An NB-ANN based Fusion Approach for Disease Genes Prediction and LFKH-ANFIS Classifier for Eye Diseases Identification

Samar Jyoti Saikia[1]
Gauhati University, Guwahati-781014, Assam, India
Assam Don Bosco University, Guwahati-781017
Assam, India

Dr. S. R. Nirmala[2]
Gauhati University, Guwahati-781014, Assam, India
KLE Technological University, Hubli-580031
Karnataka, India

***Abstract*—A key step to apprehend the mechanisms of cells related to a particular disease is the disease gene identification. Computational forecast of disease genes are inexpensive and also easier compared to biological experiments. Here, an effectual deep learning-centered fusion algorithm called Naive Bayes-Artificial Neural Networks (NB-ANN) is proposed aimed at disease gene identification. Additionally, this paper proposes an effectual classifier, namely Levy Flight Krill herd (LFKH) based Adaptive Neuros-Fuzzy Inferences System (ANFIS), for the prediction of eye disease that are brought about by the human disease genes. Utilizing this technique, completely '10' disparate sorts of eye diseases are identified. The NB-ANN includes these '4' steps: a) construction of '4' Feature Vectors (FV), b) selection of negative data, c) training of FV utilizing NB, and d) ANN aimed at prediction. The LFKH-ANFIS undergoes Feature Extraction (FE), Feature Reduction (FR), along with classification for eye disease prediction. The experimental outcomes exhibit that method's efficiency with regard to precision and recall.**

Keywords*—Disease gene identification; eye disease identification; deep learning; adaptive neuro-fuzzy inferences system (ANFIS); levy flight based krill herd (LFKH); principle component analysis (PCA)

I. Introduction

Disease genes are the dysfunction of a collection of genes, which in turn leads to Complex diseases [1] [2]. A key step towards enlightening the fundamental molecular operations of diseases is the recognition of genes concerned with genetic as well as rare diseases [3]. Prioritizing the candidate genes using experimental approaches is very costly and tedious [4]. Matrix decomposition along with Network propagation is the '2' categories under which all these existing techniques for the prediction can well be summarized [5]. In current years, technologies, say higher-throughput [6] gene expressions profiling has permitted the characterization of molecular differences betwixt healthy and disease states, bringing about the recognition of an augmenting number of disease-linked genes [7]. A great quantity of machine learning-centered computational methods was generated for predicting disease genes [8], say restricted Boltzmann machines [9], deep belief network [10], linear regressions model [11], support vectors machine [12], multilayer perceptions (MLP) [13], et cetera. These often attain greater prediction accuracy on larger data sets [14]. Nonetheless, on account of the lower statistical power brought about by means of smaller samples in bio-medical data, the issue of smaller samples typically causes poor reproducibility of prediction outcomes among disparate patients [15]. To trounce such downsides, in this paper, an NB-ANN is proposed for identifying the disease genes as well as the LFKH-ANFIS is proposed for the identification of eye-linked diseases triggered by means of those recognized disease genes.

II. Literature Review

Chen BoLin et al.[16] proffered a kernel-centric Markov random field approach. This approach was deployed for capturing the genes-diseases associations on the base of biological networks. Here, three sorts of kernels were deployed for delineating the overall relations of vertices in 5 biological networks, respectively, and weighted methodology was built with the proffered approach to merge those data. It acquired 0.771- Area under the ROC Curves (AUC) score when merging all the concerned biological data. Here, Markov Exponential Diffusions (MED) kernel rendered the low AUC performance contrasted with Laplacian Exponential Diffusions (LED) kernel on integrated '3' network situation.

Abdulaziz Yousef and Nasrollah Moghadam Charkari [17] rendered a disease gene identification technique centered on amino acids' physicochemical properties as well as classification algorithm. Amino acids physic-chemical properties were utilized to change the sequences of protein into numerical vector for the feature vector generation. Support vector data description algorithm was employed to envisage the disease genes. The rendered method performed better contrasted with the prevailing methods concerning precision, recall, along with F-measure. Data standardization was required for Principle Component Analysis (PCA) utilization. The standardization absence brought about the PCA's failure in finding optimal components which in turn affected this model's performance.

Zhen Tian et al. [18] paid attention on a framework, termed RWRB, for inferring the causal genes of disease. The Similarity Networks (SN) of 5 genes (protein) was individually constructed grounded on countless genomic data. The integrated gene SN was re-developed in respect of the SN fusion approach. The restart along with random walk algorithm

was deployed on a Phenotype-Gene (PG) bi-layer network, which integrated phenotype SN, PG association, and integrated gene SN for proffering the priority for the candidate genes (disease-associated ones). Outcomes corroborated that the RWRB was accurate when analogized to certain methods regarding the evaluation metrics. This method rendered the degraded performance with respect to Number of Successful Predictions (NSP) metric when the jumping probability is above 0.6 in disparate experiments.

Mehdi Joodaki et al. [19] put forward a gene ranking approach, named as Random Walks with Restart on a Heterogeneous Networks with Fuzzy Fusions (RWRHN-FF). Here, first, centred on disparate genomic sources, '4' gene-gene similarity networks were generated, and then, they were joined utilizing the type-II fuzzy voter scheme. The resultant gene to gene network was linked with the disease-disease similarity network. By means of integrating '4' sources via a '2'- part disease gene network, the disease-disease similarity network was created. RWRHN analyzed this network. While considering Area Under ROC Curves (AUC) as well as convergence time, the presented approach trounced the prevailing methods. On account of the bad data integration of manifold sources, the precision metric of this method was declined.

Pradipta Maji and Ekta Shah [20] recognized the disease-associated genes with the utilization of a gene selection algorithm, named SiFS. The SiFS algorithm gathered countless genes as of micro-array data as diseased genes by elevating the functional and significance similarity of the chosen gene subset. Contrarily, a similarity metric was instated for computing the functional similarity betwixt 2 genes. The experimental outcomes on disparate data sets corroborated that the algorithm recognized more disease-associated genes when analogized to prevailing disease gene selection methodologies. The similarity measure of the presented method was affected by the low coverage of human genes and reliability of protein-protein interaction (PPI).

III. Proposed Methodology

Here, a novel sequence-based fusion method (NB-ANN) is proposed aimed at disease genes identification, and the LFKH-ANFIS is proposed aimed at identifying eye-related diseases that are triggered by those disease genes, say Age-associated Macular Degenerations (AMD), cataract, glaucoma, inherited optics neuropathies, Marfan syndrome polypoidal choroidals vasculopathies, retinitis pigmentosas, Stargardt disease, along with uveal melanoma.

The proposed method's architecture is exhibited in Fig. 1. Fig. 1 exhibits the proposed methodology's architecture. In the initial phase, representation methods are used to achieve the FV as of the disease and unknown disease genes. In the 2nd phase, a dataset with positive as well as reliable negative instances is created by selecting negative protein set. In the 3rd phase, disparate FV of the same instances are categorized using the NB classifier. In the '4th' phase, ANN fuses together the NB classifiers to enhance the accuracy. After the identification of disease gene, FE is done. It extracts the features as of the identified disease genes for classifying the eye-related diseases caused via the disease genes. Next, PCA is employed for FR for removing the redundant features. Lastly, the LFKH-ANFIS algorithm takes care of the eye disease classification.

A. Disease Gene Identification

Here, the technique for identifying along with prioritizing disease genes is elucidated. The proposed work comprises '4' steps: (i) Translate equivalent gene products (proteins) into '4' numerical FV utilizing '4' sorts of protein sequence translator, (ii) choosing negative data as of unknown genes, (iii) modeling every FV utilizing NB, (iv) ANN is utilized for making the last decision via fusing the envisaging outcomes of the base NB classifiers.

B. Protein Sequence Translation

Extracting FV aimed at disease and unknown genes is the utmost vital challenges while identifying disease-gene issues utilizing a machine learning algorithm. Here, for characterizing genes, equivalent gene products (Proteins) are utilized. Hereof, to extract the vital information of protein wherein fully encoded is taken, '4' sorts of representation techniques were utilized, they are i) Normalized Moreau–Broto autocorrelation (NA), ii) Geary autocorrelations (GA), iii) auto covariances (AC), and iv) Moran auto-correlations (MA). The reason for utilizing these representation techniques is to evade losing imperative information that is concealed in the protein sequences. All of these techniques are centered upon the physicochemical properties of amino acids since sequence of amino acid determines the protein. In other words, amino acids are the building block of protein. Here, '12' physic-chemical properties are employed as a descriptor to render more information concerning the amino acid sequence. These properties comprise entropy of formations (EOF), partitions coefficient (PC), polarity (POL), amino acid compositions (AAC), residue accessible surface areas in tripeptide (RAS), transfer-free energy (TFE), CC on regressions analysis (CC), hydrophilicity (HY-PHIL), polarizability (POL2), hydrophobicity (HY-PHOB), solvations free energy (SFE), along with graph shapes index (GSI), correspondingly. Min–Max normalization technique is employed for normalizing these physicochemical properties since it ensures that all physicochemical properties have exact same scale.

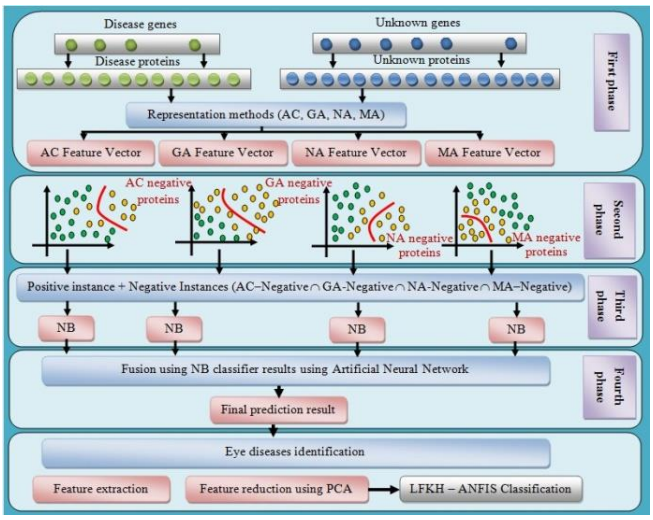

Fig. 1. Proposed Architecture.

C. Negative Data Generation

Subsequent to generating the FV for all genes, it is essential to select a negative protein set as of the unknown proteins to construct a dataset with positive as well as reliable negative instances. With regard to it, a '6' steps algorithm is proposed.

Step1: Define four negative sets as an empty set for each of the feature vectors as

$$D_{AC} = \Phi; D_{GA} = \Phi; D_{MA} = \Phi; D_{NA} = \Phi \quad (1)$$

Step2: Second, representing each protein R_i (disease and unknown proteins) into four vector: $S_{AC}^{R_i}, S_{GA}^{R}, S_{MA}^{R}, S_{NA}^{R}$ using AC, GA, MA, and NA representation methods can well be expressed as

$$S_{AC}^{R_i} = AC(R_i) \quad (2)$$

$$S_{GA}^{R_i} = GA(R_i) \quad (3)$$

$$S_{MA}^{R_i} = MA(R_i) \quad (4)$$

$$S_{NA}^{R_i} = NA(R_i) \quad (5)$$

Step3: Calculate the positive mean vector M_p of all positive proteins for every represented vectors.

$$M_p(AC) = \sum_{k=0}^{n} \frac{S_{AC}^{P_k}}{|D|} \quad (6)$$

$$M_p(GA) = \sum_{k=0}^{n} \frac{S_{GA}^{P_k}}{|D|} \quad (7)$$

$$M_p(MA) = \sum_{k=0}^{n} \frac{S_{MA}^{P_k}}{|D|} \quad (8)$$

$$M_p(NA) = \sum_{k=0}^{n} \frac{S_{NA}^{P_k}}{|D|} \quad (9)$$

Step4: Fourth, calculate the similarity, Sim_j between each unknown protein ($R_i \in U_R$) and the mean vectors (M_p) that can well be expressed as.

$$P_{U_R}^{AC}(j) = P(S_{AC}U_R, M_p(AC)) \quad (10)$$

$$P_{U_R}^{GA}(j) = P(S_{GA}U_R, M_p(GA)) \quad (11)$$

$$P_{U_R}^{MA}(j) = P(S_{MA}U_R, M_p(MA)) \quad (12)$$

$$P_{U_R}^{NA}(j) = P(S_{NA}U_R, M_p(NA)) \quad (13)$$

Step5: Fifth, for each FV, choose g negative proteins as of U_R set by selectig the g farthest proteins as of the M_p, which can well be specified as.

$$D_{AC} = sort(D_{U_R}^{AC}) \quad (14)$$

$$D_{GA} = sort(D_{U_R}^{GA}) \quad (15)$$

$$D_{MA} = sort(D_{U_R}^{MA}) \quad (16)$$

$$D_{NA} = sort(D_{U_R}^{NA}) \quad (17)$$

$$P_{AC} = select(D^{AC}(1:g)) \quad (18)$$

$$P_{GA} = select(D^{GA}(1:g)) \quad (19)$$

$$P_{MA} = select(D^{MA}(1:g)) \quad (20)$$

$$P_{NA} = select(D^{NA}(1:g)) \quad (21)$$

Manhattan distance measure is utilized as a distance measurement to gauge the distance betwixt R_j and M_p. As the number of unknown proteins is much more than disease proteins, ascertaining the appropriate number (g) of chosen negative proteins has a direct effect on the prediction model construction.

Step6: Lastly, the proteins ascertained by means of the intersection of chosen negative protein sets will be selected as reliable negative data (R_{NS}).

$$R_{NS} = P_{AC} \cap P_{GA} \cap P_{MA} \cap P_{NA} \quad (22)$$

D. Naive Bayes Algorithm

Naive Bayes (NB) stands as a probabilistic classifier stimulated by the Bayes theorem under a simple assumption, i.e., the attributes are autonomous conditionally. NB is a particularly simple algorithm to execute, and good outcomes have been attained in the utmost instances. Nevertheless, utilizing the same classifier (NB) to categorize the disparate FV of the same instances produces some uncertainties and also makes some individual errors. Therefore, a practical fusion of these classifiers will more likely lessen the overall prediction inaccuracies and renders better prediction outcomes by reducing the negative effects of noise data which proportionally increases with rising negative data ratio. Here, the ANN is utilized as a fusion method in the 4th layer. The general explanation concerning the ANN is rendered in the section below

E. Artificial Neural Network

ANN classifier comprises countless interconnected artificial neurons which have multiple interconnections connected to the adjustable weights. The inputted patterns are transmitted through the layers to solve the problem. By employing the corresponding synaptic weights, the information is mapped.

Step1: Make arbitrary weights in the interval [0, 1] and allocate it to the Hidden Layer (HL) neurons as well as the Output Layer (OL) neurons. Maintain a unity value weight for all neurons of the inputted layer for easy computation and to attain better performance together with outcomes.

Step2: Calculate the output of the hidden layer as shown in below eq.

$$H_o = B_i + \sum_{i=1}^{L} NB_i W_i \tag{23}$$

Where,

B_i - Bias value,

NB_i - Output of previous NB layer values,

W_i - Weight value of the given input features.

Step3: To find the final output unit, the hidden unit is multiplied with the weight of the hidden layer output, which is given in the equation (23).

$$O_i = B_i + \sum_{i=1}^{m} H_o W_{ik} \tag{24}$$

Where,

H_o - Hidden unit

W_{ik} - Weights of the hidden layer

O_i - Output unit.

The activation function for the output layer is estimated as

$$Active\left(o_i\right) = \frac{1}{1+e^{-O_i}} \tag{25}$$

Step4: Recognize the learning error as offered beneath

$$e_r = \sum z_i - o_i \tag{26}$$

Where,

e_r - Error rate,

z_i - Target output value,

o_i - Actual output value.

It is apparent that the NB-centered classifiers construct the model for the same dataset utilizing disparate FV. Therefore, fusing the NB-based predictors' outputs utilizing the ANN brings about concurrent utilization of optional feature descriptors along with classification procedures.

F. Feature Extraction

After the disease gene identification, this phase is done to extract the features as of the identified disease genes for classifying the eye-related diseases caused through the disease genes. The features, namely Katz Fractal Dimension (KFD), Log Energy (LE), Hurst exponent (HE), Shannon Entropy (SE), Skewness, Mean, Kurtosis, Detrended Fluctuation Analysis (DFA), Discrete Wavelet Transforms, and also Standard Deviation are extracted.

G. Feature Reduction

Following feature extraction, feature reduction is done with the utilization of PCA. PCA that conserves the existent information and eliminates the redundant constituents is employed to discover significant features. PCA acts as a linear combination where one set of variables in P_m space into another set in P_n space containing the maximum amount of variance in the data where $n < m$. This is obtained in the following steps:

Step1: Evaluate the covariance matrix "C_m" as,

$$C_m = \frac{1}{N}\sum_{k=0}^{N}\left(F_k - m\right)\left(F_k - m\right)^T \tag{27}$$

Where,

$m = \frac{1}{N}\sum_{k}^{N} F_k$ - Original feature vectors.

Step2: Determinr the eigenvectors "v_i "and Eigen values "μ_i " of the C_m by solving the subsequent decomposition

$$\mu_i v_i = L v_i \tag{28}$$

Where,

L - Matrix having the properties of eigen value and eigen vector.

Step3: Sort the outcomes in decreasing order of μ_i

Step4: Choose the indispensable components (that is, features).

H. Classification for the Identification of Eye Diseases

Here, the related eye disease prediction is performed with the utilization of the LFKH-ANFIS algorithm. Gradient-centric learning is the standard learning process in ANFIS but it is prone to trap in local minima. On this account, the ANFIS is ameliorated with the utilization of LFKH for lessening its complexity and for elevating the classification accuracy. And thereby, the ANFIS is termed as LFKH based ANFIS (LFKH-ANFIS). The ANFIS has 2 fuzzy IF-THEN rules as specified in the equations (28) and (29).

Rule i: If F_1 is A_i and F_2 is B_i then, $-_i$

$$Rules_i = cF_i + d_iF_{i+1} + e_i \quad (29)$$

Rule ii: If F_1 is A_{i+1} and F_2 is B_{i+1} then,

$$Rules_{i+1} = c_{i+1}F_i + d_{i+1}F_{i+1} + e_{i+1} \quad (30)$$

Where,

A_i, B_i, A_{i+1} and B_{i+1} - Fuzzy sets,

c_i, d_i, e_i, c_{i+1}, d_{i+1} & e_{i+1} - Predicted design parameters during training,

F_i and F_{i+1} - Disparate reduced feature values acquired as of PCA.

These provided parameters are optimized with the assist of the LFKH algorithm for attaining a better outcome. The ANFIS encompasses some layers as elucidated below,

Layer1: The first layer named a fuzzification layer gathers the input values and finds their membership functions (MF) as proffered below.

$$Z_{1,i} = \chi_i(F_i) \quad (31)$$

Where,

F_i - Input to node i,

χ_i - MF of the input F_i.

Each node here is adapted well to a functional parameter. The output acquired from each node acts as a degree of member-ship value that is provided by the input of an MF. The MF utilized in the proposed work is Gaussian kernel MF. The reason for choosing Gaussian kernel MF is to diminish the computational price of ANFIS since Gaussian kernel MF has least number of modifiable parameters. The MF used in the proposed work is specified in the succeeding equation.

$$\chi_i = \exp\left(-\frac{\|c_i - d_i\|^2}{2e_i^2}\right) \quad (32)$$

Where,

c_i, d_i and also e_i - MF parameters that could alter the MF's shape and are concerned as the premise parameters.

Layer2: This layer named the rule layer is accountable for creating the firing strengths (FS) for the rules. The incoming signals are mathematically multiplied to acquire the output that means the FS of a rule.

$$Z_{2,i} = Q_i = \chi_i(F_i) \times \chi_i(F_{i+1}) \quad (33)$$

Layer3: It aids to normalize the evaluated FSs by dividing every value with a total FS. The i^{th} node evaluates the ratio $(\overline{Q}_i)$ betwixt the i th rule's FS and the sum value of all rules' FSs to generate its output.

$$Z_{3,i} = \overline{Q}_i = \frac{H_i}{Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6},\ i = 1,2..6 \quad (34)$$

Layer4: It takes the above attained normalized values as inputs (resultant parameter sets) and it has adaptive nodes with a node function.

$$Z_{4,i} = \overline{Q}_i . Rules_i \quad (35)$$

Where,

$\overline{Q}_i$ - Normalized FSs from the former layer and

$Rules_i$ - system rule. And here, the deployed parameters are named as succeeding parameters.

Layer5: The former fourth layer proffers the defuzzificated values and these values are transmitted to the fifth layer for acquiring the final output. All incoming signals are summated to acquire overall output, and here, the circle node is labeled as $\sum$

$$Z_{5,i} = \sum_i Q_i Rules_i = \frac{\sum_i Q_i Rules_i}{\sum_i Q_i} \quad (36)$$

From the LFKH-ANFIS, the 10 classes of eye diseases for the identified disease gene, that is, Age-related Macular De-generation (AMD), cataract, Marfan syndrome, glaucoma, inherited optic neuropathies, polypoidal choroidal vasculopathies, retinitis pigmentosa, uveal melanoma, and Stargardt disease are acquired.

I. Levy Flight based Krill Herd Algorithm

The Krill Herd (KH) algorithm has the potential to effectively determine the optimum solution for certain search spaces configurations. With the futile exploration of KH's search approach, it is incompetent to assure convergence. This proposed method utilizes the Levy flight (LF) in KHA with the intention of resolving the aforesaid difficulty. Hence, the parameter tuning for ANFIS utilizing this optimization is termed as LF based KH (LF-KH). With the utilization of the Lagrangian model, the krill's location is evaluated as,

$$\frac{dXi}{dt} = H_i + F_i + D_i \quad (37)$$

Where,

H_i - Motion guided by other KI,

F_i - Foraging motion,

D_i - Physical diffusion of the i^{th} KI's.

The steps that are done in this alogorithm are,

Step1: The krill individuals (KI) endeavor to hold a high density and move on account of their mutual effects. The direction of KI motion is ascertained by the density of the local - target and repulsive - swarms. The KI movement is written as:

$$H_i^{new} = H^m \alpha_i + \omega_n H_i^{old} \tag{38}$$

Where,

H^m - Maximal induced speed,

ω_n - Motion' inertia weight in [0, 1],

H_i^{old} - Last motion-induced.

Here, α_i is evaluated as follows,

$$\alpha_i = \alpha_i^{local} + \alpha_i^{t\arg et} \tag{39}$$

Where,

α_i^{local} - Local effects of neighbors of the i^{th} individual,

$\alpha_i^{t\arg et}$ - Best solution direction as of the i^{th} individual.

The α_i^{local} in a KI movement is evaluated as:

$$\alpha_i^{local} = \sum_{j-1}^{MM} \widehat{K}_{i,j} \widehat{H}_{i,j} \tag{40}$$

$$\widehat{H}_{i,j} = \frac{H_j - H_i}{\left\| H_j - H_i \right\| + \in} \tag{41}$$

$$\widehat{K}_{i,j} = \frac{K_i - K_j}{K^{worst} - K^{best}} \tag{42}$$

Where,

K^{best} - Best-fitness (BF) values of the KIs,

K^{worst} - Worst-fitness values of the KIs

K_i - Objective function or the fitness of the i^{th} KI,

H - Associated positions,

K_j - Fitness of jth neighbors (j = 1,2,..., MM),

MM - Number of prevailing neighbors.

The least positive number termed "λ" is added to the denominator for averting the singularities. Utilizing the KIs' original behavior, a sensing distance (S_d) is evaluated as

$$S_{d,i} = \frac{1}{5M} \sum_{j=1}^{N} \left\| H_i - H_j \right\| \tag{43}$$

Where,

$S_{d,i}$ - Sensing distance for the i^{th} KI,

N - Number of KIs,

Factor 5 - Empirically acquired value. The effect of the KI with the BF on the i^{th} KI is regarded utilizing Equation (44).

$$\alpha_i^{t\arg et} = C^b \widehat{K}_{i,best} \widehat{H}_{i,best} \tag{44}$$

Where,

C^b - Effective co-efficient of the KI bearing the BF to the i^{th} KI.

This coefficient is defined because $\alpha_i^{t\arg et}$ directs the solution to the global optima and it must be more effective when analogized to other KI, that is, neighbors. Herein, the C^b is evaluated as

$$C^b = 2\left(rd + \frac{I}{I_{\max}} \right) \tag{45}$$

Where,

I - Actual iteration number,

$I_{\max}$ - Maximal iterations.

For enhancing exploration, "rd" which is a random value lies in the gamut of (0, 1) is utilized. The proposed approach utilizes the LF for the process of a random walk rather than a simpler one to overcome the incapability of KH search approach which led to its inability to ensure convergence. LF maximizes the efficiency of the searches in uncertain environments. Whilst generating a new solution $Xi^{'}$ for the i^{th} solution by performing LF, the new candidate is evaluated as,

$$Xi^{'} = X_i \oplus \alpha \operatorname{Levy}(\beta) \tag{46}$$

Where,

α - Random step size parameter

β - LF distribution parameter

$\oplus$ - Entry wise multiplication

Here, the equation (45) is rewritten as,

$$C^b = 2\left(X_i^{'} + \frac{I}{I_{\max}} \right) \tag{47}$$

Step2: The foraging motion is also known as searching motion is evaluated in respect of 2 vital effective parameters like i) food location along with ii) the prior experiences of the KIs' food location. They are evaluated as

$$Fa_i = F_s\gamma_i + \omega_I Fa_i^{old} \quad (48)$$

Where $\gamma_i = \gamma_i^{best} + B_i^{best}$ (49)

Where,

F_s - Foraging speed,

w_I - Inertia weight for foraging,

B_i^{best} - Best solution

Step3: The physical diffusion process of the KI is an arbitrary one and is the motion associated to Df_i and δ. Its equation is,

$$Df_i = Df_m\delta \quad (50)$$

Where,

Df_i - Maximal diffusion speed,

δ - Random directional vector together with its arrays of arbitrary values in (-1, 1).

The KH movement is concerned as a process on the way to the BF. So, the KI position is proffered by.

$$X_i(t+\Delta t) = X_i(t) + \Delta t\frac{dX_i}{dt} \quad (51)$$

Where,

Δt - Scale factor of the speed vector

Δt is an imperative parameter, and it must be adjusted in respect of the optimization issue. Its value is completely contingent on the provided search space.

IV. Result and Discussion

Here, the proposed system is analyzed and its performance is analogized to the existing algorithms regarding certain performance metrics. To ascertain the proposed method's robustness, to lessen the over fitting and to lessen the bias in the estimate of the classification model, 5 fold cross-validations have been employed utilizing a dataset with 10,000 instances (that is, 5000 positive and 5000 negative instances). Table I proffers the values acquired by the proposed NB-ANN predictor and some NB based classifiers regarding their prediction performance.

Table I evinces the f-measure, precision, together with recall values attained by the NB-based and fusion-based predictions. The AC-NB shows 81.6% precision, which is higher when analogized to that of GA-NB (74), NA-NB (76.54), and MA-NB (72.4). But, only the proposed fusion methodology acquires the highest precision (83.52) amongst others. Likewise, for f-measure and recall, the proposed NB-ANN classifier proffers the higher most values when analogized to other NB based approach. It is found that the fusion predictor shows the topmost performance when analogized to each NB-based predictor. As the classification of disparate FVs of the same data utilizing the same classifier generates certain uncertainties, fusing the classifier outcomes would diminish the overall classification errors. Fig. 2 evinces the comparison of NB-ANN and other existing approaches regarding f-measure, precision, together with recall.

Fig. 2 contrasts the proposed NB-ANN to some existing approaches regarding f-measure, precision, together with recall. The proposed NB-ANN acquires the 86-precision, 89.2-recall, and 88-f-measure, whereas, the existing ones acquire lower values for those measures when analogized to the proposed NB-ANN. For instance, the PUDI, ProDige, and SVM-C4.5 acquired 78.3, 72.5, and 82.4 precision values, 84.2, 78.8, and 85.2-recall values, and 80, 74.6, and 83-f-measure results. Here, the existing SVM-C4.5 shows greater performance. But, when analogized to the proposed NB-ANN, the existing ones show the least performance. From this comparison, the proposed NB-ANN is confirmed to acquire a remarkable performance for disease gene identification, and it worked well than other approaches. Then, the next experiment is performed for analyzing the proposed LFKH-ANFIS and comparing the LFKH-ANFIS with the existing techniques centered on performance regarding sensitivity, precision, specificity, recall, accuracy, f-measure, PPV, NPV, MCC, and FDR. Table II proffers the outcomes of LFKH-ANFIS and some existing algorithms.

TABLE I. Comparison of Proposed and Existing Techniques

Methods	*Precision (%)*	*Recall (%)*	*F-measure (%)*
AC-NB	81.6	73.2	74.5
GA-NB	74	84.6	80.47
NA-NB	76.54	83.2	79.8
MA-NB	72.4	88	79.6
NB-ANN	83.52	86.47	83

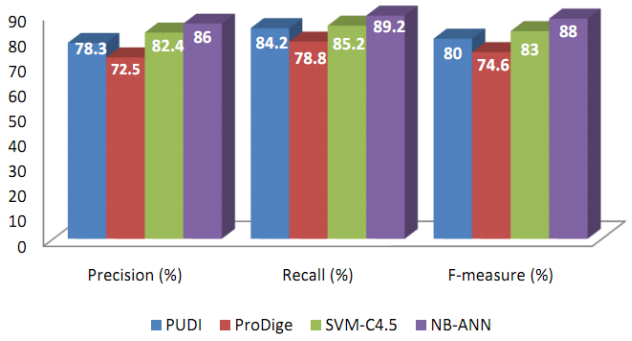

Fig. 2. Performance Graph for the NB-ANN with Existing Techniques.

TABLE II. RESULTS COMPARISON OF THE PROPOSED LFKH-ANFIS WITH EXISTING TECHNIQUES

Performance Metrics	*Proposed LFKH-ANFIS*	*ANFIS*	*DNN*	*ANN*	*KNN*
Sensitivity	0.9314	0.8442	0.7412	0.7845	0.7624
Specificity	0.9561	0.9315	0.8845	0.8412	0.9142
Accuracy	0.9947	0.8965	0.8874	0.8432	0.8398
Precision	0.9412	0.8417	0.8254	0.8347	0.7648
Recall	0.9412	0.8417	0.8254	0.8347	0.7648
F-measure	0.9412	0.8417	0.8254	0.8347	0.7648
NPV	0.9847	0.9321	0.9047	0.9075	0.9254
FPR	0.0072	0.0587	0.0784	0.0984	0.0871
FNR	0.0689	0.5245	0.7478	0.8471	0.8124
MCC	0.9343	0.6471	0.4728	0.5471	0.6547
FRR	0.0547	0.6417	0.874	0.8325	0.7841

Table II could be utilized for contrasting the results of the LFKH-ANFIS and the existing classifiers. The LFKH-ANFIS acquires 0.9412 for precision, f-measure, and recall, whereas, the existing ANFIS, DNN, ANN, and KNN proffered the values of 0.8417, 0.8254, 0.8347, and 0.7648 for precision, recall, f-measure. On considering the sensitivity, specificity, accuracy, NPV, and MCC, the LFKH-ANFIS evinces superior performance. Likewise, for MCC and NPV, the LFKH-ANFIS proffers the greatest outcomes analogized to existing algorithms. From these results, the proposed LFKH-ANFIS is confirmed to be better when analogized to other existing algorithms for eye disease identification. The error rate measures of the classification algorithm, namely FPR, FRR, and FNR, define the error that transpires at the time of performing classification.

For an effectual and excellent classification algorithm, the error rate measures must be low and that is achieved only by the proposed LFKH-ANFIS algorithm.

V. CONCLUSION

When analogized to the existing PUDI, ProDige, and SVM-C4.5, the proposed NB-ANN acquires the higher most values of precision, f-measure, and recall. Likewise, the LFKH-ANFIS shows the topmost performance by acquiring the highest results of sensitivity, precision, specificity, recall, accuracy, f-measure, NPV, and MCC when analogized to ANN, KNN, DNN, and ANFIS. And, the proposed LFKH-ANFIS acquires the lowest error rates (FNR, FPR, and FRR) for eye disease identification, which evinces the proposed method's efficiency. Therefore, the disease gene identification and the possibility of eye disease incurred by those disease genes are identified more accurately using both classification algorithms. For future work, more number of physicochemical properties of amino acids will be considered for better performance in classification. For future work, more number of physicochemical properties of amino acids will be regarded for better performance in classification.

REFERENCES

[1] Syeeda Farah, Sushma M. S, Asha T, Cauvery B, and Shivanand K. S., "DNA Based Disease Prediction Using Pathway Analysis", In IEEE 7th International Advance Computing Conference (IACC), IEEE, pp. 629-634, 2017, 10.1109/IACC.2017.0133.

[2] Ping Luo, Li-Ping Tian, Jishou Ruan, and Fang-Xiang Wu, "Disease gene prediction by integrating ppi networks, clinical rna-seq data and omim data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 1, pp. 222-232, 2017.

[3] Kuo Yang, Ruyu Wang, Guangming Liu, Zixin Shu, Ning Wang, Runshun Zhang, Jian Yu, Jianxin Chen, Xiaodong Li, and Xuezhong Zhou, "HerGePred: heterogeneous network embedding representation for disease gene prediction", IEEE journal of biomedical and health informatics, vol. 23, no. 4, pp. 1805-1815, 2018.

[4] Xiwei Tang, Xiaohua Hu, Xuejun Yang, and Yuan Sun, "A algorithm for identifying disease genes by incorporating the subcellular localization information into the protein-interaction networks", In IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 308-311, 2016, 10.1109/BIBM.2016.7822537.

[5] Lvxing Zhu, Zhaolin Hong, and Haoran Zheng, "Predicting gene-disease associations via graph embedding and graph convolutional networks", In IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 382-389, 2019, 10.1109/BIBM47256.2019.8983350.

[6] Jie Yuan, Xingpeng Jiang, Tingting He, Yan Wang, and Xiyue Guo, "Predicting disease genes based on normalized protein modules and phenotype ontology", In IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 1177- 1183, 2015, 10.1109/BIBM.2015.7359849.

[7] TengJiao 7, Wang, Wei Liu HaiLin, Tang Wei Zhang, ChangMing Xu, HanChang Sun, Hui Liu, and HongWei Xie, "Predicting potential disease-related genes using the network topological features", In Proceedings International Conference on Human Health and Biomedical Engineering, IEEE, pp. 871-876, 2011, 10.1109/HHBE.2011.6028961.

[8] Xiaochan Wang, Yuchong Gong, Jing Yi, and Wen Zhang, "Predicting gene-disease associations from the heterogeneous network using graph embedding", In IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 504-511, 2019, 10.1109/BIBM47256.2019.8983134.

[9] Jiang X, Zhang H, Duan F, and Quan X, ''Identify Huntington's disease associated genes based on restricted Boltzmann machine with RNA-seq data,'' BMC Bioinf., vol. 18, no. 1, pp. 439–447, 2017.

[10] Ngiam J, Khosla A, Kim M, Nam J, and Ng A. Y, ''Multimodal deep learning,'' in Proc. 28th Int. Conf. Mach. Learn. (ICML), Bellevue, WA, USA, pp. 1–8, 2011.

[11] Dibyendu Bikash Seal, Vivek Das, Saptarsi Goswami, and Rajat K. De, "Estimating gene expression from DNA methylation and copy number variation: A deep learning regression model for multi-omics integration", Genomics, 2020, 10.1016/j.ygeno.2020.03.021.

[12] Konstantina Kourou, Costas Papaloukas, and Dimitrios I. Fotiadis, "Identification of differentially expressed genes through a meta-analysis approach for oral cancer classification", In 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, pp. 3876-3879, 2017, 10.1109/EMBC.2017.8037703.

[13] Shin-Jye Lee, Ching-Hsun Tseng, GT–R. Lin, Yun Yang, Po Yang, Khan Muhammad, and Hari Mohan Pandey, "A dimension-reduction based multilayer perception method for supporting the medical decision making", Pattern Recognition Letters, vol. 131, pp. 15- 22, 2020.

[14] Han Zhang, Xueting Huo, Xia Guo, Xin Su, Xiongwen Quan, and Chen Jin, "A disease- related gene mining method based on weakly supervised learning model", BMC bioinformatics, vol. 20, no. 16, pp. 1-11, 2019.

[15] Xue Jiang, Jingjing Zhao, Wei Qian, Weichen Song, and Guan Ning Lin, "A Generative Adversarial Network Model for Disease Gene Prediction With RNA-seq Data", IEEE Access, vol. 8, pp. 37352-37360, 2020.

[16] BoLin Chen, Min Li, JianXin Wang, and Fang-Xiang Wu, "Disease gene identification by using graph kernels and Markov random fields", Science China Life Sciences, vol. 57,no. 11, pp. 1054-1063, 2014.

[17] Abdulaziz Yousef and Nasrollah Moghadam Charkari, "A novel method based on physicochemical properties of amino acids and one class classification algorithm for disease gene identification", Journal of Biomedical Informatics, vol. 56, pp. 300-306, 2015.

[18] Zhen Tian, Maozu Guo, Chunyu Wang, LinLin Xing, Lei Wang, and Yin Zhang, "Constructing an integrated gene similarity network for the identification of disease genes", Journal of biomedical semantics, vol. 8, no. 1, pp. 32, 2017.

[19] Mehdi Joodaki, Nasser Ghadiri, Zeinab Maleki and Maryam Lotfi Shahreza, "A scalable random walk with restart on heterogeneous networks with Apache Spark for ranking disease-causing genes using type-2 fuzzy data fusion", Biorxiv, pp. 1-20, 2019.

[20] Pradipta Maji, and Ekta Shah, "Significance and functional similarity for identification of disease genes", IEEE/ACM transactions on computational biology and bioinformatics, vol. 14, no. 6, pp. 1419-1433, 2016.