IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 May 2022, accepted 25 June 2022, date of publication 29 June 2022, date of current version 8 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3186998

RESEARCH ARTICLE

Advance Genome Disorder Prediction Model Empowered With Deep Learning

ATTA-UR-RAHMAN[1], MUHAMMAD UMAR NASIR[2], MOHAMMED GOLLAPALLI[3], MUHAMMAD ZUBAIR[4], (Senior Member, IEEE), MUHAMMAD AAMER SALEEM[5], SHAHID MEHMOOD[2], MUHAMMAD ADNAN KHAN[6], AND AMIR MOSAVI[7,8,9,10]

[1]Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia
[2]Faculty of Computing, Riphah School of Computing and Innovation, Riphah International University, Lahore Campus, Lahore 54000, Pakistan
[3]Department of Computer Information Systems, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia
[4]Faculty of Computing, Riphah International University, Islamabad 45000, Pakistan
[5]Hamdard Institute of Engineering and Technology, Hamdard University, Islamabad Campus, Islamabad 45000, Pakistan
[6]Pattern Recognition and Machine Learning Laboratory, Department of Software, Gachon University, Seongnam, Gyeonggido 13120, Republic of Korea
[7]Faculty of Civil Engineering, Technische Universität Dresden, 01067 Dresden, Germany
[8]John von Neumann Faculty of Informatics, Óbuda University, 1034 Budapest, Hungary
[9]Institute of Information Engineering, Automation and Mathematics, Slovak University of Technology in Bratislava, 811 07 Bratislava, Slovakia
[10]Institute of the Information Society, University of Public Service, 1083 Budapest, Hungary

Corresponding authors: Muhammad Adnan Khan (adnan@gachon.ac.kr) and Amir Mosavi (amir.mosavi@kvk.uni-obuda.hu)

ABSTRACT A major and essential issue in biomedical research is to predict genome disorder. Genome disorders cause multivariate diseases like cancer, dementia, diabetes, cystic fibrosis, leigh syndrome, etc. which are causes of high mortality rates around the world. In past, theoretical and explanatory-based approaches were introduced to predict genome disorder. With the development of technology, genetic data were improved to cover almost genome and protein then machine and deep learning-based approaches were introduced to predict genome disorder. Parallel machine and deep learning approaches were introduced. In past, many types of research were conducted on genome disorder prediction using supervised, unsupervised, and semi-supervised learning techniques, most of the approaches using binary problem prediction using genetic sequence data. The prediction results of these approaches were uncertain because of their lower accuracy rate and binary class prediction techniques using genome sequence data but not genome disorder patients' data with his/her history. Most of the techniques used Ribonucleic acid (RNA) gene sequence and were not often capable of handling bid data effectively. Consequently, in this study, the AlexNet as an effective convolutional neural network architecture proposed to develop an advance genome disorder prediction model (AGDPM) for predicting genome multi classes disorder using a large amount of data. AGDPM tested and compare with the pre-trained AlexNet neural network model and AGDPM gives the best results with 89.89% & 81.25% accuracy of training and testing respectively. So, the advance genome disorder prediction model shows the ability to efficiently predict genome disorder and can process a large amount of patients' genome disorder data with a multi-class prediction method. AGDPM has proved that it is capable to predict single gene inheritance disorder, mitochondrial gene inheritance disorder, and multifactorial gene inheritance disorder with respect to various statistical performance parameters. So, with the help of AGDPM biomedical research will be improved in terms to predict genetic disorders and put control on high mortality rates.

INDEX TERMS Genome disorder, AlexNet, deep learning, machine learning, artificial intelligence, data science, information systems, convolutional neural network.

The associate editor coordinating the review of this manuscript and approving it for publication was Rongbo Zhu.

I. INTRODUCTION

Approximately 2000 human diseases are caused by single causal genes (monogenic syndromes). The distinct phenotypic features of each syndrome are the biological manifestations of its underlying genes, and each differs slightly from the others. This makes establishing phenotype-gene relationships a vital biological process that assists clinicians and biologists in understanding the pathogenetic processes in the syndromes. The ability to identify the genes responsible for certain ailments simplifies patient diagnosis and provides insight into the operational network of connections and mutation. In the other words, a potential genetic illness can be detected through the disease gene identification process where the causing mutant genotypes studied. Such disease-causing genes may have single nucleotide variations, single nucleotide additions or deletions, entire gene loss, and other genetic diseases. Linkage analysis and positional cloning, followed by mutation analysis, have traditionally been used to identify disease genes. The susceptible chromosomal interval, which is the approximate location of the disease-associated candidate genes, is first identified using linkage analysis on human pedigrees. Second, the use of positional cloning to sequence a group of candidate genes in the area is discussed. Physical and transcription mapping are both included in this process. A human genetic ailment is a genetic condition caused by abnormalities in the genes or chromosomes that emerges before birth. Single-gene illnesses and complex disorders are the two types of genetic disorders. A single gene disease is produced by a single mutation in deoxyribonucleic acid structure, which results in a single fundamental deficit with severe implications. These illnesses are easily passed down through generations. As a result, mendelian diseases are occasionally used to describe these conditions.

Genetic disorders can also be multifactorial, i.e., complex diseases that reflect the pathological consequences of a mix of genetic abnormalities, lifestyle, and environmental variables, with genetic factors accounting for just a portion of the phenotypes associated with the disorders. A single gene disorder is caused by a genetic mutation in just one gene. Because this may happen in any gene, single-gene disorders can influence all aspects of functioning and are incredibly diverse [1]. Despite their clinical distinctions, all single-gene illnesses have the same biological foundation, are transmissible to offspring, and need the same critical genetic and counseling services. Accurate diagnosis, risk assessment, and information for impacted individuals and families, as well as access to risk management alternatives and help for sick adults and children. It is linked to non-nuclear mitochondrial deoxyribonucleic acid alterations. Each mitochondrial genome contains five to ten circular deoxyribonucleic acid segments. They retain their organelles as eggs during fertilization. As a result, this disease is always passed down from mother to the child. The hereditary mitochondrial condition causes mitochondrial encephalopathy, lactic acidosis, stroke-like events, and ocular disorders. Several causes of genetic illness mutations in multiple genes cause these illnesses, which are typically the result of a complicated interaction with environmental and nutritional factors. It's also known as polygenic or complicated disease [2]. Cancer, diabetes, and Alzheimer's are all caused by a multifactorial genetic disease. Machine learning is a method of genetic prediction that differs from old methods. Deep learning advancements, as well as the increase of data sets and processing capacity, have boosted its popularity in recent years. These approaches are appealing in statistical genetics [3], where the effects of a multitude of variables on a result are difficult to predict due to their ability to work in wide dimensions and identify interactions across loci without assuming additivity.

II. RELATED WORK

A. PREVIOUS WORK WITH RESPECT TO MACHINE LEARNING

Complex disorders containing a high number of genes, such as Single gene inheritance disorder (SGID), Mitochondrial gene inheritance disorder (MGID), and Multifactorial genome disorder (MGD), can have a wide range of symptoms. More precise genetic data collecting has come from recent developments in genomic technology. Hundreds of people with disorders have been detected in several large genetic investigations, such as those for SGID and MGD [4], [5]. Despite the large amount of data generated by this study, identifying the exact genes that cause disease has proved to be a difficult endeavor [6]. Genetic data have been proposed to be particularly informative because different disturbances in a single disorder module frequently produce similar phenotypes [7], and phenomena networks (where genes are appended endpoints if they indicate associated phenotypic states) are highly linked to proteins. Genome interactions and transcription factor networks [8]. Furthermore, abnormalities found in the interactome distant neighbors induce unique phenotypes [6]. Several ways for predicting disease from genes that incorporate these diverse types of data have been published [9]. A set of algorithms is utilized to merge the data into a single graph, which is then used for prediction. According to the fundamental principles of scientific studies, genetic variants implicated in intermediary variables will cease to be meaningful when put in a dependent variable with these intermediate components. When genetic variants are engaged in previously unknown pathways or processes with immeasurable intermediate components, they have the potential to enhance disease prediction beyond existing risk factors. Some disorders may be more susceptible to previously unknown paths than others. A critical, but not improbable, the point is that gene findings may lead to the identification of new etiological networks and intermediate biomarkers, which may be stronger predictors of disease than the genetic variant that led to their discovery. A binary support vector machine was utilized in this study to aggregate data from diverse sources. Because the remaining collection may contain genes

for unknown disorders, semi-supervised learning approaches [10] and adaptive and maladaptive [11] binary learning algorithms, have been suggested. In recent years, deep learning and machine learning have been successfully employed in a wide range of biological contexts. Deep learning and machine learning algorithms are powerful enough to handle large data sets with significant degrees of noise, complexity, and/or defects while generating only a few reasonable estimates regarding probability distributions and data production processes. Deep learning and machine learning approaches are primarily concerned with prediction, as opposed to the inferential approach of classic statistical methodology. In previous studies, mostly researchers worked on genome sequencing disorders using binary classes. Due to genome sequencing and binary classes data put some limitations on the results. Several researchers applied to machine and deep learning algorithms to get accurate results on genome sequence disorders, they used pre-trained algorithms and classification models to predict the diseases. Furthermore, genome sequencing results using machine and deep learning algorithms were not accurate as per the prediction requirements.

B. PREVIOUS WORK WITH RESPECT TO DEEP LEARNING

GCN-MF is a framework introduced by Xin Gao *et al.* [13] for disease-gene association task that combines Graph Convolutional Network (GCN) and matrix factorization. They were able to quantify non-linearities and leverage measured similarities thanks to the GCN. In Zeng *et al.* [14], a new relevance metric called HeteSim was utilized to priorities potential disease genes. Three heterogeneous networks were investigated in this article, based on gene-phenotype connections, protein-protein interactions, and phenotype-phenotype similarity. IN a study by Zhou [15], knowledge-based methodologies were used to predict gene-disease associations. The genes that co-occur within known gene-disease association data are used to derive gene-gene mutual information. By embedding the heterogeneous network made of genes and diseases, as well as their individual features, Li *et al.* [16] formulated a new method of disease gene prioritization based on graph convolutional networks, PGCN. A novel deep neural network model based on the features of phenotypes and genotypes has been developed by Yang *et al.* [17]. To optimise the parameters of a deep neural network and produce deep vector characteristics of diseases and genes, the model combined multi-view features of diseases and genes with feedback information from training samples. SmuDGE, a method that leverages feature learning to produce vector-based representations of phenotypes linked with an entity, was developed by Alshahrani and Hoehndorf [18]. SmuDGE can be used to compare two collections of phenotypes using a trainable semantic similarity score. For disease gene prediction, Yang *et al.* [19] created a heterogeneous disease-gene-related network (HDGN) embedding representation framework. A low-dimensional vector representation (LVR) of the nodes in the HDGN can be obtained using this approach. A heterogeneous network embedding representation algorithm was introduced by the researchers in [20], which constructed a network that integrated symptom-related associations and applied an embedding representation algorithm to derive low-dimensional node representations. It is possible to obtain candidate genes for given symptoms by comparing the similarity of their vectors. Machine learning was employed to reduce gene/non-coding RNA features in the study by Liu *et al.* [21]. Data from the Common Mind consortium was used to generate RNA-seq sequences for the dorsolateral prefrontal cortex (dlpfc). Liu *et al.* [22], researched eight common mental disorders, including ADHD, depression, anxiety, autism, intellectual disabilities, speech/language disorders, developmental delays, and oppositional defiant disorder in African Americans. 4179 samples of whole genome sequencing were obtained from blood, including 1384 of those who were diagnosed with at least one mental disorder. Table 1 shows the limitations of previous studies, and it shows that Liu *et al.* [22] applied deep learning models to center of applied genomics (CAG) biobank patients ncRNA and achieved 65 percent genetic disease prediction accuracy, and that this study has handcrafted features and data imbalance drawbacks. This work has handcrafted characteristics and unbalanced sequence drawbacks, according to Liu *et al.* [21], who used machine learning models on IncRNAs patients' data and achieved 67 percent genetic illness prediction accuracy. Yang *et al.* [20] used the LSGER deep learning model on gene sequence patients and achieved a prediction accuracy of 66.80% for genetic disease, using features such as selection approaches and imbalance sequence disadvantages. Liu *et al.* [22] used the PDGNet deep learning model on patients' genetic characteristics and were able to predict genetic disease with 73.8 percent accuracy, and this study had unbalance predict classes. The following are the study's significant contributions:

- This study presents a new deep learning model for predicting diseases caused by many genetic disorders.
- To examine the performance of the predicted class, the suggested model used a number of statistical factors.
- This research depicts the proposed model in its entirety, including its spatial and computational complexity.

III. PROPOSED MODEL

Early detection of genetic diseases is extremely beneficial to doctors and the biomedical sector in terms of prescribing medications for treatment. We propose AGDPM for the early detection of multi-class genomic abnormalities in this study. The flow of this investigation employing AGDPM and the pre-trained model of neural network AlexNet is depicted in Fig. 1. The first phase of the proposed methodology is to take the input of data and send it to the preprocessing phase. In preprocessing, phase data is processed using several pre-processing techniques to clean the outliers and fill the all-missing values. After removing the outliers of data, regression techniques were applied to pre-processed data to get top rank features to predict the genome disorder. After the best features extraction data is split into training and

TABLE 1. Limitations of previous studies.

Study	Model	Dataset	Accuracy	Limitation
Liu Y. et al [22]	Deep learning	CAG biobank (ncRNA)	65%	-Handcrafted features -Imbalance data
Yichuan Liu. et al [21]	Machine learning	IncRNAs	67%	-Handcrafted features -Imbalance sequences
Yang K. et al [20]	LSGER	Gene sequence	66.80%	-Imbalance sequences -Features selection method
Kuo Yang. et al [17]	Deep learning (PDGNet)	Genetic features	73.8%	-Imbalance predict classes

testing blocks. Training block of data sent towards AGDPM and pre-trained AlexNet to train models, at the termination of the training phase data was stored in the cloud for easy access anytime to testing block. Testing block access, the pre-processed data from pre-processed phase, and trained model from cloud to predict the genome disorder. So, both models AGDPM and AlexNet predict the disorders and we compare the results of both the newly proposed AGDPM and pre-trained AlexNet model with the help of several performance parameters. Malign patients from different genome disorders can contact doctors for treatment.

AGDPM consists of 22 layers, 7 fully connected layers, 6 batch normalization layers, 6 ReLU layers, Softmax, and a classification layer. AGDPM gets input features from genome disorder data. The first six fully connected layers consist of 100 neurons each and the last fully connected layer contains 3 neurons because prediction output contains single-gene inheritance disorder, mitochondrial gene inheritance disorder, and multifactorial gene inheritance disorder [23].

$$g^{in} = L + O + S \tag{1}$$

$$g^{out} = ReLU(g^{in}) \tag{2}$$

where L represents the input of features as matrix form with batch input, O is the number of neurons and S is for biases.

Neural networks are very sensitive to train. There are a lot of tricks and paths to overcome the limitation of data distribution. So, for data normalization AGDPM used a batch normalization layer for the normal distribution of data using zero mean and variance. Batch normalization layer to overcome the limitation of covariate shift problem [24].

$$M_y = \frac{1}{N}\sum_{p=1}^{k} g_D^{(p)} \tag{3}$$

$$Vriance_x = \left(\frac{N}{N-1}\right)\frac{1}{N}\sum_{p=1}^{k} Q_D^{2(p)} \tag{4}$$

$$z = \frac{\gamma}{\sqrt{Variance_x + e}}x + \left(\vartheta + \frac{\gamma M_y}{\sqrt{Variance_x + e}}\right) \tag{5}$$

Equations 3, 4, and 5 represents batch inference mean, batch inference variance, and batch inference shifting respectively [24].

Softmax layers are used to calculate the prediction probabilities with the help of the transformation of values between 0 and 1. So, equation 6 represents the working of the Softmax layer [25].

$$\xi\,(\vec{n})_j = \frac{\mathbf{l^{nj}}}{\sum_{\mathbf{j=1}}^{\mathbf{M}} \mathbf{l^{nj}}} \tag{6}$$

where $\vec{n}$ represent input values to the Softmax function, n_j are the values which the Softmax function takes as input like positive values, negative values, and zero which are not a valid distribution for probability. Exponential function l^{nj}converts all input values from zero to one for better probability distribution and M is the number of classes for a multi-classification purpose. Table 2, Table 3, and Table 4 elaborate the details of the feature input layer, fully connected layer, and Softmax layer respectively.

IV. DATA SET

Genome disorders dataset get from Kaggle which has open access for everyone [12]. Genome disorder data contains 22083 patients' history of single-gene inheritance disorder, mitochondrial gene inheritance disorder, and multifactorial gene inheritance disorder. This data consists of three classes, 43 independent variables and 1 is the dependent variable. So, we applied several outliers, regression, and normalization techniques to get the best features. Table 5 describes the best 24 extracted features for genome disorder prediction. So, AGDPM uses this feature-based dataset for training and testing the model.

V. PROPOSED MODEL SAPCE AND COMPUTATIONAL COMPLEXITY

In this section, we calculate the space and computational complexity of the proposed model for genome disorder prediction. From equation 7 to 18 used to calculate the space complexity of model and equation 19 to 22 used to check the computational complexity of the proposed model.

Fully connected layer:

$$|w| = n * m + k = (n+1) * m \tag{7}$$

$$|w| = \left(\sum_{n=1}^{700} +1\right) * \sum_{m=1}^{21}. \tag{8}$$

where w represents no. of total elements in weight matrix, n represents total input neurons, m represents total output

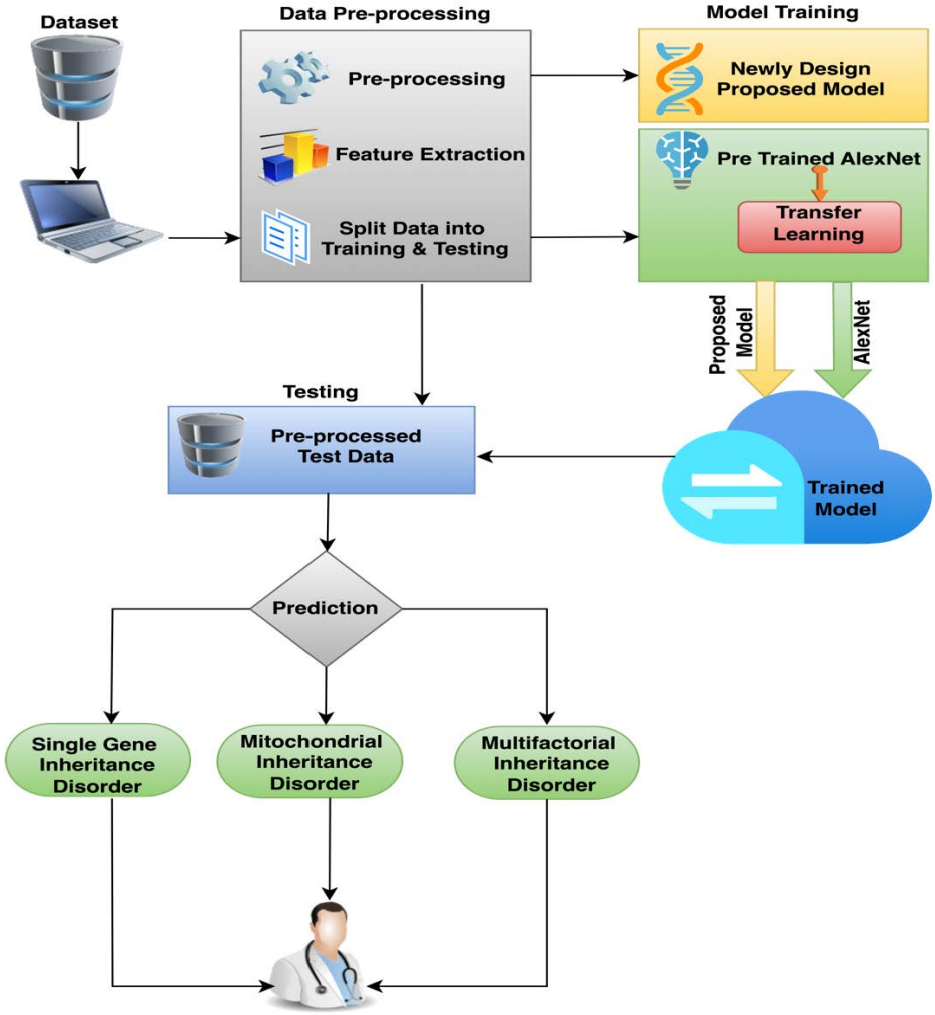

FIGURE 1. Proposed methodology for multiclass genome disorder Prediction.

TABLE 2. Feature input layer.

Feature Input Layer	
Input Size	**Normalization**
1	Zscore

TABLE 3. Fully connected layer.

Fully Connected Layer						
Input	**Output**	**Weight learns rate factor**	**Weight L2 factor**	**Bias learns rate factor**	**Weights initializer**	**Bias initializer**
Auto	50	1	1	1	glorot	zeros

neurons and k represents total bias.

$$|w| = (n+1) * m * p \; bytes \tag{9}$$

$$|w| = \left(\sum_{n=1}^{700} +1\right) * \sum_{m=1}^{21} * p \; bytes \tag{10}$$

where p represents total number of bytes per element

Model space complexity

$$100 \rightarrow (fc) * 3 \tag{11}$$

where 100 are total neurons per fully connected layer and 3 output neurons.

$$|w| = (100+1) * 3 = 303 \tag{12}$$

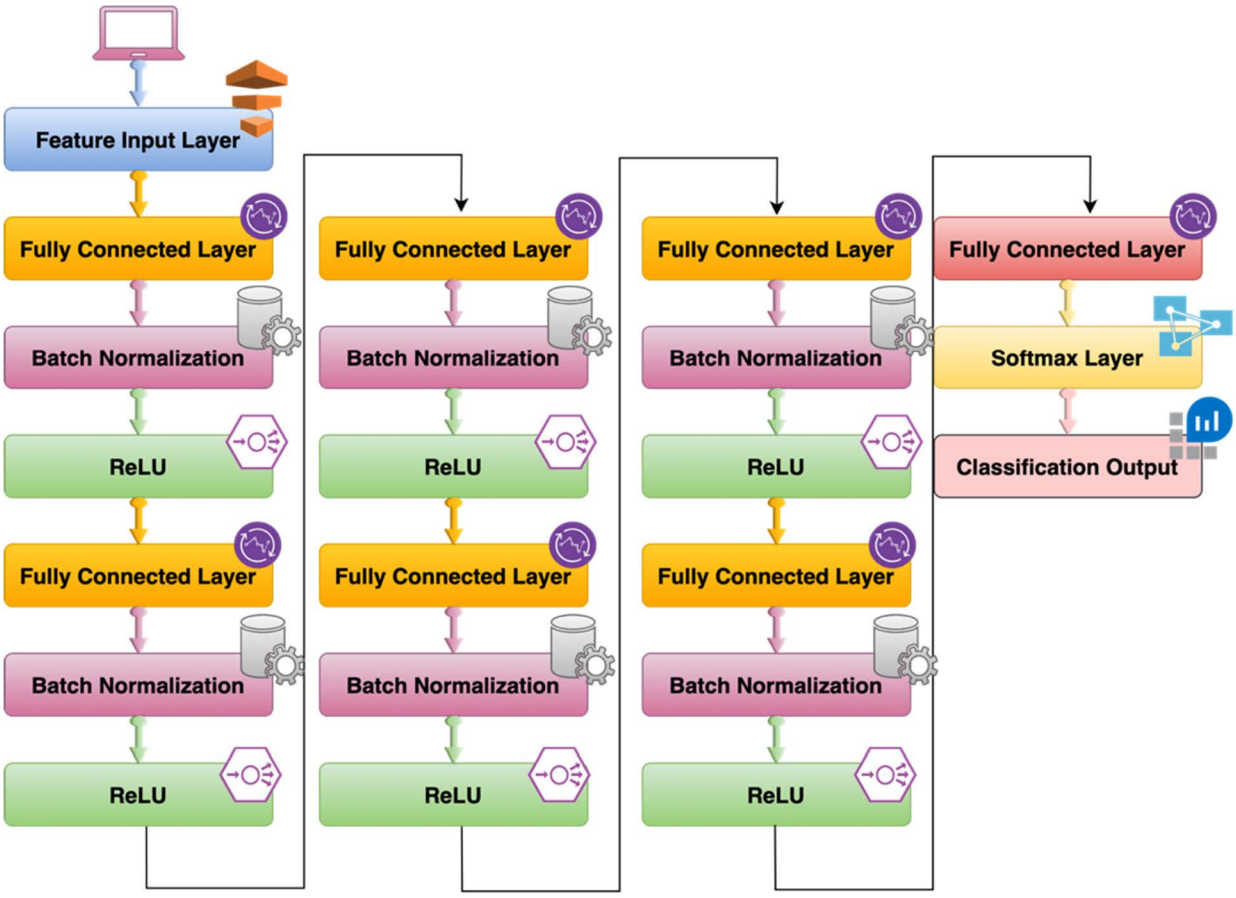

FIGURE 2. Proposed Advance genome disorder prediction model.

TABLE 4. Softmax layer.

Softmax Layer	
Mean decay	0.1
Variance decay	0.1
Epsilon	0.00001
Offset learns rate factor	1
Offset L2 factor	1
Offset initializer	Zeros
Scale initializer	Ones

From equation 4 the proposed model gets total number of elements representing matrix.

Size of weighted matrix in bytes using double precision floating point number:

$$p = 64bits/number = 8bytes/number \quad (13)$$

$$|w| = (100 + 1) * 3 * 8 = 2424bytes \quad (14)$$

where the proposed model achieved 2424 bytes model space complexity per fully connected layer using double precision floating point layer and all fully connected layers takes 16Kb to save model operations..

Single precision floating point number:

$$p = 32bits/number = 4bytes/number \quad (15)$$

$$|w| = (100 + 1) * 3 * 4 = 1212bytes \quad (16)$$

where the proposed model achieved 1212 bytes model space complexity per fully connected layer using double precision floating point layer and all fully connected layers takes 8Kb to save model operations.

Integer fixed point:

$$p = 8bits/number = 1bytes/number \quad (17)$$

$$|w| = (100 + 1) * 3 * 1 = 303bytes \quad (18)$$

where the proposed model achieved 303 bytes model space complexity per fully connected layer using double precision floating point layer and all fully connected layers takes 2Kb to save model operations. So, the proposed model occupied lesser space using integer fixed point.

TABLE 5. Dataset description [12].

No	Feature	Type
1	Patient Age	0-15
2	Genes in mother	1-Yes 2-No
3	Inherited from father	1-Yes 2-No
4	Maternal gene	1-Yes 2-No
5	Paternal gene	1-Yes 2-No
6	Status	1-Alive 2-Dead
7	Respiratory Rate	1-Normal (30-60) 2-Tachypenia
8	Heart Rate	1-Normal 2-Tachrydia
9	Follow-up	1-High 2-Low
10	Gender	1-Male 2-Female 3-Ambiguous
11	Birth asphyxia	1-Yes 2-No
12	Autopsy birth defect	1-Yes 2-No
13	Folic acid	1-Yes 2-No
14	H/O serious maternal illness	1-Yes 2-No
15	H/O radiation exposure (x-ray)	1-Yes 2-No
16	H/O substance abuse	1-Yes 2-No
17	Assisted conception	1-Yes 2-No
18	History of anomalies in previous pregnancies	1-Yes 2-No
19	No. of abortion	0-6
20	Birth defects	1-Singular 2-Multiple
21	Blood test result	1-Normal 2-Inconclusive 3-Slightly abnormal
22	Symptom 1	1-Yes 2-No
23	Symptom 2	1-Yes 2-No
24	Symptom 3	1-Yes 2-No

Model Computational Complexity:

$$\#multiply = n \tag{19}$$

$$\#add = n \tag{20}$$

where equation 21 represents total no of operations per output neurons. So,

$$Operation = 2{*}n * k \tag{21}$$

$$Operation = 2{*}\sum_{n=1}^{700} * \sum_{k=1}^{21}. \tag{22}$$

The proposed model achieved 6000 operations per second to compute the problems per fully connected layer and collectively all fully connected layers operate 42000 operations per second which is much better and faster to predict the problems.

VI. SIMULATION RESULTS

In this research MATLAB, 2021 was used for simulation and prediction purposes. AGDPM and pre-trained AlexNet model is used to train and test the dataset. At the beginning of the simulation process, the dataset total attributes split into training (70%) 15,458 attributes and testing (30%) 6661 attributes. So, we applied both AGDPM and AlexNet models on training and testing data, and choose the best-performed model to predict the single-gene inheritance disorder, mitochondrial inheritance disorder, and multifactorial inheritance disorder. Before the selection of the best-predicted model, we applied several statistical performance parameters [26]–[31] like Miss classification rate (MCR), sensitivity, specificity, Positive predicted value (PPV), Classification accuracy (CA), Negative predicted value (NPV), False-negative ratio (FNR), f1-score, Likelihood positive ratio (LPR), False positive ratio (FPR), Likelihood negative ratio (LNR) and Fowlkes-Mallows index (FMI) on predicted results. Statistical performance parameters are described below in the form of mathematical equations with respect to the confusion matrix and in these equations ∂, $\emptyset$, μ & Ω represents the true-positive, false-positive, true-negative results and false-negative results respectively.

$$\partial_i = {}^{i}/_{\Lambda_i} \tag{23}$$

$\therefore$ ב is for predicted class and Λ for true class, (24)–(38), as shown at the bottom of the next page.

Fig. 3 shows the training progress of the advance genome disorder prediction model on 100 epochs. AGDPM tested its training progress on 20, 30, 50, 60, 80, and 100 epochs and AGDPM achieved its best training progress results on 100 epochs of 89.67% accuracy with smooth convergence and 10.33% training loss rate. Table 6 shows the descriptive results of AGDPM training progress on each tested epoch. It clearly observed that at 100 epoch's model achieved the 89.67% & 10.33% accuracy and loss rate respectively.

Table 7 elaborates the predicted results of single-gene inheritance disorder, mitochondrial gene inheritance disorder, and multifactorial gene inheritance disorder using the confusion matrix.

Table 8 shows the AGDPM predicted simulation results of single-gene inheritance disorder with the help of several statistical performance parameters in which single-gene inheritance disorder achieved 83.55%, 87.83%, 74.92%, 16.45%, 75.12%, 75.33%, 87.59%, 12.17%, 25.08%, 6.15%, 0.29% and 75.13% of CA, specificity, sensitivity, CMR, f1-score, PPV, NPV, FPR, FNR, LPR, LNR, & FMI respectively. Table 9 demonstrates the AGDPM predicted

simulation results of mitochondrial gene inheritance disorder utilizing numerous statistical parameters where single-gene inheritance disorder obtained 84.28%, 86.99%, 80.58%, 15.72%, 81.23%, 81.89%, 85.98%, 13.01%, 19.42%, 6.19%,

$$\mu_i = \sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right) \tag{24}$$

$$\emptyset_i = \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_i \right) \tag{25}$$

$$\Omega_i = \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_{j \neq i} \right) \tag{26}$$

$$\text{CA} = \frac{\beth_i / \Lambda_i + \sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right)}{\beth_i / \Lambda_i + \sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right) + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_i \right) + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_{j \neq i} \right)} * 100 \tag{27}$$

$$\text{CMR} = 100 - \left(\frac{\beth_i / \Lambda_i + \sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right)}{\beth_i / \Lambda_i + \sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right) + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_i \right) + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_{j \neq i} \right)} * 100 \right) \tag{28}$$

$$\text{Sensitivity} = \frac{\beth_i / \Lambda_i}{\beth_i / \Lambda_i + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_{j \neq i} \right)} * 100 \tag{29}$$

$$\text{Specifity} = \frac{\sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right)}{\sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right) + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_i \right)} * 100 \tag{30}$$

$$\text{F1-Score} = \frac{2 \beth_i / \Lambda_i}{2 \beth_i / \Lambda_i + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_i \right) + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_{j \neq i} \right)} * 100 \tag{31}$$

$$\text{PPV} = \frac{\beth_i / \Lambda_i}{\beth_i / \Lambda_i + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_i \right)} * 100 \tag{32}$$

$$\text{NPV} = \frac{\sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right)}{\sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right) + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_{j \neq i} \right)} * 100 \tag{33}$$

$$\text{FPR} = 100 - \left(\frac{\sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right)}{\sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right) + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_i \right)} * 100 \right) \tag{34}$$

$$\text{FNR} = 100 - \left(\frac{\beth_i / \Lambda_i}{\beth_i / \Lambda_i + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_{j \neq i} \right)} * 100 \right) \tag{35}$$

$$\text{LPR} = \frac{\frac{\beth_i / \Lambda_i}{\beth_i / \Lambda_i + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_{j \neq i} \right)} * 100}{100 - \left(\frac{\sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right)}{\sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right) + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_i \right)} * 100 \right)} \tag{36}$$

$$\text{LNR} = \frac{100 - \left(\frac{\beth_i / \Lambda_i}{\beth_i / \Lambda_i + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_{j \neq i} \right)} * 100 \right)}{\frac{\sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right)}{\sum_{j=1}^{3} \left(\beth_i / \Lambda_{j \neq i} \right) + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_i \right)} * 100} \tag{37}$$

$$\text{FMI} = \sqrt{\left(\frac{\beth_i / \Lambda_i}{\beth_i / \Lambda_i + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_{j \neq i} \right)} * 100 \right) * \left(\frac{\beth_i / \Lambda_i}{\beth_i / \Lambda_i + \sum_{j=1}^{3} \left(\beth_{j \neq i} / \Lambda_i \right)} * 100 \right)} \tag{38}$$

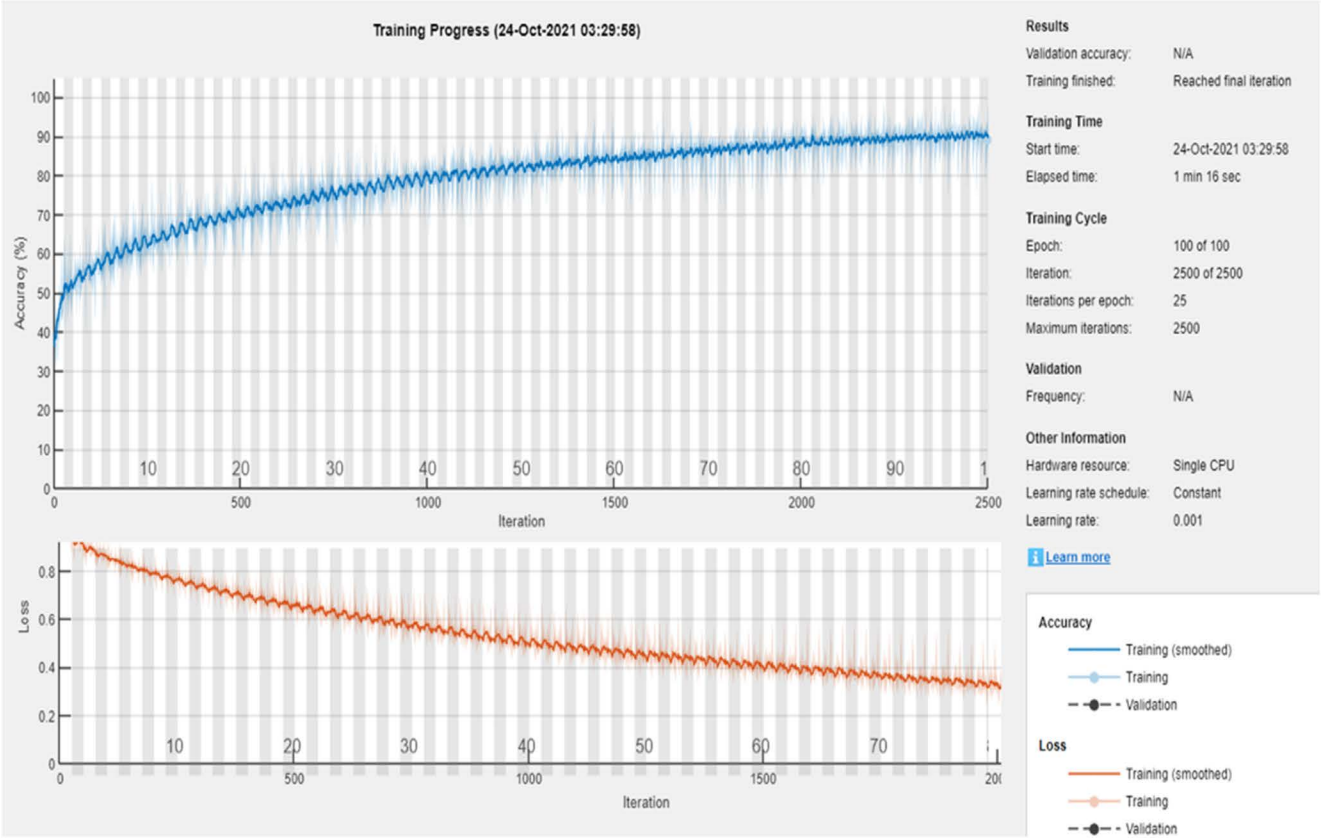

FIGURE 3. Proposed Advance genome disorder prediction model training progress.

TABLE 6. Training progress of proposed AGDPM.

No. Epochs	Training Accuracy (%)	Training Loss Rate (%)
20	68.94	31.06
30	71.09	28.91
50	71.10	28.90
60	79.69	20.31
80	81.28	18.72
100	89.67	10.33

TABLE 7. Testing confusion matrix of proposed AGDPM.

Attributes (6,661)	Single-gene inheritance disorder	Mitochondrial gene inheritance disorder	Multifactorial gene inheritance disorder
Single-gene inheritance disorder	1655	453	89
Mitochondrial gene inheritance disorder	441	2265	60
Multifactorial gene inheritance disorder	113	93	1492

0.22% and 81.23% of CA, specificity, sensitivity, CMR, f1-score, PPV, NPV, FPR, FNR, LPR, LNR, & FMI respectively. Table 10 illustrates the AGDPM predicted simulation results of multifactorial gene inheritance disorder using several statistical performance parameters, in which single-gene inheritance disorder reached 94.67%, 95.90%,90.92%, 5.33%, 89.37%, 87.87%, 97.00%, 4.10%, 9.08%, 22.16%, 0.09%, and 89.38% of CA, specificity, sensitivity, CMR, f1-score, PPV, NPV, FPR, FNR, LPR, LNR, & FMI respectively. Table 11 elaborates the comparative testing results of AGDPM and respectively. As mentioned in table 9, AlexNet attained 58.68% classification accuracy and 41.32% loss rate and AGDPM accomplished 81.25% classification accuracy and 18.75% loss rate. So, AGDPM performed better as

TABLE 8. Testing simulation results of proposed single-gene inheritance disorder using AGDPM.

Instances (6,661)	CA(%)	CMR(%)	Sensitivity(%)	Specificity(%)	F1-Score(%)
AGDPM	83.55	16.45	74.92	87.83	75.12
	NPV(%)	**FPR(%)**	**FNR(%)**	**LPR(%)**	**LNR(%)**
AGDPM	87.59	12.17	25.08	6.15	0.29
	PPV(%)	**FMI (%)**			
AGDPM	75.33	75.13			

TABLE 9. Testing simulation results of proposed mitochondrial gene inheritance disorder using AGDPM.

Instances (6,661)	CA(%)	CMR(%)	Sensitivity(%)	Specificity(%)	F1-Score(%)
AGDPM	84.28	15.72	80.58	86.99	81.23
	NPV(%)	**FPR(%)**	**FNR(%)**	**LPR(%)**	**LNR(%)**
AGDPM	85.98	13.01	19.42	6.19	0.22
	PPV(%)	**FMI (%)**			
AGDPM	81.89	81.23			

TABLE 10. Testing simulation results of proposed multifactorial gene inheritance disorder using AGDPM.

Instances (6,661)	CA(%)	CMR(%)	Sensitivity(%)	Specificity(%)	F1-Score(%)
AGDPM	94.67	5.33	90.92	95.90	89.37
	NPV(%)	**FPR(%)**	**FNR(%)**	**LPR(%)**	**LNR(%)**
AGDPM	97.00	4.10	9.08	22.16	0.09
	PPV(%)	**FMI (%)**			
AGDPM	87.87	89.38			

TABLE 11. Testing comparative results of AlexNet vs proposed AGDPM.

Model	Testing Classification Accuracy (%)	Testing Loss Rate (%)
AlexNet	58.68	41.32
AGDPM	81.25	18.75

TABLE 12. Comparative analysis with previous studies.

Study	Model	Dataset	Accuracy	Miss-Classification
Liu Y. et al [22]	Deep learning	CAG biobank (ncRNA)	65%	35%
Yichuan Liu. et al [21]	Machine learning	IncRNAs	67%	32%
Yang K. et al [20]	LSGER	Gene sequence	66.80%	33.2%
Kuo Yang. et al [17]	Deep learning (PDGNet)	Genetic features	73.8%	26.2%
The Proposed Model	**Deep learning (AGDPM)**	**Patient's clinical features data**	**81.25%**	**18.75%**

compared with the pre-trained AlexNet model for the prediction of genome disorder. Table 12 illustrates the comparative analysis of the proposed model with previous studies and it depicts, Liu *et al.* [22] applied deep learning models on center of applied genomics (CAG) biobank patients' ncRNA and achieved 65% prediction genetic disease accuracy and 35% miss-classification rate, Yichuan Liu. et al [21] applied machine learning models on IncRNAs patients' data and achieved 67% prediction genetic disease accuracy and 32% miss-classification rate, Yang *et al.* [20] applied LSGER deep learning model on gene sequence patients and achieved 66.80% prediction genetic disease accuracy and 33.2% miss-classification rate, Liu *et al.* [22] applied PDGNet deep learning model on patients' genetic features and achieved

73.8% prediction genetic disease accuracy and 26.2% miss-classification rate and at the end the proposed model AGDPM uses patients clinical features base data and achieved 81.25% prediction accuracy and 18.75% miss-classification which is far better than all previous studies because of the proposed model used perfect architecture of fully connected and convolutional layers for the prediction of this disease and also having a perfect space and computational complexity.

VII. CONCLUSION

In the field of biomedical research, artificial intelligence development has had a massive impact. In this research, we created a new model, AGDPM, and used the pre-trained AlexNet model as well. AGDPM and AlexNet were trained and tested on genome disorder data obtained from an online repository, and the performance of both models was evaluated using various statistical performance parameters. With 81.25 percent prediction accuracy, AGDPM outperforms AlexNet in the prediction of single-gene inheritance disorder, mitochondrial gene inheritance disorder, and multifactorial gene inheritance disorder. The AGDPM will significantly contribute to biomedical research to help predict genetic diseases. To get more accurate and enhanced prediction results, this research can be expanded to include more genetic disorders and more than one prediction model in the future.

ACKNOWLEDGMENT

(Atta-ur-Rahman and Muhammad Umar Nasir are co-first authors.)

REFERENCES

[1] McKusick-Nathans Institute of Genetic Medicine. *Online Mendelian Inheritance in Man Johns Hopkins University School of Medicine*. Accessed: Nov. 1, 2021. [Online]. Available: www.ncbinlmnih.gov/omim

[2] B. Irom, "Genetic disorders: A literature review," *Genet. Mol. Biol. Res.*, vol. 4, no. 2, p. 30, 2020.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

[4] S. J. Sanders, "First glimpses of the neurobiology of autism spectrum disorder," *Current Opinion Genet. Develop.*, vol. 33, pp. 80–92, Aug. 2015.

[5] Europe PMC Funders Group, "Biological insights from 108 schizophrenia-associated genetic loci," *Nature*, vol. 511, no. 7510, pp. 421–427, Jul. 2014.

[6] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabasi, "Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, no. 6224, Feb. 2015, Art. no. 1257601.

[7] A. L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: A network-based approach to human disease," *Nature Rev. Genet.*, vol. 12, pp. 56–68, Oct. 2011.

[8] M. Vidal, M. E. Cusick, and A.-L. Barabási, "Interactome networks and human disease," *Cell*, vol. 144, no. 6, pp. 986–998, Mar. 2011.

[9] X. Wang, N. Gulbahce, and H. Yu, "Network-based methods for human disease gene prediction," *Briefings Funct. Genomics*, vol. 10, no. 5, pp. 280–293, 2011.

[10] T.-P. Nguyen and T.-B. Ho, "Detecting disease genes based on semi-supervised learning and protein–protein interaction networks," *Artif. Intell. Med.*, vol. 54, no. 1, pp. 63–71, Jan. 2012.

[11] P. Yang, X. L. Li, J. P. Mei, C. K. Kwoh, and S. K. Ng, "Positive-unlabeled learning for disease gene identification," *Bioinformatics*, vol. 28, no. 20, pp. 2640–2647, 2012.

[12] A. Rishabh. *Of Genomes and Genetics HackerEarth Machine Learning Challenge*. Kaggle. Accessed: Oct. 27, 2021. [Online]. Available: https://www.kaggle.com/aryarishabh/of-genomes-and-genetics-hackerearth-ml-challenge

[13] P. Han, P. Yang, P. Zhao, S. Shang, Y. Liu, J. Zhou, X. Gao, and P. Kalnis, "GCN-MF: Disease-gene association identification by graph convolutional networks and matrix factorization," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 705–713.

[14] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and validation of disease genes using HeteSim scores," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 3, pp. 687–695, May 2017.

[15] H. Zhou and J. Skolnick, "A knowledge-based approach for predicting gene–disease associations," *Bioinformatics*, vol. 32, no. 18, pp. 2831–2838, Sep. 2016.

[16] Y. Li, H. Kuwahara, P. Yang, L. Song, and X. Gao, "PGCN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks," *bioRxiv*, vol. 2019, Jan. 2019, Art. no. 532226, doi: 10.1101/532226.

[17] K. Yang, Y. Zheng, K. Lu, K. Chang, N. Wang, Z. Shu, J. Yu, B. Liu, Z. Gao, and X. Zhou, "PDGNet: Predicting disease genes using a deep neural network with multi-view features," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 1, pp. 575–584, Jan. 2022, doi: 10.1109/TCBB.2020.3002771.

[18] M. Alshahrani and R. Hoehndorf, "Semantic disease gene embeddings (SmuDGE): Phenotype-based disease gene prioritization without phenotypes," *Bioinformatics*, vol. 34, no. 17, pp. i901–i907, Sep. 2018.

[19] K. Yang, R. Wang, G. Liu, Z. Shu, N. Wang, R. Zhang, J. Yu, J. Chen, X. Li, and X. Zhou, "HerGePred: Heterogeneous network embedding representation for disease gene prediction," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1805–1815, Jul. 2019.

[20] K. Yang, N. Wang, G. Liu, R. Wang, J. Yu, R. Zhang, J. Chen, and X. Zhou, "Heterogeneous network embedding for identifying symptom candidate genes," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 11, pp. 1452–1459, Nov. 2018.

[21] Y. Liu, H.-Q. Qu, X. Chang, L. Tian, J. Qu, J. Glessner, P. M. A. Sleiman, and H. Hakonarson, "Machine learning reduced gene/non-coding RNA features that classify schizophrenia patients accurately and highlight insightful gene clusters," *Int. J. Mol. Sci.*, vol. 22, no. 7, p. 3364, Mar. 2021.

[22] Y. Liu, H.-Q. Qu, F. D. Mentch, J. Qu, X. Chang, K. Nguyen, L. Tian, J. Glessner, P. M. A. Sleiman, and H. Hakonarson, "Application of deep learning algorithm on whole genome sequencing data uncovers structural variants associated with multiple mental disorders in African American patients," *Mol. Psychiatry*, vol. 27, no. 3, pp. 1469–1478, Mar. 2022, doi: 10.1038/s41380-021-01418-1.

[23] *Rectifier/(Neural/Networks)*. Accessed: Nov. 4, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Rectifier_(neural_networks)

[24] *Statistics#03—Standard Deviation and Variance*. Accessed: Nov. 4, 2021. [Online]. Available: https://towardsdatascience.com/statistics-03-standard-deviation-and-variance-9724f33b58df

[25] *Softmax Activation Function—How It Actually Works*. Accessed: Nov. 4, 2021. [Online]. Available: https://towardsdatascience.com/softmax-activation-function-how-it-actually-works-d292d335bd78

[26] A.-U. Rahman, S. Abbas, M. Gollapalli, R. Ahmed, S. Aftab, M. Ahmad, M. A. Khan, and A. Mosavi, "Rainfall prediction system using machine learning fusion for smart cities," *Sensors*, vol. 22, no. 9, p. 3504, May 2022.

[27] M. Saleem, S. Abbas, T. M. Ghazal, M. A. Khan, N. Sahawneh, and M. Ahmad, "Smart cities: Fusion-based intelligent traffic congestion control system for vehicular networks using machine learning techniques," *Egyptian Informat. J.*, vol. 6, pp. 1–10, Apr. 2022.

[28] M. W. Nadeem, H. G. Goh, M. A. Khan, M. Hussain, M. F. Mushtaq, and V. A. Ponnusamy, "Fusion-based machine learning architecture for heart disease prediction," *Comput., Mater. Continua*, vol. 67, no. 2, pp. 2481–2496, 2021.

[29] S. Y. Siddiqui, A. Athar, M. A. Khan, S. Abbas, Y. Saeed, M. F. Khan, and M. Hussain, "Modelling, simulation and optimization of diagnosis cardiovascular disease using computational intelligence approaches," *J. Med. Imag. Health Informat.*, vol. 10, no. 5, pp. 1005–1022, May 2020.

[30] N. Taleb, S. Mehmood, M. Zubair, I. Naseer, B. Mago, and M. U. Nasir, "Ovary cancer diagnosing empowered with machine learning," in *Proc. Int. Conf. Bus. Anal. Technol. Secur. (ICBATS)*, Feb. 2022, pp. 1–6.

[31] A.-U. Rahman, A. Alqahtani, N. Aldhafferi, M. U. Nasir, M. F. Khan, M. A. Khan, and A. Mosavi, "Histopathologic oral cancer prediction using oral squamous cell carcinoma biopsy empowered with transfer learning," *Sensors*, vol. 22, no. 10, p. 3833, May 2022.

ATTA-UR-RAHMAN received the B.S. degree in computer science from the University of the Punjab, Lahore, Pakistan, in 2004, the M.S. degree in EE from the International Islamic University, Islamabad, Pakistan, in 2008, and the Ph.D. degree in EE from ISRA University, Islamabad, in 2012. Since 2003, he has been involved in teaching and research. He is currently working as an Assistant Professor with the College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University (IAU), Dammam, Saudi Arabia. He has authored/coauthored more than 100 publications in conferences, books, and journals of good reputation. His research interests include digital communication, DSP, information and coding theory, AI, and applied soft computing.

MUHAMMAD UMAR NASIR received the B.S. degree from COMSATS University Islamabad, Pakistan, and the M.S. degree from RIU Lahore, Pakistan, by obtaining the Scholarship Award from PEEF, in 2019, and secured the Chancellor Gold Medal in his M.S. degree. He is currently pursuing the Ph.D. degree with Riphah International University, Islamabad, Pakistan. He is working as a Lecturer at Riphah International University, Lahore, Pakistan. He has supervised five M.S. scholars. He has published seven research articles in international impact factor journals and eight papers in reputed international conferences. His research interests include machine learning, medical diagnosis, soft computing, image processing, and computational intelligence.

MOHAMMED GOLLAPALLI is currently an Assistant Professor with over 18 years of both academic and industry experience which was pivotal in doing this industry field training-based education research for computer information systems (CIS) students of both undergraduate and postgraduate higher studies. He is a member of MCP and is actively involved in promoting community-based educational services. His research interests include industry-driven education, best practices, health data mining, knowledge extraction, and data science.

MUHAMMAD ZUBAIR (Senior Member, IEEE) received the M.S. degree in information technology from Hamdard University, Islamabad, Pakistan, and the Ph.D. degree in EE from the International Islamic University, Islamabad. He is currently working as a Professor at the Riphah Faculty of Computing, Riphah International University, Islamabad. His research interests include multimedia analysis and processing with a focus on medical, information security, machine learning, and deep learning.

MUHAMMAD AAMER SALEEM received the M.S. degree in information technology and the Ph.D. degree in EE from Hamdard University, Islamabad Campus, Islamabad, Pakistan. He is currently working as a Professor at the Hamdard Institute of Engineering and Technology, Hamdard University. He has been teaching graduate and undergraduate students in computer science and engineering for the past 20 years. His research interests include MUD, channel estimation, and multimedia analysis and processing with a focus on medical, information security, machine learning, and deep learning.

SHAHID MEHMOOD received the master's degree in computer science from Riphah International University, Lahore, Pakistan. He is currently pursuing the Ph.D. degree in computing with Riphah International University, Islamabad, Pakistan. He joined the Riphah School of Computing and Innovation, Riphah University, Lahore, as a Lecturer, in 2019. His research interests include machine learning, deep learning, image processing, and big data analytics.

MUHAMMAD ADNAN KHAN received the B.S. and M.S. degrees from the International Islamic University, Islamabad, Pakistan, by obtaining the Scholarship Award from the Punjab Information and Technology Board, Government of Punjab, Pakistan, and the Ph.D. degree from ISRA University, Islamabad, by obtaining the Scholarship Award from the Higher Education Commission, Islamabad, in 2016. He is currently working as an Assistant Professor with the Pattern Recognition and Machine Learning Laboratory, Department of Software, Gachon University, Republic of Korea. Before joining Gachon University, he worked in various academic and industrial roles in Pakistan. He has been teaching graduate and undergraduate students in computer science and engineering for the past 12 years. He is guiding five Ph.D. scholars and six M.Phil. scholars. He has published more than 200 research articles with a Cumulative JCR-IF and more than 450 in reputed international journals as well as international conferences. His research interests primarily include machine learning, MUD, image processing and medical diagnosis, and channel estimation in multi-carrier communication systems using soft computing.

AMIR MOSAVI received the M.Sc. degree from Kingston University London, U.K., and the Ph.D. degree in applied informatics. He was a Senior Research Fellow at Oxford Brookes University and the Queensland University of Technology. He was a recipient of the Green-Talent Award, the Alexander von Humboldt Award, the UNESCO Young Scientist Award, the Alain Bensoussan Fellowship, the GO STYRIA Award, the Estonian Dora Plus Grant, the Estophilus Scholarship, the Future Talent TU Darmstadt, the World Academy of Sciences UNESCO Award, the Marie Curie, the Endeavour-Australia Leadership, the Talented Young Scientist Award, and the European Research Consortium for Informatics and Mathematics Fellowship.