the way to identify new possibilities of detecting it at the earlier stage tailoring treatments to individual patients, and envisioning a future with improved vision.

NGS is a technology set that has transformed genomics studies by enabling high throughput and cost-effective RNA and DNA sequencing¹⁹. The application of NGS is extended to identify genetic mutations in diseases e.g., cancer and glaucoma, identification of pathogens, livestock breeding, genetic variation and evolutionary relationships, and environmental genomics²⁰. This technology facilitates the identification of genetic factors linked with glaucoma for early detection, personalized treatment plans, and innovative therapeutic approaches²¹. NGS significantly identifies the mutation associated with POAG and screening of various genes e.g., Myocilin (MYOC), and Optineurin (OPTN), while Buphthalmos (GLC3A) and Cytochrome P450 Family 1 Subfamily B Member 1 (CYP1B1) are associated with this congenital disease²².

AI is remodeling disease identification at the gene level collaborating with the newly developed NGS technology²³. It deals with rapid sequencing of the entire genome which generates enormous amounts of data that is important for identifying genetic variation associated with disease. Therefore, advanced computational algorithms are required to analyze the data's sheer volume and complexity effectively. With the ability to process large datasets, machine learning models are trained to identify patterns and correlations in genetic data²⁴, leading to the prediction of disease susceptibility, progression, and response to treatments. Machine learning algorithms can be used for the discovery of the precise genetic markers linked to various types of glaucoma which are significantly assisting in early diagnosis that involves identifying people at risk of developing a condition, pre-symptomatically.

Vast quantities of data such as retinal images, medical records, and even genetic information are used to identify complex and minute quantities of anomalous patterns. Machine learning models with their capacity to recognize intricate patterns in data, then examine this map and find tiny genetic variants associated with congenital glaucoma. The disease's minuscule irregularities in the trabecular meshwork, subtle alterations in retinal blood circulation, or even nuanced variations in tear composition can be identified using learning algorithms. This keen observation allows for quicker and more precise diagnosis, frequently detecting issues before visible symptoms appear. Every instance of congenital glaucoma is distinct, and a standardized approach to therapy is no longer adequate. AI and NGS may facilitate specific treatment plans by comprehending the distinct genetic profile of each youngster. These programs are adjusted to optimize the effectiveness of therapy while reducing adverse effects, providing each individual with a personalized path to recovery. AI and NGS can use a more comprehensive knowledge of the specific risk factors and pathways of disease development related to congenital glaucoma to accurately forecast the probable trajectory of the illness. This enables the optimization of long-term management techniques, allowing families to make well-informed choices about their child's care, and an in-depth awareness of their unique path.

Personalized medicine involves customizing treatment approaches according to an individual's distinct genetic profile, to optimize treatment effectiveness while avoiding adverse reactions. Predictive insights which are gaining an understanding of the patterns of development associated with certain genetic markers may enhance the accuracy of prediction and guide long-term strategies in managing the condition. With the existing trend in heredity studies, early diagnosis of diseases with, there is hope of managing glaucoma and preventing vision loss. The advancements made in identifying gene markers for glaucoma coupled with the development of artificial intelligence-based diagnostic model is of advantage to scholars and enthusiasts in the field. The proposed framework (GAINSeq) has implications which suggest the above included some significant advancements in ophthalmology and other further implications which concern the comprehensive parameters of medicine and health at large. The contributions we have made are listed as follows,

- A GAINSeq framework is proposed for a proficient glaucoma eye disease's identification with significant accuracy.
- In the proposed framework, machine learning models are employed to differentiate various genetic traits associated with glaucoma. This is achieved with the help of the algorithmic model's ability to unmask hidden trends and patterns in the Next Generation Sequencing (NGS) data.
- The purpose is to take proactive precautions to recognize individuals with a higher genetic inclination outcome in early therapy and avoid visual loss.

- The intended aims are to utilize prevention strategies that ensure that people who have high genetic disposition rate are given early treatments to help prevent instances of visual loss. A pre-symptomatic method is applied because the patients can avail accurate detection and subsequent effective treatment to avert cheap expensive treatments at later stages reducing overall costs for the individuals and the health care systems.
- The goal line is therefore to discover what constraints or restrictions there are to the application of progressive analysis particularly with regard to eye diseases using genetic data.

Literature review

Algorithmic ability, also termed as methodological, is defined as the creation of algorithms using data and analysis derived from data science, mathematics and computer science²⁵. This expertise is used for data analysis to steer into the appropriate algorithms that are efficient in decision making in the health domain^{26,27}. The machine learning algorithms meant for glaucoma screening and the segmentation and classification of fundus image collection is described²⁸. In cases of glaucoma, computational intelligence is normally collaborating with human intelligence²⁹. Algorithms are often implemented in particular types of health care services, for example, in ophthalmology clinics³⁰. Along with the human ophthalmologist, the application with Food and Drug Administration (FDA) approval integrated with AI diagnoses diabetic retinopathy³¹. The research illustrates the efficient use of machine learning techniques in the diagnosis of diabetic retinopathy³². A study suggested a model, with the necessary improvements, that can be used in clinical decision making for various conditions such as lung infections and diabetic retinopathy³³. However, other applications have also been created that utilities AI algorithms to diagnose glaucoma via fundus photographs³⁴. Some of portable fundus photography cameras have been developed in this context. These gadgets are movable, cheap and are developed in such a way that one does not need to be a professional in computers in order to use the gadgets. Also, a study done on telemedicine in glaucoma showed that if more mechanically controlled machines are used, then more precise and fair results would be forthcoming especially since the machines can be operated by less skilled human beings in those distant sites³⁵.

Glaucoma is one of the leading contributors to vision loss, affecting about 15% to 20% of mankind, which requires premature identification³⁶. Incorporating the AI-based approaches of glaucoma detection via fundus imaging to screening mechanisms has the potential to eliminate some of the drawbacks of human involvement including bias and fatigue while at the same time allowing for large scale screening at affordable costs³⁷. Such creativity can assist in taking critical ophthalmic services to areas that are out of reach in remote rural areas especially in countries that are deficit in the number of proficient and skilled ophthalmologists. The use of Automated image explanation's real-life practice in front-line care of eyes is expected to grow significantly in the future primarily due to its use in the screening as well as in referral decision and patient specialization³⁸. Thus, even in countries and facilities with apparent excess of resources such as the National Health Service (NHS) in the United Kingdom, application of AI for referral improvement contributes to addressing the problem of the excess demand for outpatient planning and the number of mistakes in referrals³⁹.

In a clinical scenario, the diagnosis of glaucomatous optic nerve impairment utilize optical coherence tomography (OCT) imaging, visual field with Standard Automated Perimetry (SAP) and clinical assessment of the optic disc (OD). In other words, to arrive at the definitive diagnosis of glaucoma, one has to use tests with high sensitivity and specificity⁴⁰. Probably fundus images can be used in populace-based glaucoma screening as they are comfortable and often used for OD differentiation which as it makes sense helps in glaucoma identification⁴¹. Many studies investigate and process the correct detection of nerve fiber layer defects (NFLDs) in glaucoma prognosis through retinal fundus images are reviewed in the study⁴². This paper outlines the use of architectural techniques, computational analysis techniques, and morphological analysis to determine the attenuation of neuropathic fiber⁴³. The authors developed a fresh OCT dataset and presented a new CNN, termed TurkerNeXt, and according to the conclusions and classification results, the TurkerNeXt model exhibited remarkable classification performance⁴⁴.

The investigations demonstrate that timely detection of the retinal nerve fiber layer (RNFL) abnormalities is important for clinical practice and give an idea about the challenges in precise identification of pathogenic and referential changes in the RNFL⁴⁵. This procedure is straightforward and inexpensive and has shown a prospect to identify cases in some groupings. However, the amount of experts are mandatory to grade pictures manually as well as the variations between completely different observers and inside the similar observer, along with circumstances resembling extreme refractive errors. Hence, the practice of using AI-enhanced algorithms for grading pictures to improve the diagnosis of glaucoma analysis of the synthesized image cluster is an exciting idea⁴⁶. In extension with this, artificial intelligence in ophthalmology is used to address common problems such as diabetic retinopathy, age-related macular degeneration, glaucoma, prematurity retinopathy, congenital glaucoma, and retinal vein occlusion⁴⁷.

Sinthanayothin et al. described the method for the recognition of OD, fovea, and blood vessels from hue fundus images⁴⁸. Since then, numerous works have been able to reproduce the segmentation and has considered it an indispensable requisite for using algorithms for diagnosing glaucoma based on fundus images. In the long course, these images were utilized as an input set in the diagnostic assessment of glaucoma with the help of AI approaches⁴⁹. In the pursued article the author suggests an approach for the identification of the optic disc from the retinal images. This approach employs a classifier model where the model is trained with a structured learning methodology for creating an edge map of the OD⁵⁰. After that, the layout of the OD is identified by threshold and circular Hough transform for detecting the outer boundary of the OD. In three public-available datasets, the method is rated and corroborates encouraging results, extracting the morphology-based and the template-based approaches⁵¹. The studies demonstrate how important is to localize and segment the OD with high accuracy in order to diagnose their eye deformities including glaucoma and diabetic retinopathy⁵². The

specified strategies provide rather positive outcomes; these strategies can potentially improve the correctness of OD segmentation⁵³.

Using the regions of analysis of the optic disc and the optic cup from the retinal fundus picture, a new approach to the treatment of glaucoma is provided. The support vector machine is used as classifier while segmenting the image and extracting features from the image was performed by simple linear iterative clustering (SLIC)⁵⁴. A study shows the implementation of the Histogram of Oriented Gradients (HOG) and Artificial Neural Networks (ANN) to extract textural features from fundus images, achieving high classification accuracy⁵⁵. The proposed approach ensures better capability of distinguishing between persons with glaucoma and those without it because it separates the optic disc from the cup more accurately⁵⁶. To evaluate the reliability of the deep learning system to diagnose Glaucomatous Optic Neuropathy (GON) using fundus photographs the authors employed the algorithm that provided the results with high sensitivity and specificity level⁵⁷. Furthermore, the study focused on ways of enhancing and calibrating the algorithm and the methodologies used to assess the parameters that went into categorization⁵⁸. Despite, the study has the strength of having a large sample size, however, it is limited to the hospitals in China⁵⁹. For the assessment at particular demographic groups and the number of ophthalmologic illnesses where the particular algorithm may prove beneficial, further study is needed.

More economical forms of detection are suggested, for example, Optical Coherence Tomography and Heidelberg Retinal Tomography⁶⁰. The technique involves as identifying and segmenting unique features such as the cup to the disc ratio and number of blood vessels in the digital fundus images⁶¹. The variables used to train and test the neural network classifier makes the distinction of the images as normal or showing symptoms of glaucoma efficient. The results shown in the paper prove that the identified traits are significant in clinical diagnosis of glaucoma because the derived system results insensitivity of upto 100% and specificity of upto 80% in the process of categorizing glaucoma automatically⁶². These are the findings which suggest that the suggested approach is plausible, cost efficient and effective in the diagnosis of glaucoma and can be a convincing substitute to other expensive diagnostic strategies. Diagnosing glaucoma, which is responsible for significant irreversible blindness, with the help of hyperanalytic wavelet processing and a Particle Swarm Optimization (PSO) technique⁶³. The main goal of the suggested approach is to facilitate the early identification of glaucoma with screening programs in order to prevent the development of the diseases².

The segmentation of the fundus images and the use of the structured learning may diagnose glaucoma. However, the approaches have achieved an accuracy of upto 95% with a positive diagnosis⁶⁴. Pre-processing is one of the phases that are part of computer-assisted identification of various features involved in the diagnosis of eye pathology, of which identification of the optic disc is a recognition phase. Another approach to successfully locating the optic disc may require applying structured learning in a way that involves training a classifier model in the shapes of the optic disc⁶⁵. The edge map, which was previously obtained, is then threshold so as to create a binary image. This binary image is then processed using the circle Hough transform so as to estimate the boundary of the optic disc by fitting a circle. The focus on the algorithm and its efficiency will be discussed further in the context of the evaluation of the said publicly available datasets for OD segmentation⁶⁶.

The fundus images are utilized to train a neural network and the overall efficiency of the method in diagnosing glaucoma from the 1426 cases is at 98% and is tested in terms of sensitivity and specificity^{67,68}. Another comprehensive deep learning network for evaluating the extensiveness of the injuries of the optic nerve resulting from glaucoma using the fundus images, is presented⁶⁹. Other studies focus on deep learning algorithms to detect and quantify the glaucomatous damage from images of optic disc⁷⁰. The method uses training via the quantitative spectral-domain OCT data and proved to have a strong relation in concordance with the actual observations⁷¹. Considering the results of glaucomatous damage detection, the enhanced performance by the algorithm is revealed particularly for persons of African descent thus suggesting that this method may be useful in performing remote ophthalmology programs or settings where OCT might not be used. Likewise, the CNN model meant for the detection of glaucoma with the help of retinal fundus images is developed with an average accuracy of 98%. Only 13 percent accuracy rate and is capable of sustaining this level of performance through as many cycles as noted in the study⁷².

Although tomographic techniques are more valuable in the visualization of structural changes within the eye, the unidentified interactions among the multiple inherited factors that constitute the risk and progression of glaucoma needs to be studied. This could be due to the fact that the disease has hereditary characteristics and since it may present in dynamism where the type of glaucoma may also present various genetic features for instance POAG and ACG which exhibit various genetic predisposing factors, further research is needed on the genetic aspect of the illness. Therefore, determining specific genetic characteristics related to certain types of glaucoma is significantly useful due to the possibility of recognizing the disease pre-symptomatically to perform pertinent prevention efforts. More specifically, doctors can easily intervene before the current symptoms of glaucoma in these patients and ensure that those most at risk of these complications receive appropriate treatment, patient outcomes can be improved and sightlessness prevented.

Proposed glaucoma identification scheme

In the proposed framework, the DNA and gene testing tools with modeling are combined to perform the glaucoma analysis with various machine learning models to identify the disease at early stages. The proposed framework employs a database of genetic sequences of several patients with different types of genetic eye disease. The proposed work is focused on the details of the genotypes responsible for different forms of glaucoma to be identified and categorized. To enhance the outcomes, classification models are employed such as SVM, Random Forest and Decision tree during the development of the identification system as depicted in Fig. 1. This approach proves useful in handling genetic information and makes it possible to identify biomarkers that are likely linked to the susceptibility of glaucoma progression over time.

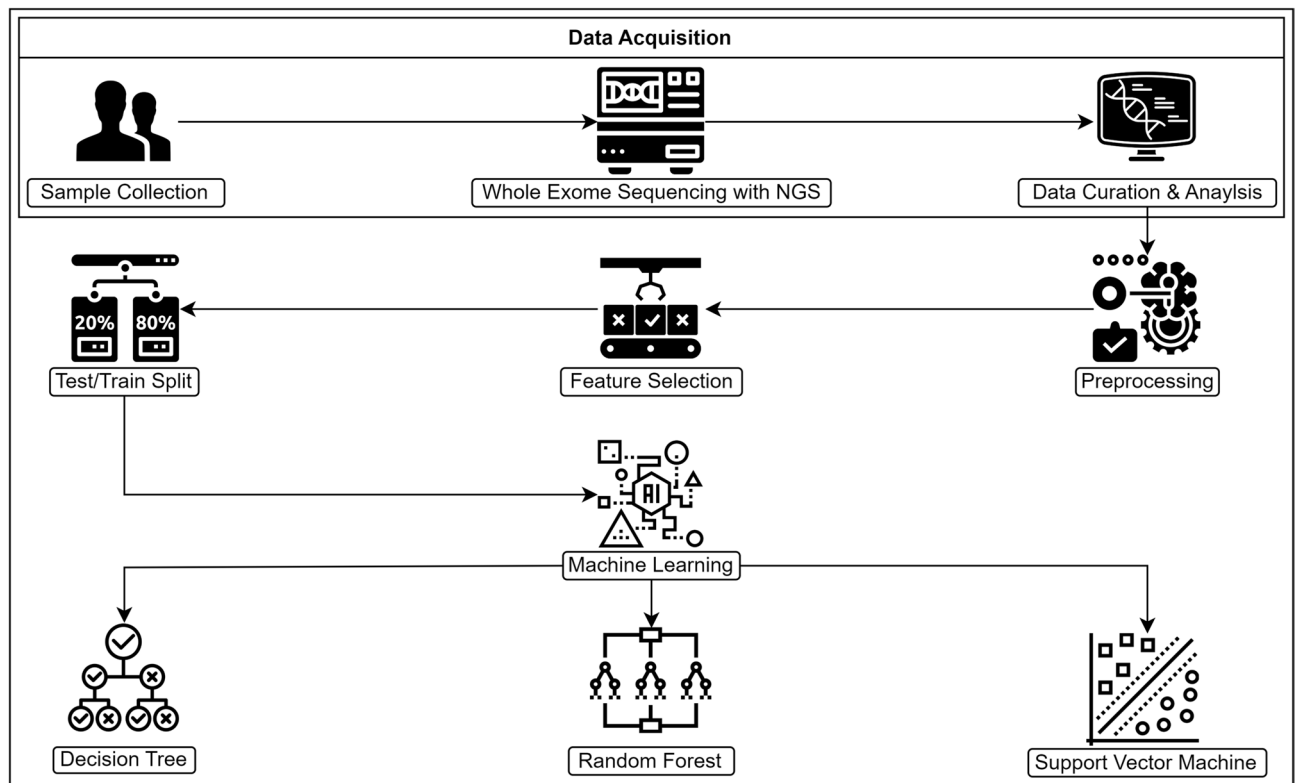


Fig. 1. The overall setup of the proposed framework for identification of glaucoma on the NGS data.

Machine learning models are trained and validated based on two different sets of data split with significant ratios. The validation of the trained models is required to generalize the impact of the models on the number of patients with diverse genetic sequencing data. Genetic information is analyzed to understand the molecular nature of diseases i.e., glaucoma, that depends on numerous genetic markers existing in human genes. The key information includes single nucleotide polymorphism (SNP), insertions, deletions, references, Grantham score, and PhyloP. In other parts of the genome, the SNPs may be substituted by a single nucleotide. These particular SNPs are considered to be associated with the risk of occurrence and the progression of the disease. Complex genetic variations exist that dictate fairly challenging problems to elucidate how these variations impact complex diseases such as glaucoma. In case, when changes exist the probability calculations call for more elaborate computational methods that are highly effective. These approaches distinguish between different genetic states at a fundamental level including, for instance, homozygous and heterozygous alleles. The former alleles represent those genes that are the same as the genes inherited from both parents while later alleles represent the genes that are different from the genes of both parents, but inherited from them. Considering the fact that both homozygous as well as heterozygous admixture genotypes are involved in the determination of the subject's genetic predisposing factor in glaucoma, it is necessary to define homozygous and heterozygous clearly. This distinction is particularly significant, especially while assessing the inheritance structure and allelic combination that defines the prevalence of the condition. For example, specific genetic differences involve situations where the expression of a gene is masked till the recessive gene from both the parent's surface with the trait.

To display the disease phenotype, one has to be a carrier of two alleles of a gene which are responsible for a certain disease. The other type is the prime model, in which carrying a variation allele referred to as heterozygous is enough to raise the probability of the illness. The methods of categorization are presented in order to provide the tools for classifying genetic information based on certain parameters. This allows the identification of clusters and relationships which are important to several features concerning the vulnerability and onset of glaucoma. This makes it easier to sort all individuals depending on their genetic past and thus allows researchers to determine the possibility of getting glaucoma depending on the genetic pleomorphism homozygous as well as the heterozygous choices made.

Data collection

Next-generation sequencing (NGS) is a set of genomic sequencing technologies that have been at the forefront of making DNA sequencing much more efficient and for a much higher throughput. This has resulted in a lot of developmental changes, especially regarding genomes and clinical analysis. Present-day technologies in NGS equipment are capable of processing millions to billions of DNA fragments at a time thereby increasing the speed and decreasing costs. This is different from the Sanger Sequencing technique which concerns the sequencing of one portion of the DNA strand at a time. While it is possible to sequence millions to billions of DNA fragments at one time, compared to the Sanger sequencing technology that can only pile-sequence

thousands of DNA fragments, next-generation or massively parallel sequencing systems exist. High-throughput sequencing equipment is employed, and the major data analyzed within this research are sourced from the NGS approach. The NGS technique is one of the most revolutionary advancements made in the field of genomics enabling the sequencing of large DNA molecules in terms of time and cost.

The process currently used in NGS involves the simultaneous clonal amplification of the DNA fragments and the identification of each fragment in terms of the sequence of nucleotides using fluorescence-tagged nucleotides. The approach is based on the sequencing by synthesis (SBS) scheme that forms a methodological basis for NGS. The method typically includes (i) Library development (ii) Cluster generation (iii) Sequencing (iv) Imaging, and (v) Data analysis. The next-generation sequence includes the formation of a DNA Library. To prepare the required fragment of DNA a procedure of fragmentation and addition of adapters to the fragment is performed. These adapters are used in amplification and incorporation of DNA fragments and further sequencing of these fragments which include specific sequences that are unique to the used sequencing platform. Afterward, these DNA fragments are attached to firm support that can include a slide of glass or flow cell by ways of cluster generation, or emulsion PCR. This is a preliminary stage of performing the clustering which is used in the generation of the clusters. Thus, the process includes the formation of clusters which are composed of several copies of a similar DNA piece, instead of single-stranded circular DNA molecules and linear chromosomes.

The sequencing strategy is based on Sequencing by Synthesis (SBS) which includes the incorporation of the fluorophore-labeled nucleotides into the DNA chain. The phenomenon referred to as the hierarchical organization is called sequencing it implies that the respective set of actions is grouped into different categories, where each of the actions is done in respect of a certain group of objects. The sequencing device employs the use of a photo-optical detector that can distinguish the fluorophore label for the complementary nucleotide of the incorporated nucleotide. The overarching method used to describe NGS is sequencing by synthesis, although this is implemented with some differences in chemical reactions and nucleotide recognition platforms across the NGS systems. Then, a surface image is taken to capture the fluorescence associated with the integration of the cluster after every cycle of integration. At the end of the specific imaging procedure, it is possible to obtain nucleotide sequence at any given position within the DNA fragment. To obtain further information from the data produced by the NGS equipment, such data needs to be processed and analyzed with the help of tools and methods linked to Bioinformatics. The following processes are included: This is through base calling which translates the sample into nucleotide sequences, read alignment which superimposes the sequenced reads, variant calling followed by downstream analysis which gives the meaning and interpretation of the sequenced information derived from the data.

NGS begins with a sample of interest, from which small DNA fragments, referred to as reads, are extracted. The readings are then sequenced using a NGS apparatus. Once data is processed and the assessment is done, it reconstructs the DNA sequence, concludes about the genetic difference and then acquires the biological nutrient. In the last few years NGS has enormously affected many areas including genetics, cancer detection, infectious disease studies and personalizing treatment in both the basic research and clinical applications. It can be stated that NGS plays a significant role in understanding the molecular mechanisms of diseases, comprehending the multifactorial genetic characteristics, and contributing to the development of the precision medicine projects because of its highly versatile, applicable, and efficient nature.

Dataset characteristics

This study was conducted in accordance with the ethical standards of the Declaration of Helsinki and was approved by the Institutional Ethics Committee of the Department of Biological and Health Sciences, Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology (PAF-IAST/2021/08). Individuals clinically diagnosed with glaucoma were recruited after a detailed clinical evaluation. Before participation, written informed consent was obtained from all individuals. Blood samples were collected using ethylenediaminetetraacetic acid (EDTA) vacutainers, after which genomic DNA was extracted and subjected to Next-Generation Sequencing (NGS) for further analysis. All genetic and clinical data used in this study were de-identified prior to analysis to ensure patient anonymity.

In this study, the employed dataset composed of fifteen participants. Genetic data and clinical data in seventy-seven aspects related to the examination of glaucoma are used⁷³. Figure 2 shows several important characteristics with correlation matrix Heatmap which visualizes the pairwise correlations between different features in the dataset. The color scale ranges from -1 to 1, where red indicates a strong positive correlation, blue indicates a strong negative correlation, and shades closer to white represent weaker correlations. The matrix helps identify linear relationships between variables. For instance, a high correlation between the feature Grantham Score and the PhyloP Score suggests that these features positively affect each other. While the feature Novel allele mean QV and the SNP Frequency feature inversely affect each other. This type of analysis is crucial for selecting features in multivariate analyses and understanding how attributes influence each other in genomic studies. These emphasized characteristics features help in understanding the hereditary feature of the glaucoma case. It is not limited to genetic markers and effects only but also touches grades of clinical diseases and demographic features. These variables relate to certain features of the disease manifestation which includes Index of Performance (IoP), morphology of optic nerve, the preexisting family history of the disease, and genetic prognosis for glaucoma respectively.

Data cleaning procedure

The initial data analysis represents that the newly collected dataset contains missing values in connection to several parameters. The occurrence of such conditions leads to possibility of getting inaccurate or biased data in the course of the study. Therefore, data wrangling is performed which include addressing of the null values and adequately managed in the data pre-processing step before the actual analysis whereby a detailed cleaning

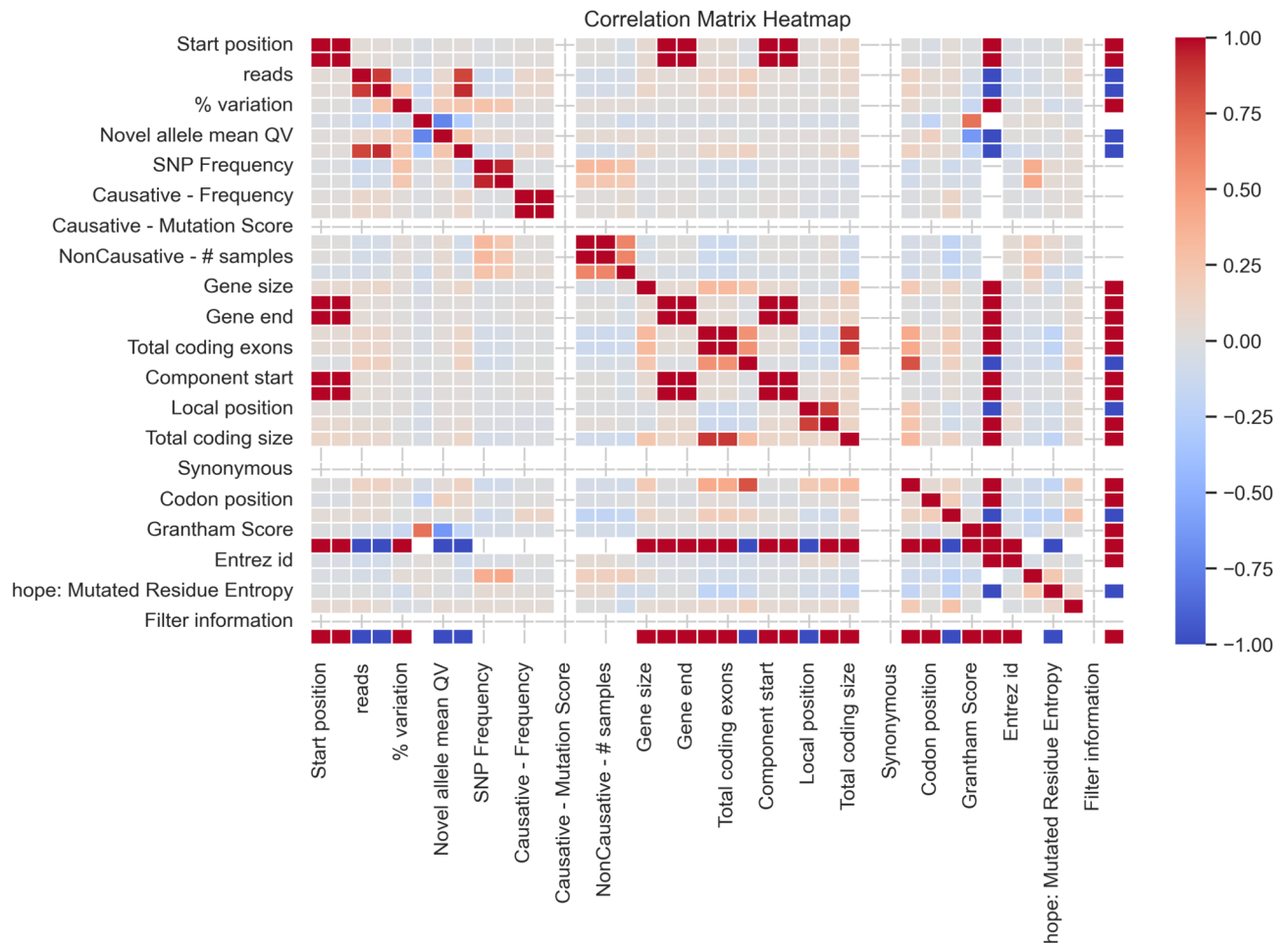


Fig. 2. The correlation matrix of various features of the extracted data through NGS technology.

of data is conducted to ensure that the level of unwanted null values is reduced to the required tolerable limit. To this end, the intended goal is to apply data imputation methods including mean or median. Some of these elements include, predicting of missing values using the data which is gathered previously. The data wrangling helps to prevent the subsequent analytical data from being sparse of valuable information that could represent the totality of what had been under analysis. The conservative approach that is adopted in handling of the issue leads to enhanced data analysis.

Features selection

The genetic variation percentage mean the measure of the genetic changes that exist in the population being analyzed. It shows how, while involved in the process of gene sequencing, they take an opportunity to create awareness by presenting the feature variation reads. These reads when aligned against a reference sequence can show a variance or variant from the sequence. Besides snips, these modifications can be described as variations and include such aspects as insertions, deletions and other genetic changes. A key component in such processes as a study of genetic variations within a given community and identification of disease associated genes in specific persons is the comparison of variations in genetic readings. By comparing these variance readings with that of other people or population, it is possible to identify if there are genetic changes that will predispose an individual to get an illness or any other favorable or unfavorable characteristics. Likewise, the genetic flags to which the matter specialist assigns a variation percentage of 80% or below would be classified as homozygous. Such markers are assumed to be dangerous if PhyloP values drop below 2 and the Grantham score crosses over 150 scores. This has laid down some useful guidelines for determining the utility of genetic variations and ascertaining whether they are or are likely to be caused by genetic selection.

The correlation of the key features is shown in Fig. 3 which provides a visual representation of the relationships and distributions for key numerical variables, including % variation, reads, Grantham Score, and PhyloP Score. Each diagonal plot shows the kernel density estimate (KDE) for the distribution of each variable, offering insights into the spread and skewness. The off-diagonal scatter plots reveal relationships between variables; for example, the plot between reads and % variation shows a dispersed relationship without a clear linear trend. This type of visualization is helpful in identifying potential patterns, outliers, or non-linear relationships that may not be evident from the correlation matrix alone. In the case of glaucoma, this makes a practical distinction

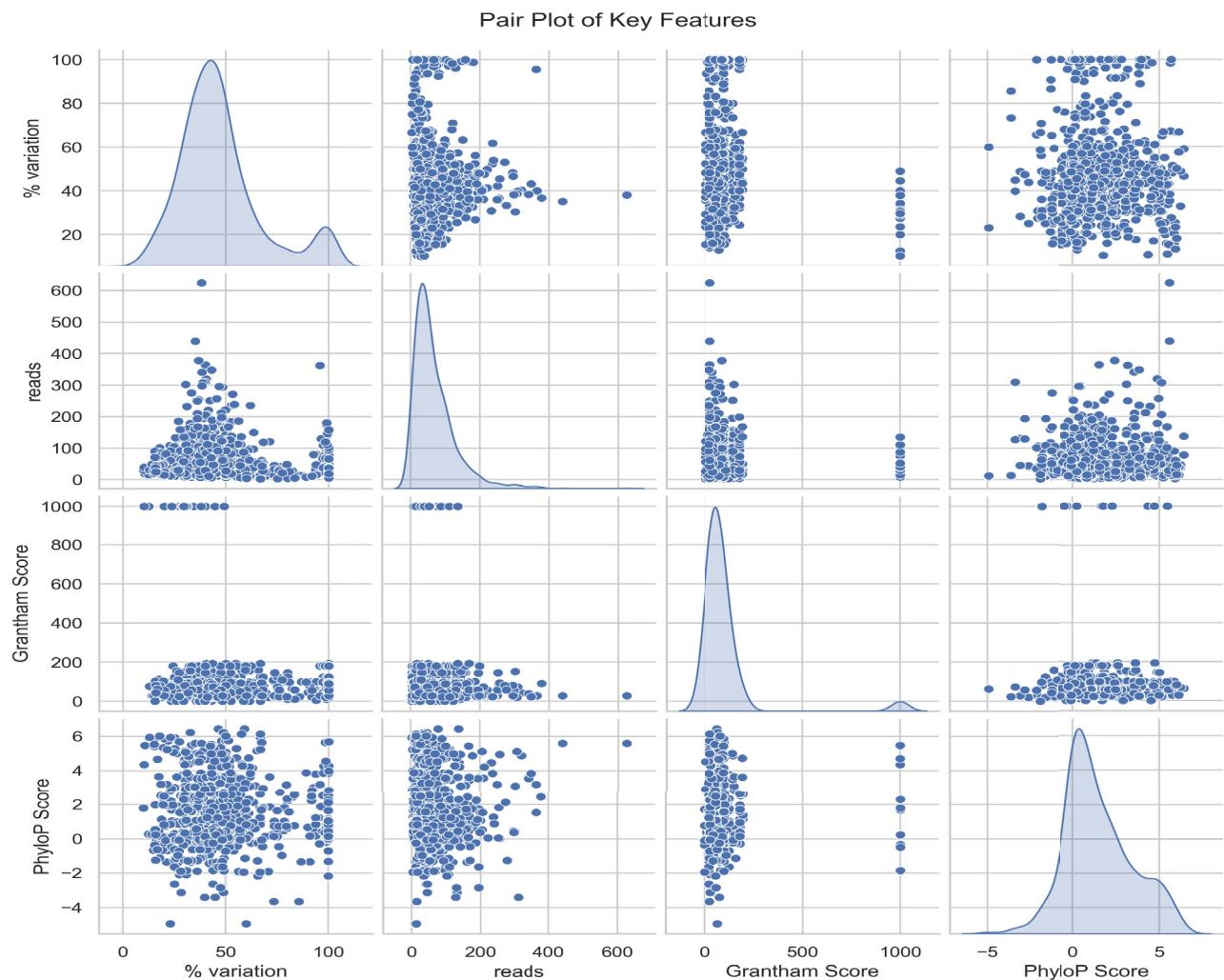


Fig. 3. The pair plot analysis of the key features of the DNA sequencing.

of the genetic factors from clinical features possible by engaging many dimensions. This makes it possible to get a better understanding of how some of the genetic factors can affect the clinical factors. This characteristic called the Work-in-Progress allows the researchers to study the relationships, the rhythms, and the links of certain characteristics. Thus, it is possible to prospect new genetic markers, to develop the prognostic markers, and to study genotype-phenotype correlations which are essential for the diagnostics and the treatment of the glaucoma. Collectively all the characteristic offers an angle of view on the disease.

Implementation of classification techniques

This article involves the use of classification methods for the entire research process. These methods employ different machine learning models to classify existing data into different classes depending on the assigned attributes. The three types of classification approach include the Support Vector Machine (SVM), Random Forest, and Decision Tree Classifier. These algorithms have a specific architectural and methodological implementation to solve the problem.

Support vector machine (SVM)

A supervised machine learning model used for categorized and continuous data analysis. SVMs define the range of distinction between data labels or outcomes. For data categorization, SVMs identify a hyperplane that maps data points into different categories or classes while defining an optimal separator for the classification tasks. The SVM aims to find the optimal hyperplane in the feature space separating different data categories while maximizing the margin. The margin is the maximum width of the parallel line around the hyperplane containing no internal support vectors. The support vectors are the data points that closely touch the boundaries of the respective classes. The application of SVM ranges from healthcare care, e.g., disease diagnosis, Disease Classification, Gene expression, Predictive Modeling, and voice and image recognition⁷⁴. Mathematically, the hyperplane is represented as Eq. 1,

$$w^T x + b = 0, \quad (1)$$

where w is the weight vector, x the feature vector, and b the bias, respectively.

The hyperplane can be easily determined for linearly separable data; however, with overlapping classes, largely used in real-world applications, SVMs maximize the distance between vectors in the feature space of interest with the corresponding support vectors. SVMs primarily excel in binary classification problems. For multiclass problems, SVMs employ strategies such as one-vs-one and one-vs-rest, where multiple binary classifiers are combined to handle computationally intensive tasks.

Random forest

Random forests are another form of the ensemble learning technique that involves the creation of numerous decision trees during the learning phase. The strategy provides the mean, be it for regression problems or for frequencies in classes for classification problems. This was why Random Forest is a compilation of decision trees, each trained independently of the others.

Random forests are another form of the ensemble learning technique that involves the creation of numerous decision trees during the learning phase. The strategy provides the mean, be it for regression problems or for frequencies in classes for classification problems. This was why Random Forest is a compilation of decision trees, each trained independently of the others.

The present article is concerned with the description of the Binary Recursive Partitioning trees, also known as the base of trees used in Random Forests. Categorical variables are divided into two categories by using presumed known binary partitions, which are referred to as splits. Trees are also used to divide the predictor region into subregions. The “root” node of the formed decision tree comprises each of the structures within the predictor space. Now, once the predictor space is partitioned, the nodes that do not divide further are known as the terminal node, and it is on this node that the final partition is made. Any nonterminal node divides into two descendant nodes, the left and the right, according to the value of one of the predictors.

$$Q = \sum_{k \neq k'}^k \hat{p}_k \hat{p}_{k'} \quad (2)$$

Considering the classification in Eq. (2), assuming there are K classes assign with the number 1, ..., K , the Gini index is a usual splitting benchmark that is used. Thus, when analysing classification in the context of Eq. (3), it is apprehended that there are of K classes marked by the numbers 1, ..., K , being the most used criterion is the Gini index which usually defines the classes K .

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^n I(y_i = k) \quad (3)$$

Decision tree

Supervised learning using the decision tree can be described as the technique of constructing the binary tree according to specified conditions as shown in Fig. 4. This is particularly the case when it is looked at as way of partitioning a domain into a number of class. The concept is based upon the act of heuristically partitioning of a data dominion (node) into two minor dominions until the it reached to final node. To do that, the split should make the information gained on the subdomains, regarding the original node more significant. Supervised learning is more developed form of learning that is usually used when classification of the data collected is necessary. In their informal sense, Information Gain is the extent of information acquired from a given source.

The notation used for the purpose is p_0 , for the probability that the next draw will be class 0, and p_1 represents the predicted value at node, which is the mean value of all the class 1 from the domain specified as D . Suppose the difference between p_0 and p_1 is getting closer to unity. It can be easily observed that a particular class occupies a position that makes any further division of the domain not required. Therefore as the absolute value of p_0 and p_1 goes to zero it implies that the two classes have same influence in that particular area. Therefore, it is important that a new equation should be formulated which will incorporate these two important variables; the father separation. Based on the aforementioned situation, we have created two subdomains; D_1 and D_2 .

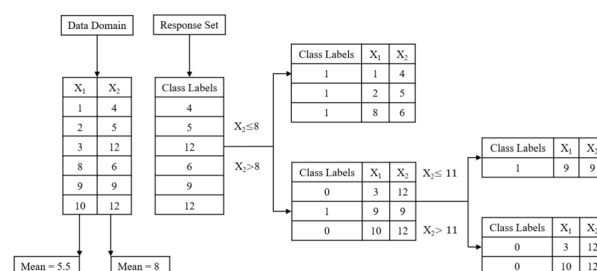


Fig. 4. Decision tree with bi-dimensional data domain and outputs a discrete result.

Probability of class 0 and class 1 from subdomain D_1 is represented by p_0 while that of class 1 is represented by p_1 . If one is efficient then the other is greater than p_0 or q_1 is greater than p_1 these are subparts of the string. In order to simplify this explanation, let us assume that value of p_0 is less than competitor's initial value of q_0 . As it is the case with the p_0 notation, q_0 can be defined as the extent of p_0 measured below some positive quantity ε , as presented in Equations. (4) and (5).

$$|q_0 - q_1| = |2q_0 - 1| = |2(p_0 + \varepsilon) - 1| = |2p_0 + 2\varepsilon + 1| \quad (4)$$

$$|q_0 - q_1| = |p_0 + 1 - p_1 + 2\varepsilon - 1| = |p_0 - p_1 + 2\varepsilon| \quad (5)$$

The mathematical Eq. (6) highlights the subsequent inequality (where q_0 is greater than p_0).

$$|q_0 - q_1| = |p_0 - p_1| \quad (6)$$

The difference in the inequality above is actually expressed by means of the absolute values that comprise quantitative measures showing the relative proportion and similarity between the classes in certain subdomains. The probabilistic measure is another statistic that is effective in achieving the objectives of the optimization objectives, especially in relation to the decision tree.

Results and discussion

This technique aims at using NGS data in order to diagnose eye diseases with biomarkers for early diagnosis. Thus, it is essential to bring some brief-focused changes in order to reflect the desired outcome properly. The AI implemented algorithms of SVC, Decision Tree, and Random Forest has been improved by the complex genomic data. The integration of the uncertainty demonstrates another and a wider range that is arguably inherent in a set of uncertainties. Hiding in this text are no mere numbers and quantitative predictions that have been coded in; they are the results of analyses, experiments and optimizations that remain integrated into our system.

This process involves identification of relevant data in nucleotides produced by NGS then followed by use of artificial intelligence in categorizing this data. This information is very useful because it's containing the important data about the predisposition to genetic disorders. In this article, the true beauty of construct of AI is rooted in solving the problems related to computer science and biology. It surpasses the point of analysis and becomes a tool in the learning process that allows one to see differences that would be otherwise extremely difficult to come by. The spirit of our findings is summarized in the first inference that the combination of AI with NGS data can impressively transform the diagnostics and go beyond traditional approaches. Earlier, the diagnosis approaches used the aspect of how patient symptoms looked like in diagnosing them.

Evaluation metrics

Every measure which is basic, or supplementary, such as accuracy, recall, F1 etc., provides a different perspective for viewing how the arithmetic operations and evaluations of our models are built. By following this approach, we may be able to make right decisions and further update the algorithms that we currently have. This offers a more critical analysis of the evaluation categories and benchmarks, and of the related concept, real-world examples, and better understanding of these parameters.

Precision

By precision, especially of classification, is the degree of how the model is able to correctly identify the positive instances of the test data. The question it seeks to answer is: "Which constituents fit the definition of the set of samples that are expected to be positive, and are positive in fact?"

- True positives (TP): this term refers to the situation which the model places a given input data point into the positive class or the right category. False positives (FP): it is understood as such instances that the model assigns to the positive risk zone when they should belong to the negative risk zone.

Precision is calculated using the Eq. (7) as follows:

$$Precision = \frac{TP + FP}{TP} \quad (7)$$

Most of the time, these measures define the proportional ($\frac{TP+FP}{TP}$) of the samples that were well classified as positive compared to the total number of positive samples in the given dataset. Higher values suggest that the model is doing a good job of getting the positive cases of the dataset most of the time thus reducing the chances of getting it wrong with the negatives.

Recall

Retrieve Sensitivity equally known as the true positive rate is a classification metric adopted for evaluating a model's capacity to classify every event from a particular class. The issue at question here is to identify which of the positive samples, and they can appear in many different modes, were correctly classified as such by the model.

- False negatives (FN): stands for situations when model actually predicts the negative class while being expected to predict the positive one.

Here is a more comprehensive Eq. (8).

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

Most of the time, these measures define the portion of the samples that were absolutely satisfied as positive compared to the total number of positive samples in the given dataset. Higher numbers suggest that the model is doing a good job of getting the positive cases of the dataset most of the time thus reducing the chances of getting it wrong with the negatives. Hence, it still has its significant role in the areas where wrong examples might lead to harm like in the case of diagnostic tests of medical ailments. When the recall rate is low, it means that the positive cases are not well identified thus making different true positive cases missed and so making false negatives. This could pose a possible issue in circumstances that involve having the means to preview all positive events.

F1-score

The F1 score is a classification performance measure for use that takes both the precision of the model as well as the recall into consideration. It provides an objective prognosis of 'how well' the experiment worked out for the corresponding model. The F1 score is calculated as the average of accuracy and recall as is evident in the formula in Eq. (9).

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (9)$$

It gives more importance to small values and that is why the F1-score is exact in-between. At the first point, we know the points of maximum gain, at the same time, point zero gives the minimum fit. Therefore, precision is 100 percent while recall is also 100 percent. The F1-score is particularly appropriate to be used since it offers a fine measure of how precise the speech measurement of a certain context is while also giving the extent of which it covers the pertinent speech measurement required. It is particularly useful in situations where there is a large number of positive cases or a large number of negative cases so that the cases are not misdiagnosed. Further, the F1-score can be optimized for the aim of comparing numerous models within the same set and especially in the cases where the several classes substantially differs from others.

Accuracy

In classification accuracy is a crucial scale because it describes the overall quality of the prediction developed by a particular model. The query it responds to is: "Out of all the samples present within the obtained dataset, how many did the model get to classify correctly?"

- True Negatives (TN): These are the instances when labeled as such there is actually no existence of the positive class according to the model.

Accuracy is evaluated by performing calculation as follows: Accuracy Ac is calculated by using a formula such as the one shown in Eq. (10).

$$Ac = \frac{TP + TN}{TotalSamples} \quad (10)$$

Accuracy is a scale which quantifies the proportion of total number of samples classify accurately, including the positive ones and the negative ones in relation to the total sum of sample in the dataset. The overall measure that is employed frequently to evaluate the effectiveness uniformly of a classification algorithm is named as accuracy. As it does not involve any complex calculations and is very simple to use, it is used more often to test the model for its efficiency especially when working on the data set which has equivalent count of classes. Nevertheless, accuracy improbable suitable especially when; the classes are skewed. Presuming, the accuracy of the algorithm is high it may mean that it has a high tendency of predicting the most frequent class for all the instances.

The comparison with existing work

Some studies that use Machine Learning Classifier (MLC) in retinal imaging for disease identification are presented in Table 1 for comparison. In each of these studies, the size of the population, the type of imaging done and the measured performance are different. Medeiros et al.⁶⁹ analyzed OCT images with an accuracy 83.7%. Kim et al.⁴⁰ also used OCT images, and reported higher performance with 98% accuracy. The respective authors Townsend et al.⁷⁵ employed cSLO images, but without published accuracy figures. The authors of Uchida et al.⁷⁶ used cSLO images, with an accuracy of 92%. This comparison shows that the diagnostic accuracy may differ among the studies due to differences of imaging technology and validation protocols contrasts in terms of imaging technique, and reported performance metrics. The GAINSeq (proposed) model used Random Forest, Decision Tree, and SVM algorithms using NGS data with an 80 : 20 training to testing ratio. The random forest score 99%, the decision tree, and the SVM models achieved accuracies of 99% and 92%, respectively.

The following Table 2 shows the comparison of different contemporary ML approaches aimed segmentation of optic cup in fundus images⁷⁷. It depicts the overall accuracy of the proposed sequence (GAINSeq) methods on the different classes of diseases compared with the accuracy documented in earlier research works. The Random Forest approach also gives a 99% accuracy level which is higher than 97% accuracy level documented in the literature. Slightly similarly, the Decision Tree technique gives a 99% accuracy, which is way higher than the earlier claimed 91%. However, the SVM method being proposed gives an accuracy of 92%, which can be seen

Authors	Data type	AI type	Accuracy
Medeiros et al. ⁶⁹	OCT images	Machine learning classifier (MLC)	83.70%
Kim et al. ⁴⁰	OCT images	Machine learning classifier (MLC)	98%
Townsend et al. ⁷⁵	cSLO images	Machine learning classifier (MLC)	–
Uchida et al. ⁷⁶	cSLO images	Machine learning classifier (MLC)	92%
GAINSeq (proposed)		Random forest	99%
	NGS data	Decision tree	99%
		SVM	92%

Table 1. Comparison of studies using MLC in retinal imaging for disease detection.

Method	Framework	Accuracy	
		GAINSeq	Literature
Random forest	Optic cup segmentation	99%	97%
Decision tree	Optic cup segmentation	99%	91%
SVM	Optic cup segmentation	92%	93%

Table 2. Comparison between existing image based work with proposed GAINSeq using NGS data on same AI models.

Method	Framework	Accuracy	
		GAINSeq	Literature
Random forest	Binary classification models	99%	72.62%
Decision tree	Binary classification models	99%	77.92%
SVM	Binary classification models	92%	71.96%

Table 3. The comparison table of prior FASTA-based work with the GAINSeq.

Classifier	Target	Precision		Recall		F1-score		Accuracy	
		Test	Train	Test	Train	Test	Train	Test	Train
Random forest	Homo	0.99	1	1	1	1	1	99%	100%
	Hetro	1	1	0.91	1	0.95	1		

Table 4. Random forest classification evaluation matrices.

to be less from the 93% stated in the current research. The comparison shows that the proposed Random Forest and Decision Tree outperform currently used ones while the proposed SVM method performs slightly worse. Table 3 provides a comparison of three binary classification models of Random Forests, Decision Trees, and SVMs that was used in the UniProt⁷⁸ dataset. The reliability of these methods is then checked in the framework suggested as well as comparison with results from literature. For the proposed framework, Random Forest and Decision Tree algorithms both get an accuracy of 99 percent which is higher than the previous experiment. On the other hand, SVM method yields a proposed accuracy of 92% slightly higher than the literature accuracy of 71.96%. Such a comparison highlights the efficacy of the suggested approaches in improving the classification accuracy in binary classification problems.

Random forest

In order to develop highly respectable and accurate classification models, one ought to have a good understanding about performance criteria. Accuracy, recall, Precision and F1 measures are some of the measures that depict the ability or power of the model in predicting over the classes as being accurate or not. Classification of datasets is very sensitive within Random Forest one of the most diverse forms of ensemble learning, with high and absolute accuracy and over-fitting. In this part, the performance assessment of the Random Forest Classifier model is explained briefly. It elaborates on how the model works in predicting the outcomes and how they came up with the measures of F1-Score accuracy, precision, and recall. Table 4 shows the assessment measurements of the random forest algorithm.

Precision measures the ability to predict positive cases by the model. The Homozygous class has a precision value at 99% out of the total samples in this class depicting that the classifier correctly classified this class samples. On the other hand, the precision of the Hetrozygous category is comparatively higher up to 1.00. All samples

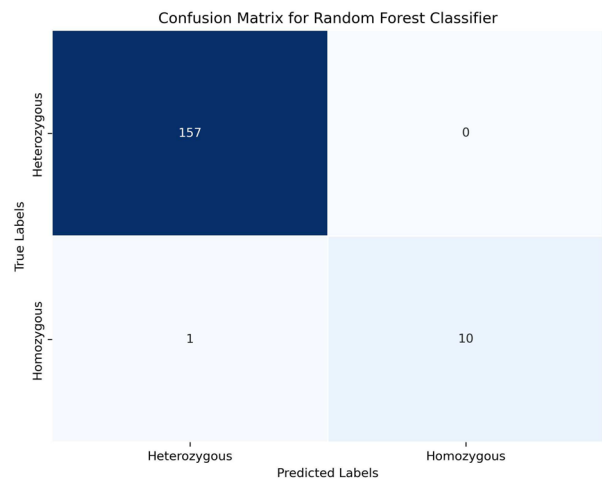


Fig. 5. Random forest confusion matrices.

Classifier	Target	Precision		Recall		F1-score		Accuracy	
		Test	Train	Test	Train	Test	Train	Test	Train
Support vector machine	Hetro	0.96	0.96	0.95	1	0.96	0.98	92%	96%
	Homo	0.38	0.93	0.45	0.62	0.42	0.75		

Table 5. Support vector classifier evaluation matrices.

predicted as ‘Homozygous’ were correctly identified in all the analyzed cases, thus the precision of the model is 100%.

Recall or sensitivity is the measure of how well the model makes correct predictions of positive cases all by itself. It suggests that for the ‘Homozygous’ class, recall is exactly 1 meaning there was no misidentification of the samples from this class; all the 15 real ‘Homozygous’ samples were correctly recognized by the model. Note that the recall for the “Hetrozygous” class is 0.91 which means that from all the actual “Hetrozygous” 91% were correctly recalled by the classifier. It is calculated as the simple average of the standard setups of precision and the measurement of the recall, and it is the best measure of the performance of a model. That is why the value of the F1-Score of the class “Homozygous” equals 1, which speaks about the perfect share the precision and the recall. Yet the F1-score of this model was 0.95 regarding the Macro-average for ‘Hetrozygous’ class. Accuracy can be described as a level of the plan and its predictions range at which the model is accurate or close to the target. Concisely, to the Random forest classifier, it has been seen a value of test accuracy as 0.99 Which means 99 percent samples are classified correctly.

From confusion matrix of the Random Forest Classifier, we get the overall performance of the model in classifying the dataset into two classes. Based on the analysis above, the results of the evaluation of the model and its defects are given in the Fig. 5. From all the samples that were classified as “Heterozygous” 157 were correctly classified while 0 were misclassified. At the same time, of them all the samples that were identified as “Homozygous” only one sample was classified as Hetrozygous.

Support vector machine

SVMs are used in the field of machine learning since they have been proved to deal with categorization tasks in multiple domain. The SVM algorithms gives very high accuracy in arranging decision planes which are used in determining the maximum margins and the kernel trick. This helps in grouping of points into different classes. SVM classification process can be described as very formal and mechanical. Here, in its simplest form, our analysis investigates hyperplanes and vectors in more detail and confines itself to the few key components that make the SVM such an effective tool for classification. In this section, the author gives a detailed account of how the SVM classifier works and the difficult it poses in preparing as well as analyzing the patterns in the acquired data. The classification performance of Support Vector Classifier (SVC) is shown in Table 5.

Precision quantify the degree of accurateness of the positive forecasts produced by the algorithm. The precision of assay for the “Heterozygous” class is 0.96, and this means 96% of the “Heterzygous” correctly classified. Yet, the precision for the “Homozygous” class is 0.38, which shows that 38 percent of samples classified to this class were actually correct. Sensitivity measures the ability of the model to identify those that have the disease, which is also known as recall. The recall for the “Heterozygous” class is 0.95 which can be interpreted that the model was inclusive of all 95% true “Heterozygous” samples. But in case of “Homozygous” class, it is 0.45, it simply means that about 45 percent of the total “Homozygous” samples are correctly identified. F1-score is also known as balanced accuracy score because this is an arithmetic mean of recall and precision. The precision and recall

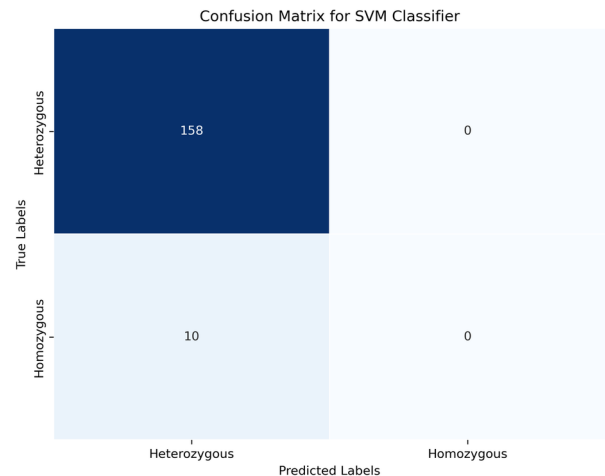


Fig. 6. Support vector classifier confusion matrices.

Classifier	Target	Precision		Recall		F1-score		Accuracy	
		Test	Train	Test	Train	Test	Train	Test	Train
Decision Tree	Hetro	0.99	1.00	1.00	1.00	1.00	1.00	99%	100%
	Homo	1.00	1.00	0.91	1.00	0.95	1.00		

Table 6. Decision tree classification evaluation matrices.

values for the “Heterozygous” class are in balanced 0.96 of the F1-score. Labeled accuracy of SVM Classifier has been found to be 0.92, which attest that about 92% of data were rightly classified.

The confusion matrix for the SVC is a measure of the model’s ability to correctly classify the data into two obviously separated classes as depicted by Fig. 6. Heterozygous samples were as expected with all 158 samples in this group correctly identified, which resulted in a perfect precision and recall for this category. However, there were no samples that were correctly placed in this category which resulted in a complete loss of ability to predict this category and all samples labelled “Homozygous” were misclassified. Such differences in the forecasts are perhaps meant to underline the fact that the model has some limitations in making a clear distinction between the ‘Homozygous’ category which also sheds light to the direction where the model has potential to enhance the forecasting capability.

Decision tree

The Decision Trees are especially beneficial in the case of predictive modeling in the machine learning environment as it can be easy to handle and easily understandable. Tree-based method divides the given data into different regions with homogenous characteristic based on certain properties by using iterative technique and tree actually explore the features space. Decision trees which use splits to illustrate decisions depending on the value of a characteristic make understanding of decisions easy. This makes them very suitable for uses like prognosis and knowledge mining among others. But, in order to understand Decision Tree Classification better, you also have to consider the tree topology as well as the pre-existing decision trees and then deduce the patterns needed for classification. This section provides the additional proximate analysis and clearer comprehension of Decision Tree Classification and its outcome in identifying complex relation between several of the data. Now, let’s discuss the peculiarities of Decision Trees and find out how one can construct the reliable and easy to understand models for prediction using the material below. The decision tree categorization results are stated in the Table 6 below.

Precision quantifies the degree of accurateness of the positive forecasts constructed by the algorithm. The precisions regards all “Heterozygous” and “Homozygous” classes are measured 0.99 and 1.00 individually. Obviously, measures are the model readiness to identify true positives or in other words it is the indicator of correct positive. The recall on the ‘Heterozygous’ class is 1.00, which symbolize the fact that the model under discussion in able to correctly identify all the actual ‘Heterozygous samples’ and therefore possesses the recognition rate of 100%. When it comes to the class “Homozygous” the recall value I calculate is 0.91, which means 91% of the original “Homozygous” samples were retrieved. The F1-score then is an arithmetic mean derived from both the precision as well as the recall and gives an estimation of the model’s performance which is fairly mutual. The values of the F1-score of both classes respectively are 1.00 and 0.95, hence there is minimal trade-off between precision and recall. Essentially, Precision is directly related to exactly the predictions that a given model is making. After training, the Decision Tree Classifier performed best by achieving accuracy of 99%, and proved that the classifier can accurately classify a sample in the set.

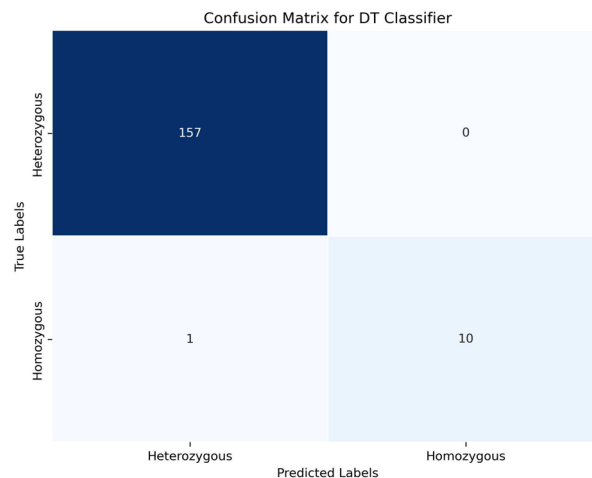


Fig. 7. Decision tree confusion matrices.

The Decision Tree (DT) is among the most popular classifiers used in the problem-solving process. An algorithm which can be easily understood and which has a relatively small amount of test error. It can be seen from the confusion matrix in Fig. 7 that the DT classifier achieves an excellent result in the classification of data and also shows a high level of perceptiveness. Every examined sample belonging to the “Heterozygous” category of the current classification task was correctly and sufficiently recognized by the computer without a single misidentification. The same applies to the “Homozygous” type of occurrences, which it correctly identifies with a recall value of 1; there were no false-negative true positive. The graphic given below evidently indicates the outperformed DT classifier in handling the movement of the algorithm.

Conclusion and future work

Such types are important because glaucoma is one of the second most common causes of blindness and is one of the greatest threats to global health. Using the visual data when diagnosing the condition may have a number of difficulties and limitations when it comes to defining the genetic aspects of the sickness. Hence, a recent study has shown that phenotype not necessarily has to manifest with symptoms and that through the use of NGS data, AI methods can be used to identify genetic signatures of glaucoma. In this example, there is an exhaustive methodology where data had to be gathered before proceeding to calculations to arrive at the proposed attributes before the model was built. Thus, a clear, effective guideline of the appropriate framework is provided for our study to randomly and systematically select appropriate genetic data of glaucoma patients.

While the current study was not designed as a clinical tool, we believe that it provides an important foundation for future translational work. By identifying potentially pathogenic mutations in patients with glaucoma, our research can help refine the genetic panels used in diagnostic screening and provide insight into disease mechanisms that could guide targeted therapies or personalized surveillance protocols in the future. Moreover, our machine learning framework offers a scalable method that could, with further validation, contribute to the development of genomic-based risk prediction models. As our findings are further corroborated, they can improve our understanding of glaucoma pathogenesis and inform early detection strategies or family-based screening programs, particularly in genetically predisposed populations.

It has been discussed in the context of the this study that the preparation stage was instrumental in dealing with such problems as missing values, outliers and other irrelevant data. This step improved the quality of the data for further analysis in the subsequent step by a great extent. Firstly, statistical analysis and machine learning were used to select the features while secondly examining the association of every characteristic to the disease. This approach was envisaged to reveal the features which possessed the worst predictive capacity. This stage was vital since it aimed at improving the general performance of the machine learning models that would make up the system. Concerning the construction of the model, it is necessary to state that we carried out the analysis and training and testing of a set of AI algorithms. Generalizability is constrained by its small and demographically homogeneous cohort. The reason for this is that we evaluated their applicability based on performance criteria such as accuracy, precision, F1 score, and recall. Therefore, the aforementioned study could be useful to support if the data presented in this paper can sufficiently support this statement that glaucoma can be timely diagnosed using artificial intelligence to interpret the NGS data with the help of these brilliant gene markers. It aids in the creation of a considerably better imaging scheme compared with traditional imaging for the gene segments that relate to the disease under study and for accurate and timely identification of the disease.

In the same way, future studies should employ the functional investigation of novel genes that have not been described before. Therefore, providing accurate information on the dynamics and trails of these genes, particularly in relation to eye-related processes, could go a long way towards unraveling the genetic antecedents of glaucoma. This may employ functional genomic methods that aims at understanding the properties of genes and the roles they play in certain ocular functions. A possible strategy could therefore include deploying optimally current machine learning technologies including Deep Learning and Ensemble methods to improve on the

chances of making accurate predictions including factors such as highly correlated genes. Although algorithmic choices (RF, DT, SVM) mitigated overfitting, validation in larger, diverse populations is essential. Future work will integrate federated learning to harmonize data across global cohorts. We plan to adopt fairness-sensitive AI frameworks to audit and correct biases in multiethnic datasets, ensuring equitable diagnostic performance across populations. Therefore, future phases of this research will involve collaboration with clinical partners to assess the utility of these genetic markers in patient cohorts. This will include functional studies of candidate variants and possibly prospective clinical trials to evaluate the performance of the predictive model in diverse populations. The composition of the datasets which was one of the key limitations, we plan to extend our research in the future phase by incorporating data from various ethnic groups including underrepresented populations that might harbor rare genetic mutations.

Data availability

The datasets generated and analyzed during the current study are available in the *Harvard Dataverse*, V1 repository, accessible at <https://doi.org/10.7910/DVN/1FVLVI>.

Received: 4 February 2025; Accepted: 26 May 2025

Published online: 02 July 2025

References

- Senthil, S. et al. Glaucoma mimickers: A major review of causes, diagnostic evaluation, and recommendations. *Semin. Ophthalmol.* **36**, 692–712 (2021) (Taylor & Francis).
- Saba, T., Bokhari, S. T. F., Sharif, M., Yasmin, M. & Raza, M. Fundus image classification methods for the detection of glaucoma: A review. *Microsc. Res. Tech.* **81**, 1105–1121 (2018).
- Tham, Y.-C. et al. Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis. *Ophthalmology* **121**, 2081–2090 (2014).
- Stuart, K. V., Pasquale, L. R., Kang, J. H., Foster, P. J. & Khawaja, A. P. Towards modifying the genetic predisposition for glaucoma: an overview of the contribution and interaction of genetic and environmental factors. *Molecular aspects of medicine* **93**, 101203 (2023).
- Doucette, L. P., Rasnitsyn, A., Seifi, M. & Walter, M. A. The interactions of genes, age, and environment in glaucoma pathogenesis. *Surv. Ophthalmol.* **60**, 310–326 (2015).
- Ito, Y. A. & Walter, M. A. Genetics and environmental stress factor contributions to anterior segment malformations and glaucoma. In *Glaucoma* (Rumelt, S. Ed.). Chap. 3. <https://doi.org/10.5772/54653> (IntechOpen, 2013).
- Rao, K. N., Nagireddy, S. & Chakrabarti, S. Complex genetic mechanisms in glaucoma: An overview. *Indian J. Ophthalmol.* **59**, S31–S42 (2011).
- Naj, A. C., Schellenberg, G. D. & (ADGC), A. D. G. C. Genomic variants, genes, and pathways of Alzheimer's disease: An overview. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **174**, 5–26 (2017).
- Wang, Z., Wiggs, J. L., Aung, T., Khawaja, A. P. & Khor, C. C. The genetic basis for adult onset glaucoma: Recent advances and future directions. *Prog. Retinal Eye Res.* **90**, 101066 (2022).
- Sato, M. et al. The association between intraocular pressure dynamics during dark-room prone testing and intraocular pressure over a relatively long-term follow-up period in primary open-glaucoma patients. *Graefes Arch. Clin. Exp. Ophthalmol.* **262**, 949–956. <https://doi.org/10.1007/s00417-023-06257-0> (2024).
- Zukerman, R. et al. Molecular genetics of glaucoma: Subtype and ethnicity considerations. *Genes* **12**, 55 (2020).
- Choquet, H., Paylakhi, S., Kneeland, S. C. et al. A multiethnic genome-wide association study of primary open-angle glaucoma identifies novel risk loci. *Nat. Commun.* **9**. <https://doi.org/10.1038/s41467-018-04555-4> (2018).
- Gharahkhani, P., Jorgenson, E., Hysi, P. et al. Genome-wide meta-analysis identifies 127 open-angle glaucoma loci with consistent effect across ancestries. *Nat. Commun.* **12**. <https://doi.org/10.1038/s41467-020-20851-4> (2021).
- Lee, S.S.-Y. & Mackey, D. A. Glaucoma-risk factors and current challenges in the diagnosis of a leading cause of visual impairment. *Maturitas* **163**, 15–22. <https://doi.org/10.1016/j.maturitas.2022.05.002> (2022).
- Valenzi, E. et al. Disparate interferon signaling and shared aberrant basaloid cells in single-cell profiling of idiopathic pulmonary fibrosis and systemic sclerosis-associated interstitial lung disease. *Front. Immunol.* **12**, 595811. <https://doi.org/10.3389/fimmu.2021.595811> (2021).
- Stuart, K. V., Pasquale, L. R., Kang, J. H., Foster, P. J. & Khawaja, A. P. Towards modifying the genetic predisposition for glaucoma: An overview of the contribution and interaction of genetic and environmental factors. *Mol. Asp. Med.* **93**, 101203. <https://doi.org/10.1016/j.mam.2023.101203> (2023).
- Al-Khatib, K., Al-Mufti, N., Al-Rafei, M. & Al-Hashimi, M. Congenital glaucoma: A review of diagnosis and management. *Saudi J. Ophthalmol.* **31**, 91–96 (2017).
- Choon, Y. et al. Artificial intelligence and database for Ngs-based diagnosis in rare disease. *Front. Genet.* **14**, 1258083. <https://doi.org/10.3389/fgene.2023.1258083> (2024).
- Satam, H. et al. Next-generation sequencing technology: Current trends and advancements. *Biology* **12**, 997. <https://doi.org/10.3390/biology12070997> (2023).
- Nafea, A. M. et al. Application of next-generation sequencing to identify different pathogens. *Front. Microbiol.* **14** (2024).
- Milla, E. et al. Next-generation sequencing-based gene panel tests for the detection of rare variants and hypomorphic alleles associated with primary open-angle glaucoma. *PLoS ONE* **19**, e0282133. <https://doi.org/10.1371/journal.pone.0282133> (2024).
- Gupta, V. & Selvan, H. Juvenile-onset open-angle glaucoma. In *Childhood Glaucoma: Current Trends and Future Prospects*. 21–33 (Springer, 2024).
- Ouanes, K. Chapter 27—The role of artificial intelligence and machine learning in Ngs. In (Dash, H. R., Elkins, K. M. & Al-Snan, N. R. eds.) *Next Generation Sequencing (NGS) Technology in DNA Analysis*. 531–538. <https://doi.org/10.1016/B978-0-323-99144-5.00002-0> (Academic Press, 2024).
- Vilhekar, R. S. & Rawekar, A. Artificial intelligence in genetics. *Cureus* **16**, e52035. <https://doi.org/10.7759/cureus.52035> (2024).
- Hill, R. K. What an algorithm is. *Philos. Technol.* **29**, 35–59 (2016).
- Syed, S., Ahmed, R., Iqbal, A., Ahmad, N. & Alshara, M. A. Mediscan: A framework of u-health and prognostic AI assessment on medical imaging. *J. Imaging* **10**, 322 (2024).
- Ali, A., Iqbal, A., Khan, S., Ahmad, N. & Shah, S. A two-phase transfer learning framework for gastrointestinal diseases classification. *PeerJ Comput. Sci.* **10**, e2587 (2024).
- Shoukat, A., Akbar, S., Hassan, S. A. E., Rehman, A. & Ayesha, N. An automated deep learning approach to diagnose glaucoma using retinal fundus images. In *2021 International Conference on Frontiers of Information Technology (FIT)*. 120–125 (IEEE, 2021).
- Leung, K. et al. A case for the use of artificial intelligence in glaucoma assessment. *Ophthalmol. Glaucoma* (2021).

30. Islam, M. M., Yang, H.-C., Poly, T. N., Jian, W.-S. & Li, Y.-C.J. Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. *Comput. Methods Prog. Biomed.* **191**, 105320 (2020).
31. Benjamins, S., Dhunoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. *NPJ Digit. Med.* **3**, 118 (2020).
32. Kaya, M. K. & Tasci, B. Pretrained models and the role of feature selection: An artificial intelligence-based approach in the diagnosis of diabetic retinopathy. *Turk. J. Sci. Technol.* **18**, 511–517 (2023).
33. Singh, L. K., Khanna, M., Garg, H., Singh, R. & Iqbal, M. A three-stage novel framework for efficient and automatic glaucoma classification from retinal fundus images. *Multimed. Tools Appl.* **83**, 85421–85481 (2024).
34. Zhang, L., Tang, L., Xia, M. & Cao, G. The application of artificial intelligence in glaucoma diagnosis and prediction. *Front. Cell Dev. Biol.* **11**, 1173094 (2023).
35. Albahri, A. S. et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Inf. Fusion* **96**, 156–191 (2023).
36. Singh, L. K., Garg, H., Pooja, & Khanna, M. Performance analysis of machine learning techniques for glaucoma detection based on textural and intensity features. *Int. J. Innov. Comput. Appl.* **11**, 216–230 (2020).
37. Coan, L. J. et al. Automatic detection of glaucoma via fundus imaging and artificial intelligence: A review. *Surv. Ophthalmol.* **68**, 17–41 (2023).
38. González-Gonzalo, C. et al. Trustworthy AI: closing the gap between development and integration of AI systems in ophthalmic practice. *Prog. Retinal Eye Res.* **90**, 101034 (2022).
39. Dawoodbhoy, F. M. et al. Ai in patient flow: Applications of artificial intelligence to improve patient flow in NHS acute mental health inpatient units. *Heliyon* **7** (2021).
40. Kim, S. J., Cho, K. J. & Oh, S. Development of machine learning models for diagnosis of glaucoma. *PloS one* **12**, e0177726 (2017).
41. Lin, M. et al. Automated diagnosing primary open-angle glaucoma from fundus image by simulating human's grading with deep learning. *Sci. Rep.* **12**, 14080 (2022).
42. Khalil, T., Akram, M. U., Khalid, S., Dar, S. H. & Ali, N. A study to identify limitations of existing automated systems to detect glaucoma at initial and curable stage. *Int. J. Imaging Syst. Technol.* **31**, 1155–1173 (2021).
43. Acharya, U. R. et al. Automated diagnosis of glaucoma using texture and higher order spectra features. *IEEE Trans. Inf. Technol. Biomed.* **15**, 449–455 (2011).
44. Arslan, S. et al. Attention turkernext: Investigations into bipolar disorder detection using oct images. *Diagnostics* **13**, 3422 (2023).
45. Ganokratanaa, T., Ketcham, M. & Pramkeaw, P. Advancements in cataract detection: The systematic development of lenet-convolutional neural network models. *J. Imaging* **9**, 197 (2023).
46. Gonzalez-Gonzalo, C. et al. Trustworthy AI: Closing the gap between development and integration of AI in ophthalmology. In *Eur. J. Ophthalmol.* **31**, 14–14 (Sage Publications Ltd, 2021).
47. Singh, L. K., Khanna, M. & Singh, R. Application of artificial intelligence and automation techniques to health service improvements. In *Applied Intelligence in Human-Computer Interaction*. 67–81 (CRC Press, 2023).
48. Sinthanayothin, C., Boyce, J. F., Cook, H. L. & Williamson, T. H. Automated localisation of the optic disc, fovea, and retinal blood vessels from digital colour fundus images. *Br. J. Ophthalmol.* **83**, 902–910 (1999).
49. Lim, W. S. et al. Use of multimodal dataset in AI for detecting glaucoma based on fundus photographs assessed with oct: Focus group study on high prevalence of myopia. *BMC Med. Imaging* **22**, 206 (2022).
50. Fan, Z. et al. Optic disk detection in fundus image based on structured learning. *IEEE J. Biomed. Health Inform.* **22**, 224–234 (2017).
51. Krishna, N. & Nagamani, K. Glaucoma diagnosis using state of art image analysis techniques: a comprehensive survey. In *2020 International Conference on Inventive Computation Technologies (ICICT)*. 247–251 (IEEE, 2020).
52. Abdullah, F. et al. A review on glaucoma disease detection using computerized techniques. *IEEE Access* **9**, 37311–37333 (2021).
53. Kheradmandi, N. & Mehranfar, V. A critical review and comparative study on image segmentation-based techniques for pavement crack detection. *Construct. Build. Mater.* **321**, 126162 (2022).
54. Kishorjit Singh, N., Johnny Singh, N. & Kanan Kumar, W. Image classification using slic superpixel and faagkfcmm image segmentation. *IET Image Process.* **14**, 487–494 (2020).
55. Singh, L. K. et al. Histogram of oriented gradients (HOG)-based artificial neural network (ANN) classifier for glaucoma detection. *Int. J. Swarm Intell. Res. (IJSIR)* **13**, 1–32 (2022).
56. Ali, R. et al. Optic disk and cup segmentation through fuzzy broad learning system for glaucoma screening. *IEEE Trans. Indus. Inform.* **17**, 2476–2487 (2020).
57. Christopher, M. et al. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Sci. Rep.* **8**, 16685 (2018).
58. Singh, L. K., Khanna, M., Thawkar, S. & Singh, R. A novel hybridized feature selection strategy for the effective prediction of glaucoma in retinal fundus images. *Multimed. Tools Appl.* **83**, 46087–46159 (2024).
59. Bunod, R., Augstburger, E., Brasnu, E., Labbe, A. & Baudouin, C. Intelligence artificielle et glaucome: Une revue de la littérature. *J. Français d'Ophtalmol.* **45**, 216–232 (2022).
60. Mehta, N., Tsui, E., Lee, G. D., Dedania, V. & Modi, Y. Imaging biomarkers in diabetic retinopathy and diabetic macular edema. *Int. Ophthalmol. Clin.* **59**, 241–262 (2019).
61. Qureshi, I., Khan, M. A., Sharif, M., Saba, T. & Ma, J. Detection of glaucoma based on cup-to-disc ratio using fundus images. *Int. J. Intell. Syst. Technol. Appl.* **19**, 1–16 (2020).
62. Murtagh, P., Greene, G. & O'Brien, C. Current applications of machine learning in the screening and diagnosis of glaucoma: A systematic review and meta-analysis. *Int. J. Ophthalmol.* **13**, 149 (2020).
63. Raja, C. & Gangatharan, N. A hybrid swarm algorithm for optimizing glaucoma diagnosis. *Comput. Biol. Med.* **63**, 196–207 (2015).
64. Zedan, M. J. et al. Automated glaucoma screening and diagnosis based on retinal fundus images using deep learning approaches: A comprehensive review. *Diagnostics* **13**, 2180 (2023).
65. Alawad, M. et al. Machine learning and deep learning techniques for optic disc and cup segmentation—A review. *Clin. Ophthalmol.* **747**–764 (2022).
66. Dow, E. R. et al. From data to deployment: The collaborative community on ophthalmic imaging roadmap for artificial intelligence in age-related macular degeneration. *Ophthalmology* **129**, e43–e59 (2022).
67. Elangovan, P. & Nath, M. K. Glaucoma assessment from color fundus images using convolutional neural network. *Int. J. Imaging Syst. Technol.* **31**, 955–971 (2021).
68. Akbar, S., Hassan, S. A., Shoukat, A., Alyami, J. & Bahaj, S. A. Detection of microscopic glaucoma through fundus images using deep transfer learning approach. *Microsc. Res. Tech.* **85**, 2259–2276 (2022).
69. Medeiros, F. A., Jammal, A. A. & Thompson, A. C. From machine to machine: An oct-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. *Ophthalmology* **126**, 513–521 (2019).
70. Bajwa, M. N. et al. Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning. *BMC Med. Inform. Decis. Mak.* **19**, 1–16 (2019).
71. Alam, M. et al. Quantitative optical coherence tomography angiography features for objective classification and staging of diabetic retinopathy. *Retina* **40**, 322–332 (2020).
72. Juneja, M. et al. Gc-net for classification of glaucoma in the retinal fundus image. *Mach. Vis. Appl.* **31**, 1–18 (2020).

73. Iqbal, M., Iqbal, A. & Ayub, H. GAINSeq: Glaucoma Pre-Symptomatic Detection Using Machine Learning Models Driven by Next-Generation Sequencing Data. <https://doi.org/10.7910/DVN/1FVLVI> (2025).
74. Guido, R., Ferrisi, S., Lofaro, D. & Conforti, D. An overview on the advancements of support vector machine models in healthcare applications: A review. *Information* **15**, 235 (2024).
75. Townsend, K. et al. Heidelberg retina tomograph 3 machine learning classifiers for glaucoma detection. *Br. J. Ophthalmol.* **92**, 814–818 (2008).
76. Uchida, H., Brigatti, L. & Caprioli, J. Detection of structural damage from glaucoma with confocal laser image analysis. *Invest. Ophthalmol. Vis. Sci.* **37**, 2393–2401 (1996).
77. Bouacheria, M., Cherfa, Y., Cherfa, A. & Belkhamisa, N. Automatic glaucoma screening using optic nerve head measurements and random forest classifier on fundus images. *Phys. Eng. Sci. Med.* **43**, 1265–1277 (2020).
78. Ganeshpurkar, A. et al. Improved machine learning scoring functions for identification of electrophorus electricus's acetylcholinesterase inhibitors. *Mol. Divers.* 1–25 (2022).

Acknowledgements

The authors would like to thank Prince Sultan University for paying the Article Processing Charges (APC) of this publication. They would also like to thank Prince Sultan University for their support.

Author contributions

Conceptualization, M.I. and A.I.; methodology, M.I. and A.I.; software, M.I.; validation, A.I., H.A., M.K., and N.A.; formal analysis, A.I. and M.I.; investigation, A.I., M.K.; resources, H.A., M.I., and A.I.; data curation, M.I., H.A., and A.I.; writing-original draft preparation, M.I. and A.I.; writing-review and editing, A.I., M.K., M.A.; visualization, M.I.; supervision, A.I.; project administration, A.I., N.A., and M.A.; funding acquisition, N.A. and M.A.; All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.I.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025