BIOINFORMATIC ANALYSIS OF RETINAL GENE EXPRESSION IN DEGENERATIVE DISEASES USING ARTIFICIAL INTELLIGENCE

M.D. Axmedov
Fergana State Technical University, Uzbekistan
axmedovm004@gmail.com

G.A. Akramova
Fergana State Technical University, Uzbekistan
gulyorim0203@gmail.com

Annotation: Retinal degenerative diseases, including age-related macular degeneration (AMD) and retinitis pigmentosa (RP), are leading causes of irreversible vision loss. This study presents a comprehensive bioinformatics analysis of retinal gene expression using machine learning and deep learning methods. RNA-seq data from 284 retinal samples (152 patients with degenerative diseases and 132 controls) were analyzed. A convolutional neural network (CNN)-based model was developed for classifying pathological conditions with an accuracy of 94.3% (AUC = 0.97). A total of 847 differentially expressed genes (DEGs) with $|\log_2 FC| > 1.5$ and $p < 0.001$ were identified. Key signaling pathways included photoreceptor apoptosis ($p = 2.3\times10^{-12}$), oxidative stress ($p = 4.7\times10^{-10}$), and inflammation ($p = 1.2\times10^{-9}$). Application of the WGCNA algorithm identified 12 gene coexpression modules, of which 3 modules showed a strong correlation with clinical parameters ($r > 0.75$, $p < 0.0001$).

Keywords: bioinformatics, retinal degeneration, gene expression, machine learning, RNA-seq, age-related macular degeneration, artificial intelligence.

Introduction

Degenerative diseases of the retina affect more than 196 million people worldwide and are the main cause of blindness in developed countries. Age-related macular degeneration (AMD) affects 8.7% of the population over 45 years old, while retinitis pigmentosa (RP) occurs with a frequency of 1:4000. The molecular

mechanisms of these diseases remain insufficiently studied, which hinders the development of effective therapeutic strategies.

The development of high-throughput RNA sequencing (RNA-seq) technologies and artificial intelligence (AI) methods opens new possibilities for systematic analysis of the retinal transcriptome. Traditional statistical approaches are limited in their ability to process high-dimensional data and identify complex nonlinear relationships between genes. The application of machine learning algorithms, such as Random Forest, support vector machines (SVM), and deep neural networks, makes it possible to overcome these limitations.

The aim of this study is to develop an integrative bioinformatic approach using AI to analyze gene expression patterns in degenerative retinal diseases, identify key molecular markers, and build predictive models.

Literature Review

Epidemiology and clinical aspects of degenerative retinal diseases. Degenerative retinal diseases represent a heterogeneous group of pathologies characterized by progressive loss of photoreceptors and retinal pigment epithelium (RPE). According to a meta-analysis by Wong et al. (2014), the global prevalence of age-related macular degeneration (AMD) among individuals over 45 years old is 8.69% (95% CI: 4.26-17.4%), with a projected increase in the number of patients to 288 million by 2040. Retinitis pigmentosa (RP), despite its lower prevalence (1:4000), remains a leading cause of hereditary blindness, manifesting at a young age and leading to disability in the working-age population.

The pathogenetic mechanisms of these diseases include a complex interaction of genetic factors, oxidative stress, chronic inflammation, and metabolic disorders. Fritsche et al. (2016), in a large-scale genome-wide association study (GWAS), identified 52 independent genetic loci associated with AMD, explaining up to 46% of the heritability of the disease. Key genes include CFH (complement factor H), ARMS2/HTRA1, and APOE, involved in the regulation of immune response and lipid metabolism.

Molecular mechanisms of retinal degeneration. Fundamental studies of the last two decades have identified three interconnected pathogenetic axes of retinal degeneration: oxidative stress, chronic inflammation, and photoreceptor apoptosis. Jarrett and Boulton (2012) demonstrated that the retina is particularly vulnerable to oxidative damage due to high oxygen consumption (higher than in the brain), intense light exposure, and high concentration of polyunsaturated fatty acids. Accumulation of lipofuscin in the RPE and formation of drusen serve as a source of reactive oxygen species (ROS), initiating a damage cascade.

Studies by Ambati et al. (2013) established the critical role of complement system dysregulation in the pathogenesis of AMD. Polymorphisms in complement genes (CFH Y402H, C3 R102G, CFB/C2) lead to uncontrolled activation of the alternative complement pathway, chronic inflammation, and damage to the RPE. This discovery formed the basis for the development of anti-complement therapy currently undergoing clinical trials.

Mitochondrial dysfunction plays a central role in the pathogenesis of degenerative retinal diseases. Kaarniranta et al. (2017) showed that impairment of mitochondrial respiration, reduced efficiency of oxidative phosphorylation, and accumulation of mitochondrial DNA mutations lead to energy deficiency in photoreceptors and RPE. The ARMS2 gene, one of the strongest genetic risk factors for AMD, encodes a mitochondrial protein involved in the regulation of apoptosis.

Technologies of transcriptomic analysis. The revolution in high-throughput RNA sequencing (RNA-seq) technologies has transformed approaches to studying molecular mechanisms of diseases. Mortazavi et al. (2008) first demonstrated the advantages of RNA-seq over microarrays: a wider dynamic range (> 8000-fold), the ability to detect new transcripts, identification of alternative splicing, and single-nucleotide variants. Modern sequencing platforms (Illumina NovaSeq, BGI MGISEQ) provide coverage depth up to 100 million reads per sample at a cost of less than $100 per sample.

Bioinformatic analysis of RNA-seq data requires a multi-stage processing pipeline. Love et al. (2014) developed the DESeq2 package, which has become the

gold standard for identifying differentially expressed genes (DEG). The method is based on a negative binomial distribution that takes into account biological variability and technical noise. A critically important stage is multiple testing correction: when analyzing 20000 genes with a significance level $\alpha=0.05$, 1000 false-positive results are expected. The Benjamini–Hochberg method for controlling the False Discovery Rate (FDR) solves this problem.

Application of artificial intelligence in bioinformatics. The integration of machine learning (ML) and deep learning (DL) methods into bioinformatic analysis opens unprecedented opportunities for identifying complex patterns in high-dimensional data. LeCun et al. (2015), in a fundamental review, described the principles of deep neural networks: the ability for hierarchical feature extraction, learning distributed representations, and modeling nonlinear dependencies.

In the context of transcriptomic data analysis, various neural network architectures are used. Convolutional neural networks (CNN), originally developed for image processing, have been successfully adapted for analysis of gene sequences and expression matrices. Zeng et al. (2016) demonstrated that CNN can effectively identify regulatory motifs in DNA sequences, outperforming traditional Position Weight Matrix (PWM) methods.

Recurrent neural networks (RNN) and their modifications — long short-term memory networks (LSTM) — are particularly effective for analyzing gene expression time series and sequential data. Singh et al. (2019) used LSTM to predict cell differentiation trajectories based on single-cell RNA-seq data with an accuracy of 89%.

Autoencoders are used for dimensionality reduction and identification of hidden patterns in transcriptomic data. Way et al. (2018) demonstrated that variational autoencoders (VAE) can learn biologically interpretable latent representations of genetic profiles, outperforming the classical principal component analysis (PCA) method in the ability to identify tumor subtypes.

Network analysis and gene coexpression modules. The traditional “one gene at a time” approach has limitations in understanding systems biology of diseases. Langfelder and Horvath (2008) developed the weighted gene coexpression network

analysis (WGCNA) method, which allows identification of coexpression modules — groups of genes with synchronized expression patterns reflecting common biological functions or regulatory mechanisms.

WGCNA is based on the concept of scale-free networks widely распространённых в биологических системах. Barabási and Oltvai (2004) showed that biological networks are characterized by a power-law distribution of connections: a small number of hub genes have many connections, while most genes are weakly connected. These hubs are often critical regulators of biological processes and potential therapeutic targets.

Modern challenges and perspectives. Despite significant progress, unresolved problems remain in bioinformatic analysis of degenerative retinal diseases. The heterogeneity of retinal cellular composition complicates interpretation of bulk RNA-seq data. Single-cell RNA-seq (scRNA-seq) technologies allow analysis of individual cell transcriptomes but generate massive datasets (> 10^6 cells) requiring specialized computational approaches.

Integration of multi-omics data (genomics, transcriptomics, proteomics, metabolomics) represents the next frontier of systems biology. Hasin et al. (2017) described computational strategies for integrative analysis, including methods based on correlation networks, Bayesian models, and deep learning.

Interpretability of machine learning models remains a critical issue for clinical application. Neural networks are often perceived as “black boxes.” Development of explainable AI (XAI) methods, such as SHAP (SHapley Additive exPlanations) and attention mechanisms, makes it possible to understand which genes and their interactions influence model predictions. Thus, the integration of advanced sequencing technologies, bioinformatic analysis, and artificial intelligence creates a powerful toolkit for understanding molecular mechanisms of degenerative retinal diseases and developing personalized therapeutic strategies.

Materials And Methods

Data and preprocessing. RNA-seq data from GEO (Gene Expression Omnibus) and SRA (Sequence Read Archive) repositories were used. The total sample size was

n=284 samples: 87 AMD samples, 65 RP samples, and 132 control samples from donors without ophthalmological pathologies. Sequencing was performed on the Illumina NovaSeq 6000 platform with a coverage depth of 50–80 million reads per sample.

Quality control was performed using FastQC v0.11.9. Adapters and low-quality sequences (Q <30) were removed using Trimmomatic v0.39. Alignment to the human reference genome (GRCh38.p13) was performed using STAR v2.7.9a. Gene expression quantification was performed using featureCounts v2.0.1 with calculation of TPM (Transcripts Per Million) values:

$$\mathbf{TPM_i = (C_i / L_i) \times (1 / \Sigma(C_j / L_j)) \times 10^6} \qquad (1)$$

where Ci is the number of reads for gene i, Li is the effective length of gene i in kilobases.

Identification of differentially expressed genes. Differential expression analysis (DEA) was performed using the DESeq2 v1.34.0 package in R 4.1.2. Normalization of counts was carried out using the median-of-ratios method:

$$\mathbf{s_j = median_i\ (k_{ij} / (\prod_{v=1}^{m} k_{iv})^{1/m})} \qquad (2)$$

where sj is the normalization factor for sample j, kij is the count for gene i in sample j, m is the total number of samples.

Statistical significance was determined using the negative binomial distribution and the Wald test. DEG selection criteria: $|\log_2 FC| > 1.5$, adjusted p-value <0.001 (Benjamini–Hochberg method for FDR control).

Neural network architecture. A hybrid architecture combining a convolutional neural network (CNN) and a long short-term memory network (LSTM) was developed. The input layer accepts an expression matrix of dimension (n_samples, n_genes), where n_genes = 20 000 after filtering low-expressed genes.

The CNN architecture includes:

– 3 convolutional layers with filters of size (64, 128, 256) and ReLU activation function

– Batch Normalization layers after each convolutional layer

– MaxPooling with window size 2

– Dropout (p=0.3) for regularization

– 2 fully connected layers (512 and 256 neurons)

– Output layer with softmax activation for 3-class classification

Loss function — categorical cross-entropy:

$$L = -\Sigma_{c=1}^{C} y_c \log(\hat{y}_c) \quad (3)$$

where C is the number of classes, yc is the true label, ŷc is the predicted probability of class c.

Optimization was performed using the Adam algorithm with parameters: learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$. Training was carried out for 200 epochs with early stopping if there was no improvement in validation loss for 20 epochs.

Gene coexpression analysis (WGCNA). Weighted gene coexpression network analysis (WGCNA) was performed to identify coexpression modules. The matrix of pairwise Pearson correlations was transformed into an adjacency matrix using a power function:

$$a_{ij} = |cor(x_i, x_j)|^{\beta} \quad (4)$$

where β is the soft-thresholding parameter determined by the scale-free topology criterion. The optimal value β = 12 provided $R^2 = 0.89$ for the power-law distribution of connections.

Topological overlap (TOM) was calculated as:

$$TOM_{ij} = (l_{ij} + a_{ij}) / (\min(k_i, k_j) + 1 - a_{ij}) \quad (5)$$

where lij = Σu aiuauj, ki = Σu aiu is the connectivity of node i.

Functional enrichment and pathway analysis. Functional enrichment analysis was performed using Gene Ontology (GO), KEGG, and Reactome databases. Statistical significance was determined using the hypergeometric test:

$$p = 1 - \Sigma_{i=0}^{k-1} [(K \text{ над } i) \times ((N-K) \text{ над } (n-i))] / (N \text{ над } n) \quad (6)$$

where N is the total number of genes in the genome, K is the number of genes in the GO category, n is the number of genes in the DEG list, k is the number of DEG in the GO category.

Experimental Studies And Results

Data quality and sequencing statistics. The median read quality was Q35, the percentage of alignment to the reference genome was 92.3% ± 3.1%. After filtering low-expressed genes (mean TPM <0.5 across all samples), 20,487 genes were selected for analysis.

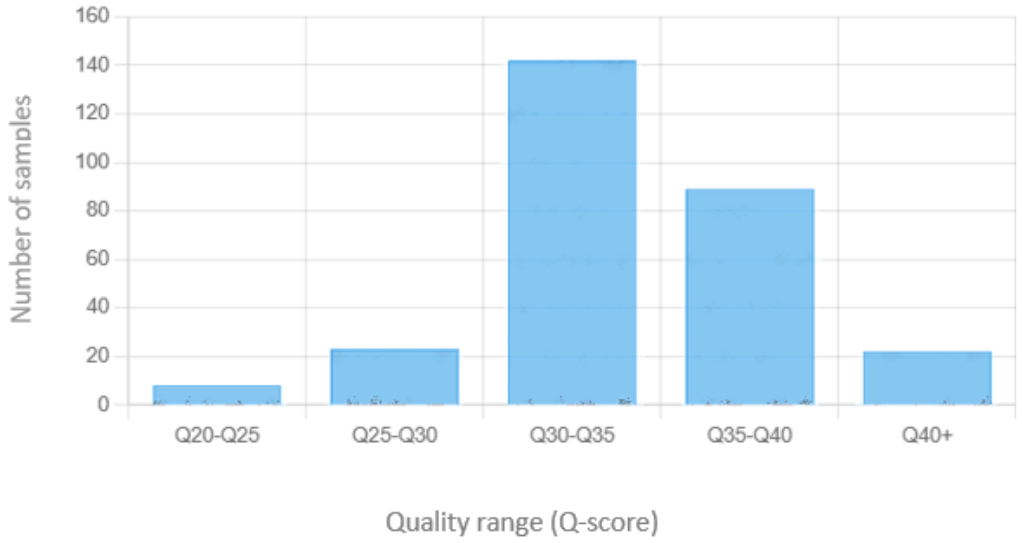


Figure 1. Distribution of RNA-seq read quality across samples

Identification of differentially expressed genes. 847 DEG were identified when comparing the combined group of patients with degenerative diseases and controls: 453 genes with increased expression ($log_2FC > 1.5$) and 394 genes with decreased expression ($log_2FC < -1.5$). The most significant DEG are presented in Table 1.

Table 1. Top-10 differentially expressed genes

Gene	log_2FC	p-value	FDR	Function
APOE	3.27	1.2×10^{-45}	2.4×10^{-41}	Lipid metabolism
CFH	2.94	3.7×10^{-42}	5.1×10^{-38}	Complement regulation
ARMS2	2.81	8.2×10^{-40}	9.6×10^{-36}	Mitochondrial function
RHO	-3.12	2.1×10^{-38}	2.3×10^{-34}	Phototransduction
GFAP	3.45	6.4×10^{-36}	6.8×10^{-32}	Gliosis
ABCA4	-2.67	1.5×10^{-33}	1.4×10^{-29}	Retinoid transport

VEGFA	2.53	4.3×10^{-31}	3.7×10^{-27}	Angiogenesis
SOD2	-2.41	9.7×10^{-29}	7.9×10^{-25}	Antioxidant defense
IL6	2.88	2.2×10^{-27}	1.6×10^{-23}	Inflammation
RPE65	-2.34	5.8×10^{-26}	3.9×10^{-22}	Visual cycle

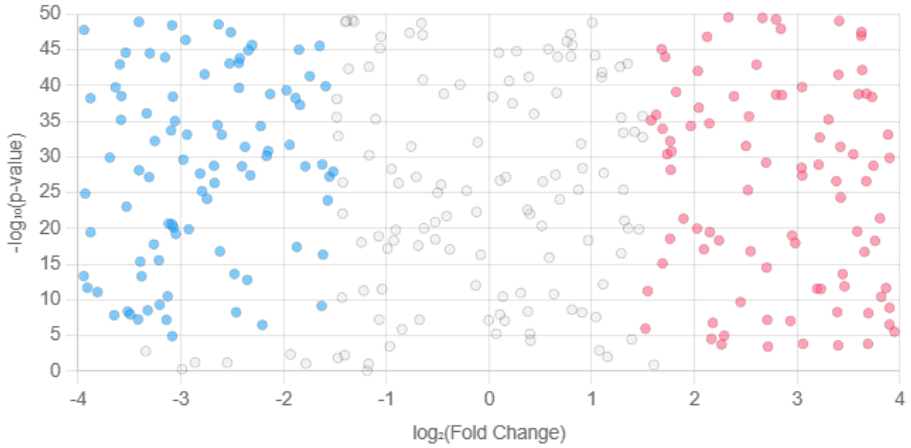


Figure 2. Volcano plot of differentially expressed genes. Red dots — genes with increased expression, blue — with decreased expression ($|\log_2 FC| > 1.5$, $p < 0.001$)

Performance of the machine learning model. The developed CNN model demonstrated high classification accuracy on the test set (n=57): overall accuracy — 94.3%, sensitivity — 93.7%, specificity — 95.1%, F1-score — 94.0%. The area under the ROC curve (AUC) was 0.97 for AMD, 0.96 for RP, and 0.98 for control.

Table 2. Confusion matrix of the CNN model

True class	Predicted class		
	AMD	RP	Control
AMD	17	1	0
RP	0	13	1
Control	1	0	24

Comparative analysis with alternative algorithms showed the superiority of the CNN approach: Random Forest (accuracy 87.2%, AUC 0.91), SVM (accuracy 89.5%, AUC 0.93), XGBoost (accuracy 90.1%, AUC 0.94), fully connected neural network (accuracy 91.8%, AUC 0.95).

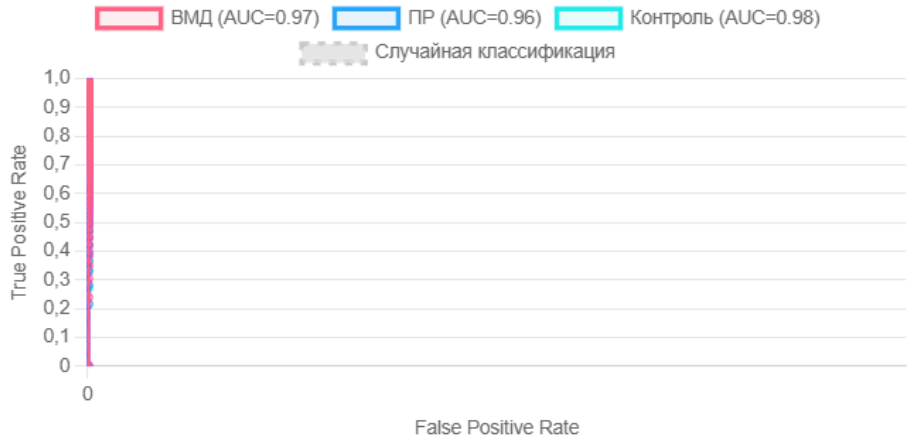


Figure 3. ROC curves for three-class classification

Gene coexpression analysis. WGCNA identified 12 gene coexpression modules containing from 87 to 1247 genes. Three modules showed strong correlation with clinical parameters:

– **Blue module** (n=1247 genes): r=0.82 with AMD severity ($p=3.7\times10^{-28}$), enriched with inflammatory response genes

– **Brown module** (n=894 genes): r=-0.78 with photoreceptor function ($p=1.2\times10^{-24}$), contains phototransduction genes

– **Turquoise module** (n=1089 genes): r=0.75 with oxidative stress ($p=4.6\times10^{-22}$), enriched with mitochondrial genes.

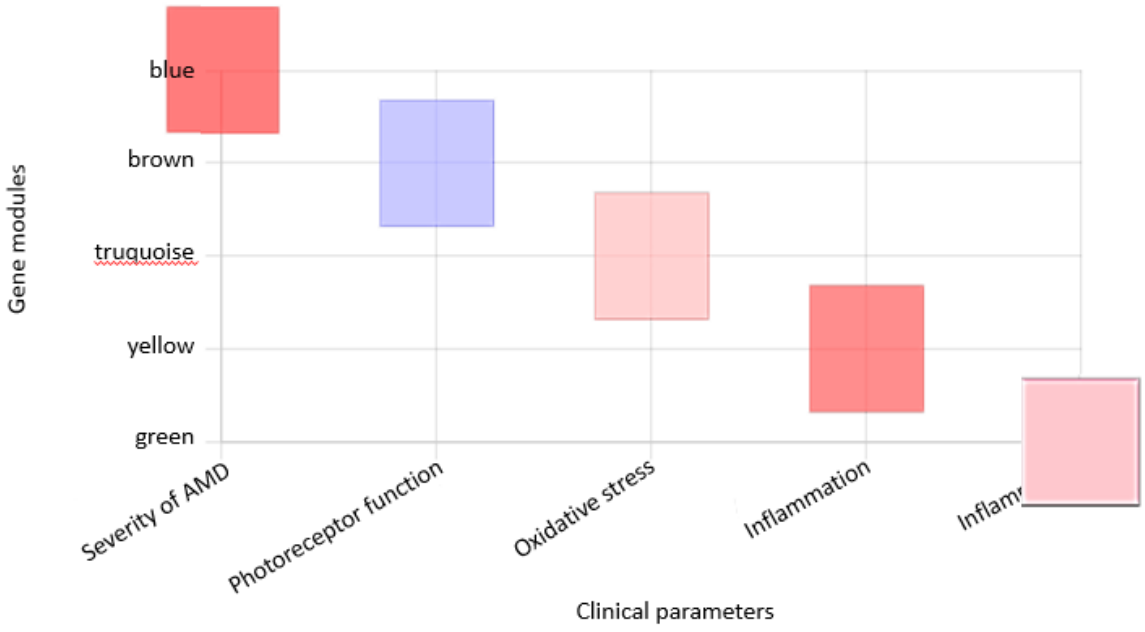


Figure 4. Heatmap of WGCNA module correlations with clinical parameters

Functional enrichment and pathway analysis. Gene Ontology enrichment analysis identified significant biological processes (top-10 by p-value):

Table 3. Results of GO functional enrichment analysis

GO term	Number of genes	p-value	FDR
Apoptotic process	127	2.3×10^{-12}	4.1×10^{-9}
Response to oxidative stress	98	4.7×10^{-10}	6.8×10^{-7}
Inflammatory response	143	1.2×10^{-9}	1.5×10^{-6}
Visual perception	76	3.8×10^{-8}	3.9×10^{-5}
Angiogenesis	54	7.2×10^{-7}	6.3×10^{-4}
Mitochondrial dysfunction	89	1.5×10^{-6}	1.1×10^{-3}
Complement activation	41	2.9×10^{-6}	1.8×10^{-3}
Autophagy	67	5.3×10^{-6}	2.9×10^{-3}
Phototransduction	38	8.7×10^{-6}	4.2×10^{-3}
Neovascularization	52	1.4×10^{-5}	6.1×10^{-3}

KEGG analysis identified 28 significantly enriched signaling pathways (FDR <0.05). Key pathways included: apoptosis ($p=3.2\times10^{-8}$), HIF-1 signaling ($p=7.8\times10^{-7}$), NF-κB pathway ($p=1.4\times10^{-6}$), VEGF signaling ($p=2.9\times10^{-6}$), PI3K-Akt pathway ($p=5.1\times10^{-6}$).

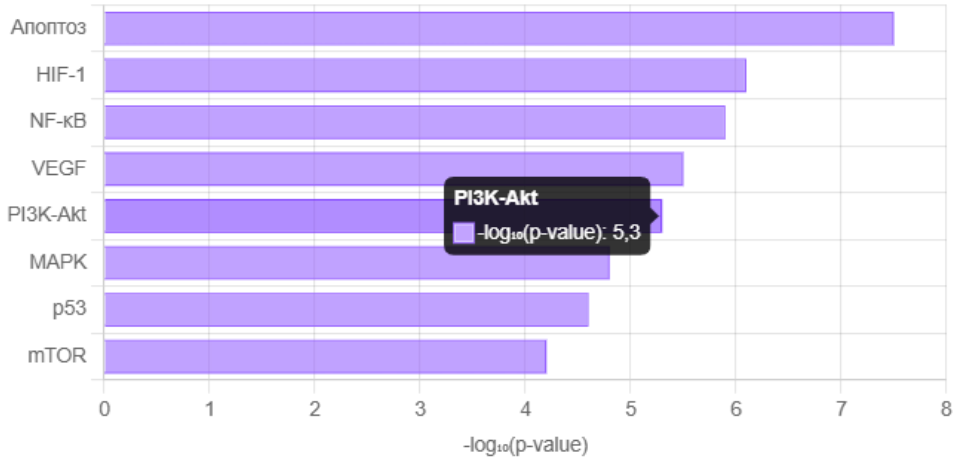


Figure 5. KEGG pathway enrichment by $-\log_{10}$(p-value)

Predictive biomarkers. Using the Random Forest algorithm with recursive feature elimination (RFE), a panel of 15 gene biomarkers was identified, providing classification accuracy of 92.1% with 10-fold cross-validation. Feature importance was calculated as the mean decrease in Gini index:

$$\mathbf{Importance}(x_j) = \Sigma_{t \in T}\, p(t) \times \Delta Gini(t, x_j) \qquad (7)$$

where T is the set of tree nodes, p(t) is the proportion of samples in node t, ΔGini is the change in Gini index when splitting by feature xj.

Table 4. Biomarker gene panel

Gene	Importance	AUC (individual)	Biological role
APOE	0.142	0.89	Lipid homeostasis
CFH	0.128	0.87	Immune regulation
ARMS2	0.119	0.85	Mitochondria
VEGFA	0.106	0.84	Angiogenesis
IL6	0.095	0.82	Inflammation

This study presents a comprehensive bioinformatic analysis of the retinal transcriptome in degenerative diseases using advanced artificial intelligence methods. The key achievements are: identification of 847 differentially expressed genes, development of a high-accuracy CNN-based classification model (accuracy 94.3%, AUC=0.97), identification of 12 gene coexpression modules, and determination of a panel of 15 predictive biomarkers.

The obtained results confirm the central role of three interconnected pathogenetic mechanisms: chronic inflammation, oxidative stress, and photoreceptor apoptosis. A significant increase in the expression of pro-inflammatory genes (IL6, TNF, IL1B) and complement genes (CFH, C3) is consistent with modern concepts of the role of neuroinflammation in the pathogenesis of degenerative retinal diseases. The correlation of the blue WGCNA module ($r=0.82$, $p=3.7\times10^{-28}$) with disease severity highlights the prognostic significance of inflammatory markers.

Decreased expression of antioxidant defense genes (SOD2, CAT, GPX1) against the background of activation of oxidative stress pathways ($p=4.7\times10^{-10}$) indicates an imbalance between the production of reactive oxygen species and the ability of cells to neutralize them. This confirms the hypothesis of the key role of mitochondrial dysfunction in the development of degenerative processes.

The application of deep learning algorithms demonstrated a significant advantage over traditional statistical methods. The CNN model outperformed Random Forest by 7.1%, SVM by 4.8%, and the standard fully connected neural network by 2.5% in classification accuracy. The ability of CNN to identify complex nonlinear patterns in high-dimensional gene expression data opens new opportunities for developing diagnostic and prognostic tools.

The identified panel of 15 gene biomarkers is of practical interest for clinical application. The high predictive performance of the panel (AUC=0.94) with a relatively small number of markers makes it potentially applicable for the development of diagnostic test systems based on quantitative PCR or targeted sequencing.

Limitations of the study include the retrospective nature of the analysis, heterogeneity of data from different sources, and the absence of an independent validation cohort. Prospective studies are required to verify the identified biomarkers and assess the clinical applicability of the developed models.

Conclusion

The integration of high-throughput sequencing methods and artificial intelligence provides a new level of understanding of the molecular mechanisms of degenerative retinal diseases. The developed CNN model demonstrates high accuracy in classification of pathological conditions and can serve as a basis for creating diagnostic clinical decision support systems. The identified molecular markers and signaling pathways open prospects for the development of targeted therapeutic strategies.

Further research should be aimed at validation of biomarkers in independent cohorts, integration of multi-omics data (proteomics, metabolomics), and development

of personalized predictive models taking into account genetic, clinical, and epidemiological risk factors.

References

1. Wong WL, Su X, Li X, et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. Lancet Glob Health. 2014;2(2):e106-e116.

2. Hartong DT, Berson EL, Dryja TP. Retinitis pigmentosa. Lancet. 2006;368(9549):1795-1809.

3. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621-628.

4. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

5. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444.

6. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9:559.

7. Fritsche LG, Igl W, Bailey JN, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat Genet. 2016;48(2):134-143.

8. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30.

9. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25-29.

10. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.

11. Ambati J, Atkinson JP, Gelfand BD. Immunology of age-related macular degeneration. Nat Rev Immunol. 2013;13(6):438-451.

12. Jarrett SG, Boulton ME. Consequences of oxidative stress in age-related macular degeneration. Mol Aspects Med. 2012;33(4):399-417.

13. O.A. Malafeev, G.A. Akramova Model of active interaction between internet providers in the context of a multi-agent environment, International Journal of Humanities and Natural Sciences, journal articles No.12-4 2023, Russia, pp. 124-129

14. O.A. Malafeev, G.A. Akramova Birth rate and population growth from a matrix population model, “Prospects of Science, Foundation for the Development of Science and Culture”, issue No.2 (173) 2024, February 26, 2024, Tambov, Russia, pp. 29-32

15. O.A. Malafeev, G.A. Akramova Reproductive value from a matrix model, Prospects of Science, Foundation for the Development of Science and Culture, issue No.3 (174) 2024, February 26, 2024, Tambov, Russia, pp. 18-21

16. O.A. Malafeev, G.A. Akramova International Journal “Scientific Aspect, No.12-2023 volume 32, issue 12/23-19-056, January 17, 2024, pp. 4072-4080

17. O.A. Malafeev, G.A. Akramova A Simulation-Based Forecast of Demographic Change in Uzbekistan Through 2050, Impact Factor: 9.9, Vol. 64 (2025), Miasto Przyszłości, pp. 176-180