Applied Cell Biology

ISSN : 2320-1991 Online
ISSN : 2320-1983 Print
Volume 10 Issue 3
DOI: https://dx.doi.org/10.53043/2320-1991.acb90029

Review Article
Appl Cell Biol, 10(3), 2022 [77-94]

Seq_B_LSTM_CNN_HPO: Rare Mendelian Diseases to Genotypes Associations from Multiple Data Sources

Mohamed Elhajabdou[1]*, Amr Maged Ehelw[2], Hassan Eldib[3] and Mohamed Elhabrouk[4]

[1]Faculty of Engineering, Arab Academy for Science and Technology and Maritime Transport, Alexandria, Egypt

[2]Faculty of Engineering, Alexandria University, Alexandria, 21544, Egypt

***Corresponding author:** Mohamed Elhajabdou, Faculty of Engineering, Arab Academy for Science and Technology and Maritime Transport, Alexandria, Egypt

ABSTRACT

Motivation: Genotype-Phenotype annotations have become a crucial tool for studying the abnormalities in phenotype diseases. These abnormalities and relations can help to understand more the complex, and hidden information. This information clearly describes the genetic mutations causes in the organisms such as human. Several systems and algorithms have been proposed and implemented to solve this issue, since the digital information is provided for free online from different resources that describe the human mutations and the different variations in genes. Machine learning, especially deep artificial neural network, has proven its ability to overcome the limitations of these traditional algorithms and remarkably performing at extraordinary accuracies compared to conventional methods such as statistical techniques and others.

Results: In this paper, a multilabel hyper-artificial neural networks model classifier is proposed and implemented for predicting rare mendelian diseases. It is called Seq_B_LSTM_CNN_HPO. The proposed system trained on more than 50 features obtained from four data sources, Gene Ontology (GO), Human Phenotype Ontology (HPO), UniProtKB, and Gene Expressions to learn complex features and relations. The proposed system was tested on UniProtKB dataset and compared with different proposed systems in the fields. The experiment was performed on human organism for variety of analytical study in order to find new relations between phenotype diseases. The tabulated results are evaluated using six different unique evaluation metrics with outstanding results scores of Fmax, Precision, Recall, AUPR, AUROC, Smin with scores of 0.894, 0.902, 0.886, 0.711, 0.631, 0.384 which outperformed several proposed systems in the literature.

Data and Source Code Availability: The source code is provided at GitHub repository and the dataset is uploaded at Google_Drive

INTRODUCTION

Phenotype [1] is a collection of several characteristics that can be physically observed in the organism. These include but are not limited to the appearance such as eye and hair color as well as height. They also include behavior such as body movements, etc... Genotype is a genetic characteristic of the organism.

Review Article

The growth of genetics data is expanding exponentially compared to the phenotypes data. This is due to the fact that they require huge research and clinical trials on real patients especially for humans which requires time and money for such application. Moreover, the phenotype information is very complex to understand from one view. This could lead the researcher to report the results based on his/her background and research area neglecting the other factors. As an example, a certain data may give extensive information about a blood study neglecting age, and other essential factors. As a result, the data generated form such studies are biased.

Phenotype is a result of interaction between environment development and genotype [2]. The process from genotype to phenotype is by the transcription from DNA to RNA and translation from RNA to different proteins. Proteins/genes can change the patterns and the time of the gene expressions. Any mis-functionality of any of these proteins/genes' functions can lead to phenotype diseases.

In this paper a hyper model Deep Neural Network model is trained and tested on data for phenotype Mendelian phenotype diseases. The data contains 55 unique features obtained from multiple databases, which are Gene Ontology annotations [3], UniProtKB [4], Human Phenotype Ontology (HPO) annotations [5] and Gene expression data annotations. The proposed system is compared against methods from the top 10 CAFA2 challenge, and other methods proposed in the field. The results are tabulated and performance is tested using six different evaluation metrics (AUROC, AUPR, Fmax, Prec, Rec, and Smin).

RELATED WORK

Genotype-Phenotype databases

Genotype-phenotype databases give insights and more details about the genetic development, variations and other characteristics. These characteristics could be diagnosed due to the interaction between the genes/proteins with the environment. This interaction could produce abnormalities in different forms which leads to disorder diseases.

Several kinds of diseases are provided, in Cancer Genomics Hub (CGH) [6], which is a database provided to study cancer genomics diseases and its variations. CGH is one of the largest databases provided in cancer phenotype diseases which currently contains more than 2 Petabytes of data. Another very large-scale database provided for the same disease called Catalogue of Somatic Mutations in Cancer (COSMIC) [7]. It contains more than 2 million unique variations, which study specifically the mutations in human cancer.

ClinVar [8] is a database that studies the genomics variations and its relations with the observed human health, and the history of the interpretations. ClinVar contains more than 150K variants (genetic and phenotype variants). Another database that is specialized in immunodeficiency diseases is IDbases [9]. IDbases studies the mutations in large number of genes, which can lead to abnormalities in cellular functions that could cause the weakening of the immune system response in the human body. IDbases provide information about more than 130 genes collected from more than 7000 patients.

There are also rare diseases databases such as DECIPHER [10] and myPhenoDB [11]. DECIPHER is focused on a rare disease called harbor submicroscopic deletions or duplications in chromosomal cloning. It is a disorder affecting the copy number of dosage-sensitive genes, or mis functionality in some genes expressions that cause a disease related to intellectual disability [12]. MyPhenoDB contains more than 40GBytes of data related to Mendelian diseases and other rare diseases. Mendelian disorder diseases are produced due to some modifications (mutations) applied to the genes. This is a condition that can be observed since birth such as cancer and other diseases. This kind of mutations occurs in germline cells which is the sex cells (eggs and sperm). Several databases are provided for this kind of study such as Online Mendelian Inheritance in Man (OMIM) [13] and Orphanet [14], which contain more than 20000 and 10000 entries related to genes or diseases, respectively.

Genotype-Phenotype diseases association

Therefore, genotype to phenotype diseases association has becomes one of the essential purposes of studying genetics. Understanding these relations has become more challenging, which has led to several state-of-the-art approaches proposed in the fields for designing and implementing a solid associations system that finds and detects these complex relations. One of the best ways to detect these kinds of relations is by measuring the similarities and finding common patterns in the proteins/gene's sequences.

In a study [15], 9 different classifiers (Random Forest, XGBoost, LSTM, DCNN and others) have been used for genotype-phenotype prediction. The type of phenotypes data, used in this paper, focuses on eye-color and type-2 diabetes, where the dataset was constructed from 806 people. A binary classification is targeted where the dataset is split into two classes 402 people with brown eye color and the remaining ones had blue-green eyes. The 9 classifiers have been evaluated using accuracy, precision, recall and AUC. The data sources used are the genotype data, SNP's data, and phenotypes data. The data were collected from different sources [16], AncestryDNA [17]

and ftdnaillumina [18]. These are companies used for DNA test and other tasks. Stacked ensamples of LSTM outperformed other algorithms with accuracy of 0.96 and AUC score of 0.98.

A graph-based link prediction is implemented for pheno-genotype association [19]. The authors compared five different supervised classical machine learning algorithms. The dataset was collected from Orphanet HPO annotations dataset. Several preprocessing techniques were used for further enhancement such as feature extraction from graphs node2vec algorithm, which is based on random walk with restart. After other preprocessing techniques (such as data cleaning and analysis), the proposed approach discovered the accuracies of the five machine learning algorithms (Logistic regression, XGBoost, LightGBM - another gradient boosting methods, Neural Networks and Random Forest). The performance evaluation metrics used were F1score, precision and recall. The prediction is also a binary classification task between two only classes. The results showed that Random Forest is outperformed the other algorithms with F1-score of 0.97 on the test set.

PHENOstrucut [20] was proposed for human phenotype ontology prediction terms based on machine learning algorithms. The proposed study implemented a multilabel classification task for HPO prediction using structures Support Vector machine. The dataset used for this approach is Human Phenotype Ontology (HPO) project [5] trained on three sub-ontologies (Organ, Inheritance and Onset). The features used in this approach was different data sources, protein-protein interaction network from BioGRID [21], STRING [22], GeneMANIA databases [23]. Variant human genome was fully extracted from UniProtKB database. These variants have been found in patients. The proposed system achieved AUC score of 0.74, which outperformed binary SVM and Clus-HMC-Ens [39] with scores of 0.66 and 0.65, respectively.

In Deeppheno [24], a hierarchical classifier is designed and implemented in order to predict a multi-label classification phenotype-genotype. The proposed system used 5fold cross validation performance analysis using two different datasets. The first dataset was obtained from CAFA2 challenge, while the other one from human phenotype ontology project. The performance evaluation metrics used were Fmax, precision, recall and AUROC. The proposed system was trained and tested on the two datasets and on three subontologies (Organ, Inheritance and onset). Deeppheno achieved scores of 0.766, 0.445, 0.457, 0.445, 0.47, 114.045 for AUROC, AUPR, Fmax, Precision, Recall, and Smin.

In HPO2GO [25], the authors proposed a study to show another method for genotype-phenotype prediction using implicit semantic similarity.

METHODS

Datasets used

The datasets used in the experiment are divided into two groups according to the type of the task whether genotype annotations or phenotype annotations diseases prediction.

Input amino acid sequences (UniProtKB Dataset): UniProtKB is a central bank of extensive protein information extracted from several resources. Various institutes, universities, and research labs have collaborated in order to enrich the annotations that can describe the protein sequence using diverse mandatory data. The more annotations information is added, the more cross-references and clear vision indirect relations to other proteins could be found. In addition, the quality of the annotations is an important factor that gives the intuition of the researcher the right way to extract hidden information. UniProtKB provides wide and diversified information such as the protein amino acid sequences in FASTA format, protein name and its description, taxonomic data, citation information, and more. UniProtKB consists of two datasets, UniProtKB-SwissProt and UniProtKB-TrEMBL. In order to provide high-quality annotations, the review process is added which is represented in the UniProtKB-SwissProt. On the other hand, to respond to the increased dataflow from the genomics projects and research UniProtKB-TrEMBL was created. The UniProtKB-SwissProt dataset is utilized in this experiment for protein function prediction.

The latest release (April 2021) of UniProtKB/SwissProt contains 564638 sequence entries. While the maximum length of the sequences in the dataset is up to 35213 alphabets. As shown in Figures 1 and 2 the number of sequences in each length ranges in both UniProtKB/SwissProt and UniProtKB/TrEMBLE.

Gene ontology: Gene Ontology (GO) is a comprehensive formal computational representation of protein functions. The mission of GO is to identify and classify the functions of the protein in the three domains, Molecular Function (MF), Biological process (BP), and Cellular Component (CC). GO annotation exemplify the target, in another word the predicted functions of the proteins will be the GO annotations. The protein may have one or more GO annotations from the three domains.

- Molecular Functions (MF) are those functions belonging to the activities that occur at the molecular level (simple processes), such as catalytic activity or binding activities. MF describes the action of the activities rather than the entities (molecules or complexes) that perform the actions.

Review Article

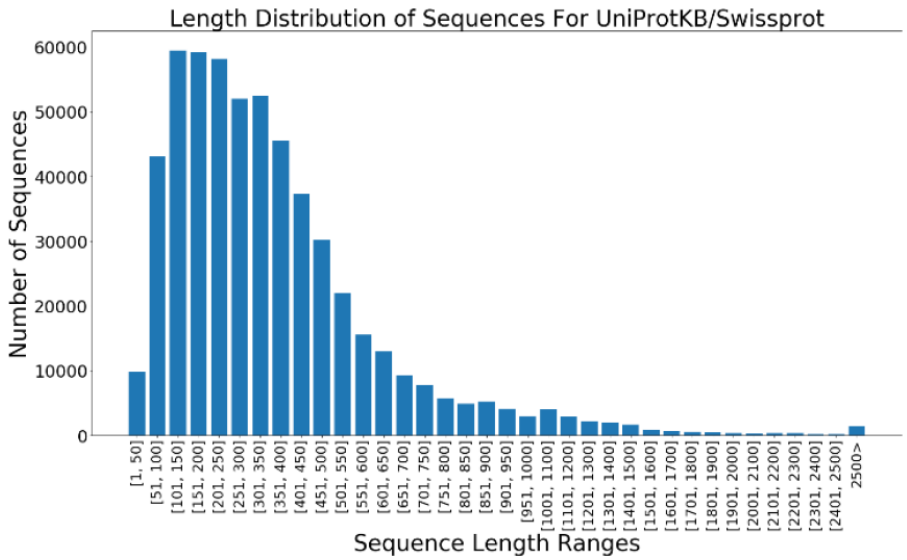

Figure 1: Number of sequences for each range length from 50 to 35213in UniProtKB/SwissProt.

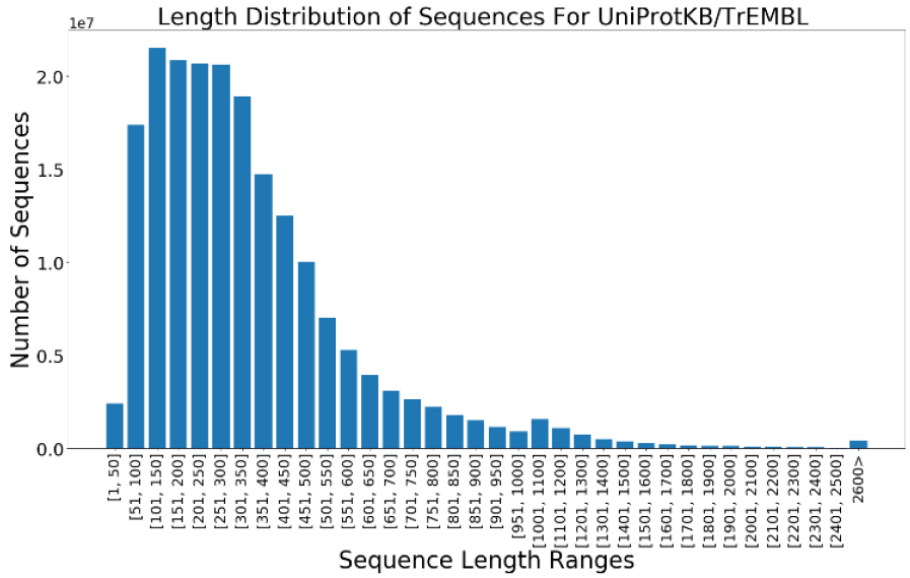

Figure 2: Number of sequences for each range length from 50 to 35213 in UniProtKB/TrEMBLE.

- However, they do not describe "where" or "when" or "in what context" the action takes place.
- Biological Process (BP) represents the largest and complex processes accomplished by multiple molecular activities. Those processes represent MF activities or chemical reactions that are involved inside and/or outside cells. An example for BP is DNA repair, signal transduction, ... At present, the GO does not try to represent the dynamics of dependencies that would be required to fully describe a pathway.
- Cellular Component (CC) are the locations relative to cellular structures in which a gene product performs a function. CC are those structures of which cells are composed. CC represents complex places that exist inside (e.g., ribosome) and/or outside (e.g., mitochondrion) cells where activities are performed. Unlike other aspects of GO, cellular component classes refer not to processes but rather to where they take place within cellular anatomy.

In this paper, GO annotations were used for the 2021 release.

The annotations are divided into three sub-ontologies 11153 for MF, 28741 for BP, and 4184 for CC. as shown in Figure 3. In addition, in Figure 4 shows the rapid growth of the annotations per year over four years.

Phenotype annotations (Human phenotype ontology): Human phenotype ontology (HPO) [5] is an annotation of relevant phenotypes-diseases. More than 4000 genes with more than 200,000 annotations are produced from different resources such as diagnostic, translation research, and several projects in computational biology in the clinical phenotype. Same as GO, HPO is represented in Direct Acyclic Graph (DAG). HPO focus on rare Mendelian diseases. Mendelian diseases are a genetic disorder disease exists when alternations in one gene or any abnormalities in the genome level. The probability of a person being affected by this kind of disease is one from thousand or one from a million. HPO term characterizes the clinical abnormality, same as GO, HPO term may be assigned into five columns as follows:

- Phenotypic Abnormality (PA) is the root of the abnormal organs such as abnormalities of the skeletal system, in blood, in the eye, in the nervous system, and more.
- Mode of Inheritance (MOI) represents the patterns or modes of the genetic disorder disease which describe how this disease takes from one generation to the next. An example of that Gonosomal inheritance, a mode observed for signs related to a gene encoded on the sex X, or Y chromosomes.
- Clinical Modifier (CM) this sub-ontology defines and characterizes the phenotypic abnormality in terms of

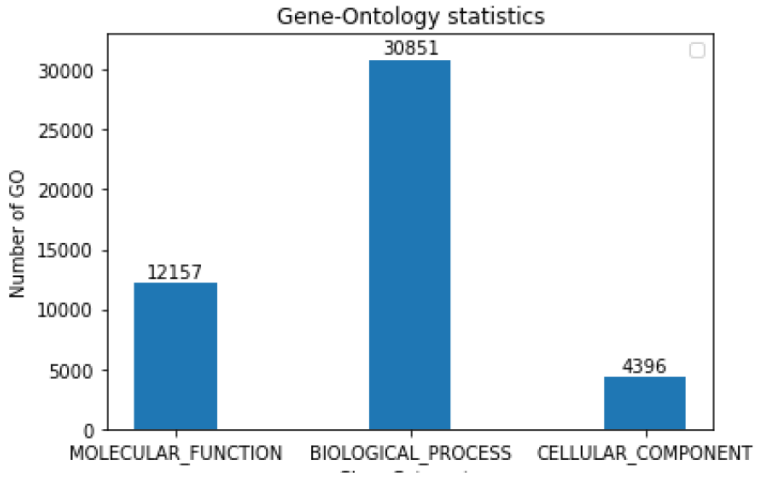

Figure 3: Gene Ontology annotations statistics.

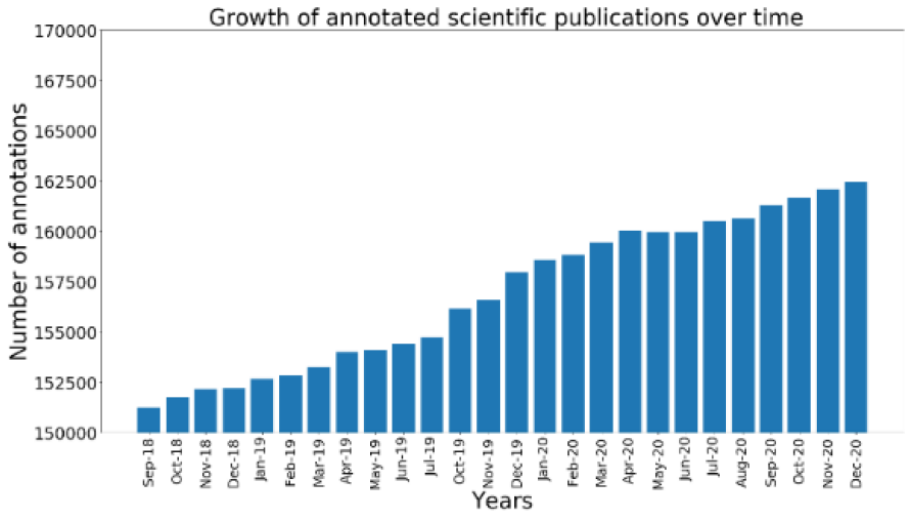

Figure 4: Rapid growth of Annotations over 4 years.

several terms such as severity which is the intensity or degree of manifestation, type of the pain if it sharp, dull, or tender, a position that describes the anatomical localization.

- Clinical course (CS) this sub-ontology describes the path or the disease from the beginning, its progress over time until the death or resolution. In this term, the disease is described by various aspects such as speed, age ranges, and more.

Genes expressions: Protein sequences are not the only features used to construct the training data. Gene's expressions are downloaded from Expression Atlas database [26]. Expression Atlas database provide more than three thousand proteins/genes expressions studies, these studies are done over 33 different species. One of these studies a project named Genotype Tissue Expressions (GTEx) [27]. This project collected more than 1600 postmortem samples covering more than 54 body sites from 175 individuals (for each gene) to study how genetic expressions varies from different tissues. The project provides 53 features, collected from experiments done over more than 50 unique tissues for each gene.

A merging process are done to extract the dataset both GTEx dataset and UniProtKB have the same gene information, using Gene-ID and gene names.

Generalized process

In this section the overall process of the system will be described and discussed in details. As shown in Figure 5 the process of the proposed system steps can be categorized as follows:

- Selected Features
- Preprocessing Techniques
- Proposed Neural networks Design and implementation
- Predicting the phenotype diseases from the predicted GO's.

Preprocessing techniques: Preprocessing step is very important to make a successful and robust machine system. As we are dealing with a biological raw data, several steps should to be done, these steps as follows:

Filtering process

- Evidences Codes.

Filtering the data from the noisy, unwanted samples prevents enhances the results remarkably. In UniProtKB dataset evidences tags are available with 90% of the overall data. evidences tags describing the exporter of that entities such as an experiment from research paper, cross reference from different database and more. The more different evidences codes attached to that entity, the more we can trust this information. Therefore, to design a trusted predictor the data information's must be trusted and verified, the constructed dataset is filtered with the records that have one of these evidences' codes ['EXP', 'IDA', 'IPI', 'IMP', 'IGI', 'IEP', 'TAS', 'IC','HTP', 'HDA', 'HMP', 'HGI', 'HEP'].

- Organism names.

As discussed, earlier UniProtKB supports more than 13K species, this experiment supports human organism.

Integer encoding

Since the proteins are represented in the dataset as a sequence of alphabets constructed from 20 amino-acids which means they are represented as a letter, these letters need to be converted into a numeric feature in order to prevent the limitations of the CNN [28] and B-LSTM [29] as well during the learning process.

One hot encoding

The neural network model may learn from the ordering of the sequences by finding a relationship which is unwanted behavior that may result a worst unexpected poor performance accuracy that leads to overfitting or biased during the learning process. In order to prevent this behavior One hot encoding is proposed to solve this issue, one hot encoding creates new feature space for each amino acid sequence, whereas all of these dimensions are orthogonal to each other. Each dimension represented as a binary vector of zeros except at the corresponding sequence index will be equal to one.

Normalization

The Expressions as well as the encoded amino acid sequences are in different ranges, normalization process is very essential in the experiment. The normalization technique facilitates and helps the loss function to converge smoothly into global or local minimums during the training process. Therefore, normalization technique called MinMax Norm is used which makes the ranges values between [0-1].

Proposed neural networks design and implementation: This section, describes the proposed algorithm details, its design, ways of implementing this approach, and how the proposed model makes prediction. As shown in the Figure 5. The proposed system constructed from several neural network models and layers.

CNN: Convolutional neural network is a regional features extraction model [30]. Several surveys in the literature over the

Review Article

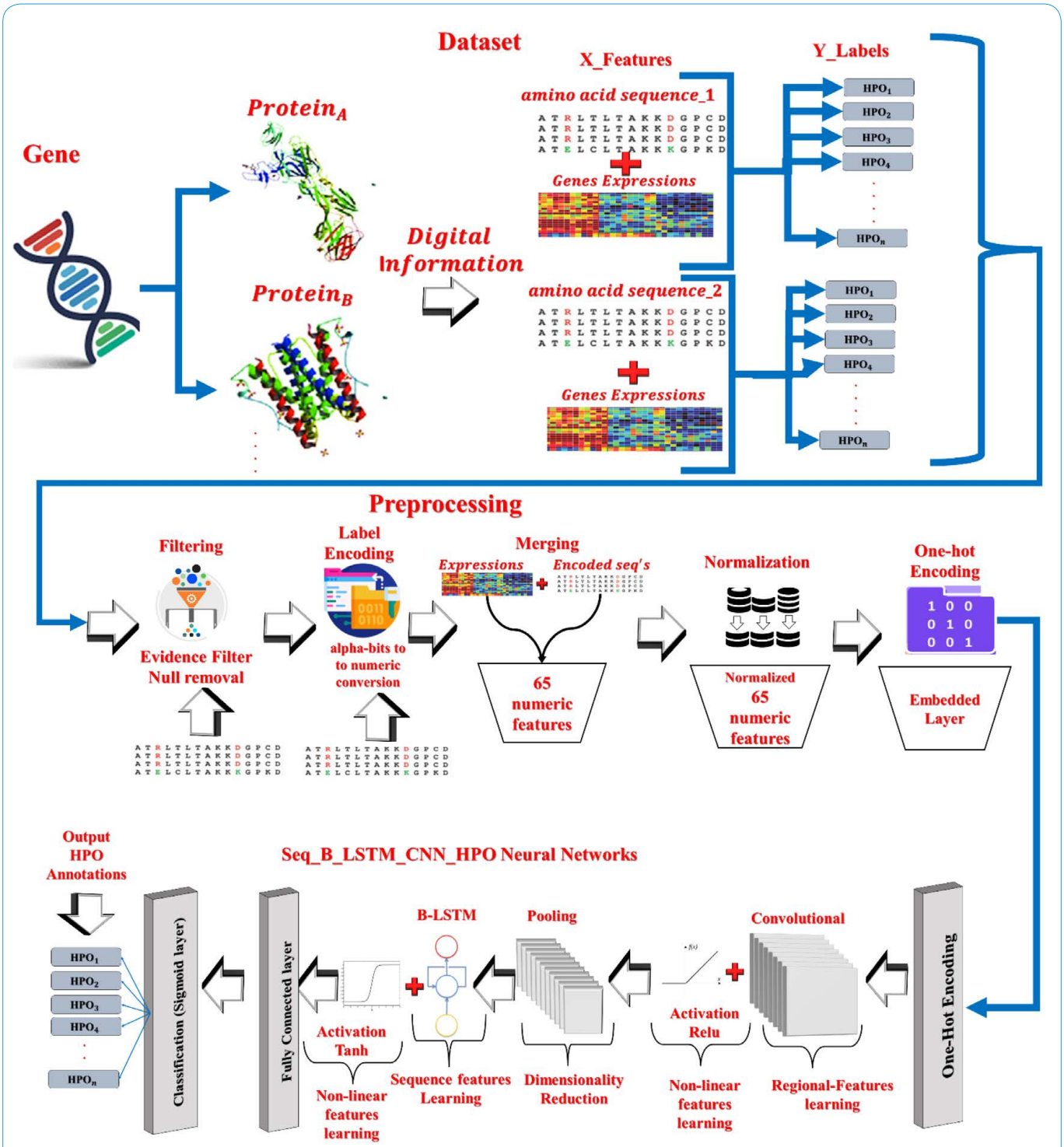

Figure 5: The proposed overall system process from the input data to the neural network model. As shown in the figure, the process starts by downloading the input data from the internet and then associating both annotations (Y_label) which are [Genetics (GO), phenotype HPO] and X features which are [sequences, expressions]. Now that the data from different resources is collected and migrated into a single file, preprocessing techniques begin in order to filter the data and also prepare it for the training process. Finally, the processed data is fed into Hyper-model Seq_B_LSTM_CNN_HPO for learning patterns. The final output is predictions for each amino acid multiple HPO annotations were predicted with the prediction score.

last 10 years proves that CNN can perform remarkable results in several fields according the datatypes. It can be used in images-based approaches (computer vision) and very suitable for that, nevertheless it can be used for text processing which performs very well performance accuracy. The reason behind that is the variety of the types of CNN designs [31], whereas there is 3D-CNN, and 2D-CNN are used for images and video processing applications, while 1D [32] is the one that is suitable for text applications. In the experiment 1D CNN is used to be a part from the hybrid model. 1D CNN can only work with numerical information (numbers integers or float), and since the protein amino acid sequences are treated as a sequence of alpha-bits, integer and one hot encoding are used in the preprocessing techniques. CNN neural networks model uses a series of convolutional layers to extract regional features from each amino-acids sequences as well as from the expression's features. The size of the kernel determine how small features can be detected in the given features vectors (amino acid sequences, expressions). For instance, a 3X3 kernel a very detailed features are detected easily, but may leads to increase the total number of parameters learned from this layer, while when the kernel size is increased may start to lose much smaller details, therefore fine tuning the kernel size is a tradeoff task that need to be choose carefully. As shown in Table 1 records each value of kernel size for each CNN layer. Regardless the kernel size tunning CNN will produce very large scale of trainable parameters [33,34] which needs to be reduced. The CNN layer is activated with RELU activation function to add more non-linearity which lead to be able to learn more complex features. Max-pooling layer comes afterwards to reduce this huge dimensionality while maintaining the features information without losing any details. In Table 1 shows the max-pooling layers size for each CNN layer.

Table 1: shows the proposed system layers types with its shape.

Layer name	Input shape
Conv1D	(None, 2593,20)
Conv1D_1	(None, 2585, 20)
Conv1D_2	(None, 2577, 20)
max_pooling1d	(None, 162, 20)
max_pooling1d_1	(None, 161, 20)
max_pooling1d_2	(None, 161, 20)
B-Lstm	(None, 8)
B-Lstm_1	(None, 16)
B-Lstm_2	(None, 24)
Flatten	(None, 8)
Flatten_1	(None, 16)
Flatten_2	(None, 24)
Concatenate	(None, 48)
dense_out	(None, 2600)

B-LSTM: Long Short-Term Memory (LSTM) [34] is an improved version of Recurrent neural network (RNN) [35] that solves the vanishing gradient issue in the original RNN. LSTM is a neural network that can learn features from long sequences. Unlike the other neural networks architecture, LSTM have three types of gates:

- Input gates
- Output gates
- Memory gates.

Using memory cell or gate LSTM can learn and forget things by charging 0 for information that need to be removed which is no-longer invalid for the sequences learning, and charging 1 for information that have high weights of relevant information. The reason behind the remembrance ability that it takes a shared input weights from what had computed from the previous time step for instance the shared information of X3 is the previous states X0+X1+X2. LSTM is activated with two different activations functions sigmoid and tanh. As shown in Figure 6a sigmoid activation function is like a guard for the three gates of LSTM (in, out, memory/forget) which outputs values from (0-1) to control the flow of the information learning throughout the gates. While to control the vanishing gradient problem which is very popular issue in RNN architectures. Tanh controls the huge increase of the weights exponentially with respect to the increasing rate of the number of layers. The reason behind that is the tanh activation function as shown in Figure 6b and 6c the output has positive and negative parts which can decrease the weights during the derivatives. Although RELU can be replaced with tanh, but tanh activation function proves its ability for faster convergence during the learning to the local or global min. In addition, the gradient computations are much cheaper during the calculations [36,37].

The bidirectional-LSTM (B-LSTM) [38] is an improved version of regular LSTM. B-LSTM contains two RNN, in the opposite direction, one learns the features from the previous states, while the other learns the biological feature in the next states. These two networks are connected to the same output layer to generate information by these ingenious integrations different complex features are learned in all the directions (forward/future and backward/past).

Integration B-LSTM and CNN: 1D CNN with kernel sizes as shown in Table 1, is a neural network architecture is a regional features extractor which can gives more details about several regions (chunks) in the amino acids. RELU activation function is activating each CNN layers to add non-linearity complex features learning and overcome the overfitting as well. These features have high parameters number that would

Review Article

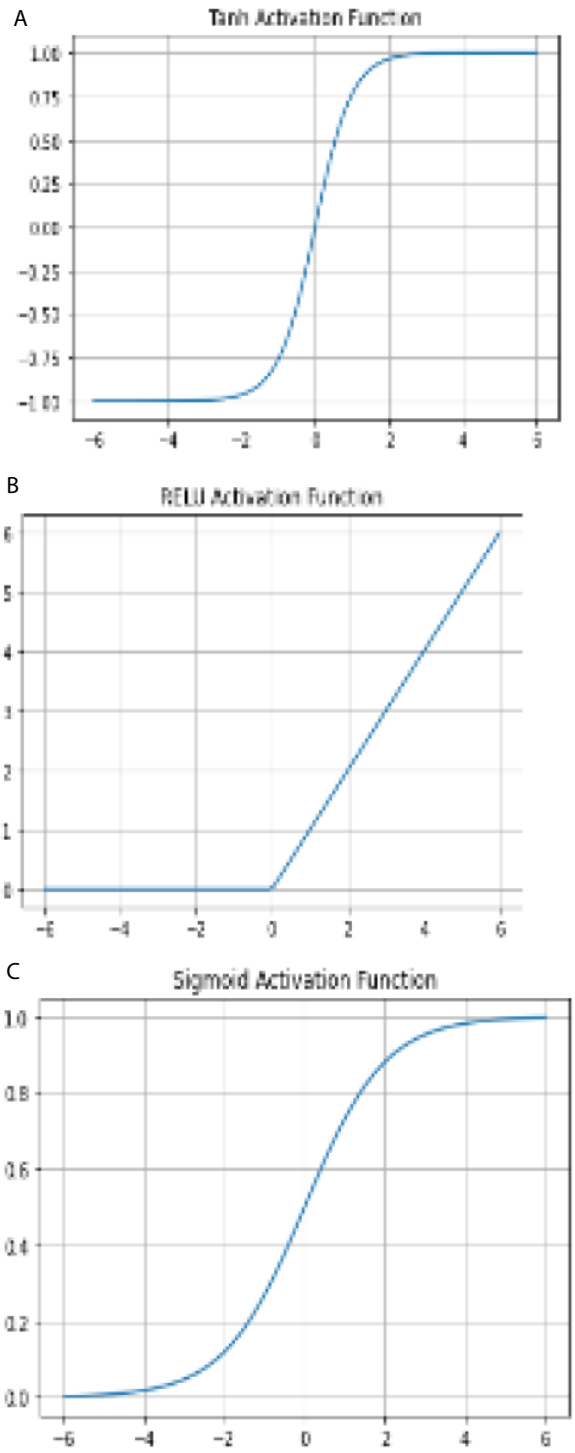

Figure 6: a, b and c shows the output form, in each acti..

Significantly increase the size of computations, therefore Max-pooling layer with sizes as shown in Table 1 are added to reduce the dimensionality without losing any information. Then these features are passed and given as input to B-LSTM to learn in different directions several sequential features. An activation function tanh and sigmoid are used with B-LSTM as discussed earlier. Finally, the output layer will be also a Sigmoid that constructed from a maximum 4600, which this number corresponds to how many HPO (class/label) predicted per amino acid. The proposed model trained on each of the five organisms individually for the three classes of HPO annotations and one model trained for all of the five organisms. There are three stages for deployment, one for training, one for validating and one for testing the result. As shown earlier in the dataset section, the distribution of the training validating and testing are 70%, 20%,10% for each organism respectively. the 70% which they are the training dataset are used to train the model in iterative process. After each step the loss function is calculated. Validation set comes into account after each epoch. When the training phase is finished the test set is used as an unseen data to evaluate the proposed model performance using different evaluation metrics that will be discussed in the results and evaluation section.

The proposed model is trained on GPU 1080 8GB RAM. On windows operating system, using python programming language with the following libraries in the Table 2 each one with its own usage.

The loss function which is the result of the difference between the prediction and the true using binary cross entropy which suitable for independent multi-class classification tasks. The weights optimization algorithm is ADAM that is improved version that can works with the nonconvex optimization applications which proves its performance over the literature. In Addition, ADAM optimization algorithm uses less amount of memory compared to the other optimizers. Batch Gradient Descent (BGD) is used to train and optimize the loss function.

EVALUATION METRICS

In this section the proposed system will be evaluated using five different evaluation metrics on UniProtKB dataset, these metrics are precision, recall, Fmax, AUPR, and ROC curve. Each of these metrics will be discussed in details including the reasons of choosing these metrics as well.

The proposed algorithm designed to deal with multi-label classification. Multi-label means that for each X features (amino acid sequence and metric expression) may contain one or more labels (classes). These labels are HPO Annotations which can

Table 2: The frameworks and libraries used in this research.

Library name	Usage
TensorFlow-keras (Python)	Neural networks implementation
Pandas (Python)	Data manipulations
Matplotlib (Python)	Data visualization
NumPy (Python)	Data manipulations

Review Article

be one of the four types (classes) of HPO Annotations. The distribution of the classes is not equally as discussed earlier which may cause a problem called imbalanced data that leads our classification to be imbalanced classification. Imbalance classification is a behavior occurs due to several reasons but not limited to:

1. Data biased.
2. Incorrect labeling of domain.
3. Rare events for that type of class.

The solution of this problem is to use different evaluation metrics which includes but not limited to: Precision, Recall, Fmax, AUPR, and ROC.

Confusion metric construction

In order to use the previously discussed evaluation metric's, the confusion metric need to be constructed first. Confusion metric is a matrix that gives insights from the data to show the rate of correct and incorrect predictions on each class as shown in Table 3.

Recall

Recall or sensitivity is a measurement metric that measures how many True Positive (TP) with respect to the True Positive (TP) and False Negative (FN) as shown in the following equation.

Recall =TPTP+FN (1)

It's a measurement metric that measure how many positive classes the model able to predict correctly. High recall means that the proposed model ends up to predict many samples that may include many false (predictions) or many true positive as well. Therefore, recall focuses on the quantity, how many classes (HPO) can be detected or classified for the specific amino acid regardless if this classification true positive or false.

Precision

Precision focuses on the quality of the proposed algorithm which measure how many TP's with respect to the TP and FP as shown in the following equation.

Precision =TPTP+FP (2)

Table 3: Confusion matrix construction details.

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Precision measures how the model is good in true positive prediction. High Precision filter out the false negative from the predicted classes for the specific sequence, while it reduces the number of predicted classes in general, in other word it ends up with low number of predicted classes but it can be guarantee that all of them are true positive.

F1-score

It's the harmonic mean between precision and recall, because it takes into account the false positive rate and false negative

$$F1-score = \frac{2*Percision*Recall}{Percision+Recall} \quad (3)$$

A high F1-score means that the proposed system predicts many classes and all of them are true positive, therefore this score needs to be maximized. F-max is calculated by selecting the threshold that perform the best (higher) F1-score. This threshold can be found using ROC curve.

$$FPR = 1 - \frac{TN}{TN+FP} \quad (4)$$

$$TPR = \frac{TP}{TP+FN} \quad (5)$$

Received Operating Characteristic (ROC)

ROC curve measures the behavior of the proposed system in terms of False Positive Rate (FPR) and TPR (Ture Positive Rate) on different thresholds. As shown in equation X_1 and X_2 calculate the FPR and TPR. From the ROC curve the best thresholds.

AUPR

Area Under Precision and Recall (AUPR) curve shows the balance between the Precision and recall for different thresholds. A higher AUPR refers to that both Precision and recall are high. High Precision and high recall mean that the proposed system returns many results, and the majority of them are correct.

RESULTS AND BIOLOGICAL EVALUATION

In the evaluation results different comparisons are conducted and performed to evaluate the proposed system Seq_B_LSTM_CNN_HPO against several methods proposed in the field that shows the advantages of the proposed model in different aspects and views, these are presented in the following subsections.

UniProtKB-SwissProt comparison

The proposed system Seq_B_LSTM_CNN_HPO is evaluated using six different evaluation metrics which they are AUROC,

Review Article

AUPR, Fmax, Precision, Recall, and Smin on the test set. Test set is constructed from the UniProtKB-SwissProt dataset with 393 samples which is the ratio of 20% from the dataset. These samples are unseen during the training process, which means that the proposed system cannot know any information about these set.

In the experiment the hyper-model neural network is trained and tested on human organism and compared with other different neural networks models which they are CNN only, B-LSTM only. The reason of these types of comparisons to shows the performance improvement in the accuracy from integrating two different artificial neural networks models to take the advantageous from each algorithm.

As shown in the Table 4 and Figure 7 the proposed system is evaluated on the three sub-ontologies organs, onset and inheritance.

For the CNN model in the organs sub ontology achieve a 0.527, 0.085, 0.164, 0.164, 0.165, and 52.245 for AUROC, AUPR, Fmax, Precision, Recall, and Smin respectively. As shown both precision and recall are very low values which means we have small number of prediction classes and have high probability that almost of these predictions are false negative which is also means we cannot guarantee that the system can predict true positive very well for the organs sub-ontology. Therefore F-max would be very low value, because it corresponding linearly proportional as shown in Equation 3. The high AUROC is worst because it corresponds to the AUPR which is very low for CNN. High AUPR refers to a high precision and high recall which means that the proposed system predicts many predicted classes at the same time we guarantee that all these classes are true positive.

on the other hand, for the onset sub-ontology achieves scores of 0.284, 0.181, 0.453, 0.293, 1.000, 1.715. which is improved prediction version than organs with high recall 1 but low in precision. This means that the system predicts a very intensive predicted classes but with low possibility having a true positive association between the gene and the corresponding phenotype ontology. As the F-max is a linearly related on both recall and precision, the amount of improvement of the recall improved F1-score from 0.164 on organs to 0.453 in onset. Therefore, we can expect that the AUPR will also improve and the AUROC is lowered, which is truly happens AUROC is lowered from 0.527 in organs to 0.284 in onset. Which inversely proportional the AUPR improved from 0.085 in organs to 0.181.

Finally in inheritance, 1D-CNN didn't predict any inherence sub-ontology. Which show us the draw-backs from the effect of learning regional features as it the ability in 1-DCNN.

For B-LSTM the same six evaluation metrics are used on the same test-set dataset, as shown in Table 4 and Figure 7 on the three sub-ontologies. For the organs sub-ontology, the results show that the proposed system achieves scores of the B-LSTM+CNN (the proposed hyper-model system) shows that improvement in the performance by:

- Predicts the three subontologies classes successfully
- The scores outperformed both B-LSTM and CNN.

In onset sub-ontology the proposed system outperformed CNN and B-LSTM with scores of 0.514, 0.078, 0.168, 0.147, 0.196 and 52.203 for AUROC, AUPR, Fmax, Precision, Recall and Smin respectively. it can be detected that the F-max is slightly better than CNN and B-LSTM which indicates that the true positive prediction accuracy is improved and at the same time the number of predicted classes are increased.

Table 4: Comparison analysis for our proposed system against CNN and LSTM.

Method	F-max	Precision	Recall	AUPR	AUROC	Smin
	Organ sub-ontology					
B_LSTM_Only	0.158	0.132	0.196	0.076	0.515	52.241
CNN_only	0.164	0.164	0..165	0.085	0.527	52.245
Seq_B_LSTM_CNN_HPO	**0.168**	**0.448**	**0.147**	**0.078**	**0.514**	**52.203**
	Inheritance sub-ontology					
B_LSTM_Only	No predictions	No predictions	No predictions	No predictions	No predictions	No predictions
CNN_only	No predictions	No predictions	No predictions	No predictions	No predictions	No predictions
Seq_B_LSTM_CNN_HPO	**0.676**	**0.549**	**0.879**	**0.423**	**0.90**	**1.290**
	Onset sub-ontology					
B_LSTM_Only	0.538	0.32	0.91	0.182	0.251	1.83
CNN_only	0.453	0.293	1	0.181	0.284	1.715
Seq_B_LSTM_CNN_HPO	**0.894**	**0.902**	**0.886**	**0.711**	**0.631**	**0.384**

Review Article

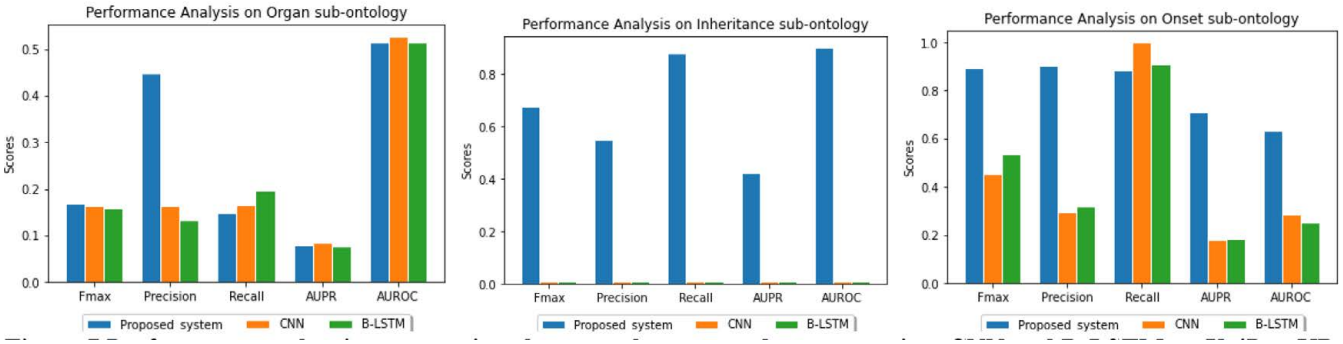

Figure 7: Performance evaluation comparison between the proposed system against CNN and B-LSTM on UniProtKB-SwissProt comparison using 'Fmax', 'Precision', 'Recall', 'AUPR', 'AUROC'.

For inheritance sub-ontology it also outperformed both CNN and B-LSTM with scores of 0.887, 0.423, 0.676, 0.549, 0.879, and 1.290 for AUROC, AUPR, Fmax, Precision, Recall and Smin respectively. the Fmax is improved remarkably than the others which corresponds to the precision and recall. As a result, we can find that the AUPR with score 0.423 is improved than 1-D CNN which did not predicts any inheritance sub-ontology at all, as well as in B-LSTM with score of 0.217.

For Onset sub-ontology, the performance is significantly improved with scores of 0.631, 0.711, 0.894, 0.902, 0.886 and 0.384 for AUROC, AUPR, Fmax, Precision, Recall and Smin respectively.

CNN+ B-LSTM achieves F-max with score of 0.711 which is improved with 107% than 1-D CNN with score of 0.453. on the other hand, compared with B-LSTM the improvement is 93%. Precision is remarkably improved than the other algorithms which indicates a very high chance that all the predicted classes are true positive. Likely that the recall score is also enhanced which means that we can have high number of predicted classes and almost of them is the true positive which is the best scenario we can get.

Smin is another score that measures the performance with different view, the lower the value the better performance accuracy of the proposed system. Smin in CNN+ B-LSTM is the lowest in the three subontologies than the two other algorithms, as shown in Figure 7 and Table 4.

As a result, this experiment proves the significant performance when more than different neural network models integrated, which give the ability to learn different kind of features and patterns for our case (regional features from the 1-D CNN and sequential rotational features from the B-LSTM. This evaluation is done using six different evaluation metric measures different point of views in the system to prevent the biased evaluation as possible.

One last note for the organs sub-ontology its noted that the scores in the three algorithms are low, this due to that the number of annotations is much larger than the genes which is the ratio is very high with respect to the other sub-ontologies. 213000/2768 annotations to genes for organs, 2668/3600, 926/1700 annotations to genes for inheritance and onsets respectively. which this proves the dilemma of low number of annotations needed. This in the future will be improved by finding more than three or four datasets.

Biological evaluations

As mentioned earlier, the proposed system is trained and tested on the dataset annotations with release 2019. The output from the system is multiple HPO annotations for single gene. Seq_CNN_B_LSTM_HPO neural networks able to generate 782 annotations for 394 genes. As shown in supplementary material excel sheet (outcomes.xlsx) the output for the prediction for all the 394 with their annotations and scores.

In order to evaluate the proposed model through another advanced point of view, we compared our predicted annotations with the dataset HPO of release 2021-Aug, which have HPO annotations more than 2019 release. Surprisingly 362 genes are successfully matched with 2021-Aug dataset (Human_phenotype_Ontology) with 607 HPO annotations out of 782. This means that out of 77% of the predicted annotations from our proposed system are true positive. The fact of performing this kind of evaluation to prove that the proposed model predictions could make a huge impact in genes to the mendelian diseases associations, and confidently can be used by the scientists in the biological labs.

As shown in the Table 5 the Top 5 Genes with their HPO annotations associated with the top GO annotations. For gene with entity number 284111 named SLC13A5 In the experiment we took the top five HPO annotations which they are matched

Review Article

Table 5: Shows the Top 5 Genes predictions from the proposed system Seq_B_LSTM_CNN_HPO.

Gene- ID	Gene name	No of annotations	Proteins names	Top Annotated mendelian diseases matched with 2021 dataset	Top GO annotations associated with the annotated HPO	Locations
284111	SLC13A5	5	Solute carrier family 13-member 5	{'HP:0002059', 'HP:0000007', 'HP:0012444', 'HP:0002500', 'HP:0001273'}	GO:0015137, GO:0005343, GO:0017153, GO:0015141, GO:0015742, GO:0098656, GO:0071285, GO:0015746	1-Plasma membrane 2-Cytosol 3-nucleoplasm
54664	TMEM106B	5	Transmembrane protein 106B	{'HP:0000006', 'HP:0030212', 'HP:0012444', 'HP:0003828', 'HP:0002500'}	GO:0051117, GO:0048813 GO:0007042 GO:1905146 GO:0007041, GO:0032418 GO:0007040 GO:1900006	1-Endosome 2-Lysosome 3-integral component of membrane
9373	PLAA	5	Phospholipase A-2-activating protein	{'HP:0012762', 'HP:0002120', 'HP:0002352', 'HP:0000007', 'HP:0003828'}	GO:0016005, GO:0043130, GO:0071222, GO:0006954, GO:0016236, GO:1900045	1-synapse 2-Cytoplasm 3-Nucleus
1436	CSF1R	5	Macrophage colony-stimulating fac- tor 1 receptor	{'HP:0002171', 'HP:0002352', 'HP:0000007', 'HP:0002500', 'HP:0000006'}	GO:0005524, GO:0019955, GO:0005011, GO:0045217, GO:0008283, GO:0071345, GO:0036006	1-nucleoplasm 2- integral component of plasma membrane 3- cell surface Source 4- CSF1-CSF1R com-plex 5- intracellular membrane-bounded orga- nelle 6- receptor complex
2253	FGF8	5	Fibroblast growth factor 8	{'HP:0001273', 'HP:0000830', 'HP:0001360', 'HP:0000006', 'HP:0002418'}	GO:0042056, GO:0005104, GO:0008083, GO:0009653, GO:0009887, GO:0035909, GO:0001974	1- Extracellular region or secreted 2- Plasma Membrane 3- cytoplasm
56479	KCNQ5	4	Potassium voltage-gated channel sub- family KQT mem- ber 5	{'HP:0002059', 'HP:0000006', 'HP:0002352', 'HP:0003828'}	GO:0005516, GO:0005251, GO:0005249, GO:0071805, GO:0034765	1-integral component of plasma membrane 2-plasma membrane 3-voltage-gated potassium channel complex 4-clathrin coat 5-integral component of membrane

Review Article

with 2021 HPO dataset annotations. This gene has a protein named (Solute carrier family 13 member 5) located in the Plasma membrane, Cytosol and nucleoplasm. which have 77 GO annotations in MF such as citrate transmembrane transporter activity (GO:0015137), and organic acid: sodium symporter activity (GO:0005343), and BP such as alpha-ketoglutarate transport (GO:0015742), and anion transmembrane transport (GO:0098656). For that protein we found five annotations matched with the 2021 dataset which they are:

HP:0002059: Cerebral atrophy→ this kind of disease decrease in size of cells or tissue, affecting the cerebrum. Cerebral atrophy has a relationship (is_a) with HP:0007369 Atrophy/ Degeneration affecting the cerebrum.

HP:0000007: Autosomal recessive inheritance → a mode of inheritance that is observed for traits related to a gene one of the autosomes.

HP:0012444: Brain atrophy → Partial or complete wasting (loss) of brain tissue that was once present.

HP:0002500: Abnormality of the cerebral white matter→ this is actually in the region of the central nervous system which interconnect with the lower brain centers.

HP:0001273: Abnormal corpus callosum morphology → Abnormality of the corpus callosum.

As shown in Figure 8 the annotations predictions by the proposed system for the first gene SLC13A5 that have protein named Solute carrier family 13 member 5.

Seq_CNN_B_LSTM_HPO is able to predict annotations in the DAG up to level 6, which indicates that the system is learning very deep features from expression and sequences. In addition, most of the predicted HPO annotations have a strong indirect relation between the HPO annotations which proves the robustness of the predictor. To the best of our knowledge this kind of evaluation is rarely performed in the literature and it also demonstrates the high accuracy of the proposed model. The Tree DAG for the other four genes is in the supplementary martials.

Table 6: Comparison analysis for our proposed system against the state of the art approaches.

Method	F-max	Precision	Recall	AUPR	AUROC	Smin
	Organ sub-ontology					
PhenoPPIOrth	0.20	0.27	0.15	---	0.52	---
Clus-HMC-Ens	0.41	0.39	0.43	---	0.65	---
PHENOstruct	0.42	0.35	0.56	---	0.73	---
RANKS	0.30	0.23	0.43	---	0.87	---
HTD-RANKS	0.37	0.30	0.49	---	0.88	---
TPR-W-RANKS	0.40	0.34	0.48	---	0.89	---
Seq_B_LSTM_CNN_HPO	**0.168**	**0.448**	**0.147**	**0.078**	**0.514**	**52.203**
	Inheritance sub-ontology					
PhenoPPIOrth	0.12	0.16	0.10	---	0.55	---
Clus-HMC-Ens	0.73	0.64	0.84	---	0.73	---
PHENOstruct	0.74	0.68	0.81	---	0.74	---
RANKS	0.56	0.43	0.81	---	0.90	---
HTD-RANKS	0.57	0.44	0.81	---	0.90	---
TPR-W-RANKS	0.57	0.45	0.80	---	0.91	---
Seq_B_LSTM_CNN_HPO	**0.676**	**0.549**	**0.879**	**0.423**	**0.90**	**1.290**
	Onset sub-ontology					
PhenoPPIOrth	0.25	0.68	0.24	---	0.53	---
Clus-HMC-Ens	0.35	0.27	0.48	---	0.58	---
PHENOstruct	0.39	0.31	0.52	---	0.64	---
RANKS	0.41	0.30	0.67	---	0.83	---
HTD-RANKS	0.42	0.30	0.69	---	**0.86**	---
TPR-W-RANKS	0.48	0.38	0.66	---	0.75	---
Seq_B_LSTM_CNN_HPO	**0.894**	**0.902**	**0.886**	**0.711**	**0.631**	**0.384**

Review Article

Seq_B-LSTM_CNN_HPO neural networks model shows its robustness and ability of predicting actual successful annotations that actually exists and verified by the biologists' scientists although the lack of the number of annotations shown in dataset section (1796, 12, 23) genes available, with (2768, 2668, 926) HPO annotations for the three subontologies. We believe that the proposed system will be able to perform a magnificent result if more data exits which will be in the future work.

State of the-art methods comparison

The results from the previous experiment are also compared and conducted against different methods proposed in the field shown in Table 6 and Figure 8. The reason from this comparison to show the accuracy of the proposed system against the other methods accuracies in the literature.

For the onset sub-ontology, the proposed system achieves the 1st place in F-max against six other algorithms proposed in the field with score of 0.894 while the 2nd place is TPR-WRanks [38] achieves 0.48 score and the other achieves (0.42,0.41,0.39, 0.35, 0.25) for PhenoPPIOrth [40], ClusHMC-Ens [39], PHENOstruct, RANKS [38], HTDRANKS [38] algorithms respectively. On the other hand, for Precision and Recall the proposed system outperformed the other algorithms with scores of 0.902, 0.886 for precision and recall respectively, while the 2nd place PhenoPPIOrth achieves 0.68 for precision and for 2nd place Recall HTD-RANKs achieves 0.69. for AUPR the proposed system performs 0.71 which is good indicator of the prediction accuracy while the other algorithms did not evaluate using this metric. In addition, for Smin the proposes system performs very good score of 0.384 which the lower the value the better performance.

For the Inheritance sub-ontology, the proposed system achieves the 3rd place in F-max with score of 0.676 while the 2nd is Clus-HMC-Ens achieve score of 0.74, and the 1st place PHENOstruct achieves score 0.74, which our result is very close to the 1st place. For precision the proposed system performs also the 3rd place with score of 0.549, while the 2nd Clus-HMC-Ens performs score of 0.64, and the 1st place PHENOstruct performs 0.68. For Recall the Seq_B_LSTM_CNN_HPO comes to the 1st place with score 0.879 while the 2nd place is Clus-HMC-Ens with score of 0.84 and the 3rd place PHENOstruct achieves 0.82. the proposed system performs 0.423 for AUPR while the others did not used this metric in the evaluation. In addition, for Smin the proposed system performs 1.290 which is also indicates a good result in the performance.

For organs sub-ontology the proposed system Seq_B_LSTM_CNN_HPO comes to the last place in Fmax score with value 0.168 while the 1st place is PHENOstruct achieve score of 0.42 and the second place Clus-HMC-Ens score 0.41. For Precision the proposed system achieves the 1st place with score of 0.448 while the second place is ClusHMCEns with score of 0.39 and the 3rd place is PHENOstruct performs 0.35. For Recall the proposed system made the last place with score of 0.147 while the 1st place PHENOstruct achieve score of 0.56. Therefore, the AUPR will be lower too with score of 0.078. AUROC and Smin are measured with scores 0.514, 52.203 respectively.

CONCLUSION

The proposed system is an accurate robust model for genes to Phenotype association. This experiment has been evaluated using five different evaluation metric that are designed for imbalanced data evaluation which they are, Fmax, Precision, Recall, Smin, AUPR, AROC. Seq_B_LSTM_CNN_HPO neural network model is a hyper-model which integrates two different neural network architectures. The proposed system has been trained and evaluated on Human Phenotype Ontology dataset on the three subontologies. In the first evaluation the proposed system is compared against CNN_only model, and B_LSTM_only model, the results shows that hyper-model integration performs significant improvement which is shown in the tables. The second evaluation is done by training and test the model on old dataset release 2019 (two years or timestamp). Surprisingly on a small set (test-set) with 394 genes, 362 genes are matched with Seq_B_LSTM_CNN_HPO predictions, with 607 out of 782 HPO annotations are matched 100%. The third evaluation is comparing Seq_B_LSTM_CNN_HPO with the state-of-the-art approach's in CAFA, which is performs improved results in the three sub-classes.

The novelty of this approach is the integration of two different functionalities from two different neural networks systems CNN and B-LSTM in order to perform and enhance the impressive prediction performance although the lack of the dataset annotations as well as the misbalancing distribution of the dataset between the three sub-classes.

Due to the lack of the Mendelian diseases annotations, we believe that the proposed system can perform better results by integrating more than dataset such as protein-protein interaction networks from STRING database and 3D structure from PDB database and more to increase the amount of sample as well as the annotations which will be in the next publications. In addition, the study focuses only on one organism (human), as we know any type of diseases exists in the human organism have indirect relations with other diseases in other organisms such as mouse, rat in future research.

Review Article

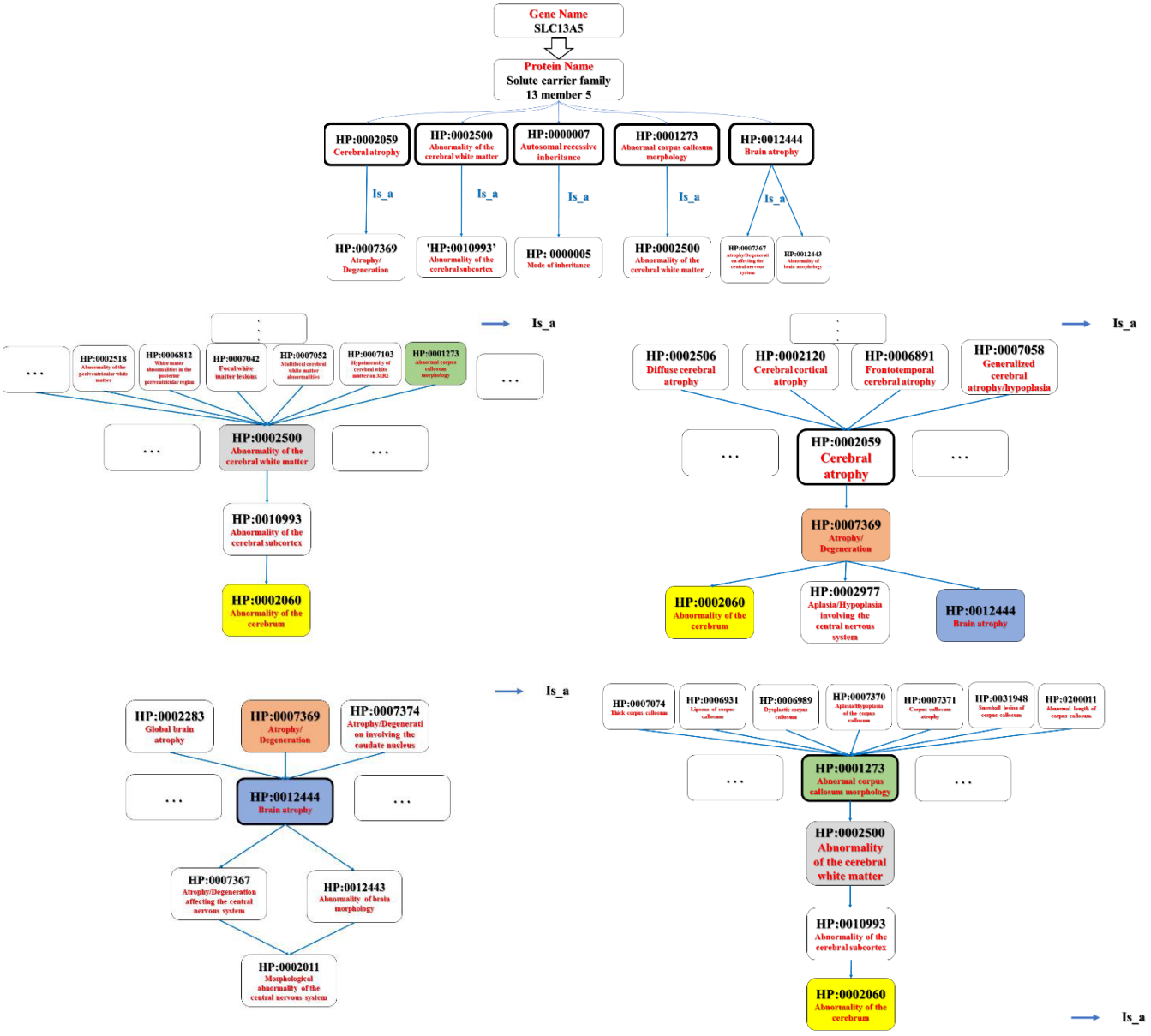

Figure 8: DAG biological annotations predictions by the proposed system for the first gene SLC13A5 that have protein named Solute carrier family 13 member 5. The common colored cells shows that the proposed system predicts the annotations at different levels (up to level 10 (depth)).

REFERENCES

1. Deans, Andrew R et al. "Finding our way through phenotypes." PLoS biology vol. 13,1 e1002033. 6 Jan. 2015, doi:10.1371/journal.pbio.1002033
2. Baye TM, Abebe T, Wilke RA. Genotype-environment interactions and their translational implications. Per Med. 2011;8(1):59-70. doi:10.2217/pme.10.75
3. The Gene Ontology Consortium, The Gene Ontology resource: enriching a GOld mine, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D325–D334, https://doi.org/10.1093/nar/gkaa1113
4. Boutet, Emmanuel, et al. "Uniprotkb/swiss-prot." Plant bioinformatics. Humana Press, 2007. 89-112.
5. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, Brudno M. The human phenotype ontology in 2017. Nucleic acids research. 2017 Jan 4;45(D1):D865-76.

6. Cheng, Phil F., Reinhard Dummer, and Mitchell P. Levesque. "Data mining The Cancer Genome Atlas in the era of precision cancer medicine." Swiss medical weekly 145 (2015): w14183.
7. Wellcome Trust Sanger Institute. "Catalogue of Somatic Mutations in Cancer (COSMIC)." (2017).
8. Landrum, Melissa J., et al. "ClinVar: public archive of interpretations of clinically relevant variants." Nucleic acids research 44.D1 (2016): D862-D868.
9. Piirilä, Hilkka, Jouni Väliaho, and Mauno Vihinen. "Immunodeficiency mutation databases (IDbases)." Human mutation 27.12 (2006): 1200-1208.
10. Firth, Helen V., et al. "DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources." The American Journal of Human Genetics 84.4 (2009): 524-533.
11. Sobreira N, Schiettecatte F, Boehm C, Valle D, Hamosh A. New tools for Mendelian disease gene identification: PhenoDB variant analysis module; and GeneMatcher, a web-based tool for linking investigators with an interest in the same gene. Hum Mutat. 2015 Apr;36(4):425-31. doi: 10.1002/humu.22769. PubMed: 25684268.
12. Tyson, C., et al. "Submicroscopic deletions and duplications in individuals with intellectual disability detected by array-CGH." American Journal of Medical Genetics Part A 139.3 (2005): 173-185.
13. Amberger, Joanna S., et al. "OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders." Nucleic acids research 43.D1 (2015): D789-D798.
14. Weinreich, Steffanie S., et al. "Orphanet: a European database for rare diseases." Nederlands tijdschrift voor geneeskunde 152.9 (2008): 518-519.
15. Muneeb, Muhammad, and Andreas Henschel. "Eye-color and Type-2 diabetes phenotype prediction from genotype data using deep learning methods." BMC bioinformatics 22.1 (2021): 1-26.
16. Annas, George J., and Sherman Elias. "23andMe and the FDA." New England Journal of Medicine 370.11 (2014): 985-988.
17. Putman, Angela L., and Kristen L. Cole. "All hail DNA: the constitutive rhetoric of AncestryDNA™ advertising." Critical Studies in Media Communication 37.3 (2020): 207-220.
18. Muneeb, Muhammad, and Andreas Henschel. "Eye-color and Type-2 diabetes phenotype prediction from genotype data using deep learning methods." BMC bioinformatics 22.1 (2021): 1-26.
19. Patel, Rushabh, and Yanhui Guo. "Graph Based Link Prediction between Human Phenotypes and Genes." arXiv preprint arXiv:2105.11989 (2021).
20. Kahanda, Indika, et al. "PHENOstruct: Prediction of human phenotype ontology terms using heterogeneous data sources." F1000Research 4 (2015).
21. Stark, Chris, et al. "BioGRID: a general repository for interaction datasets." Nucleic acids research 34.suppl_1 (2006): D535-D539.
22. Szklarczyk, Damian, et al. "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible." Nucleic acids research (2016): gkw937.
23. Warde-Farley, David, et al. "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function." Nucleic acids research 38.suppl_2 (2010): W214-W220.
24. Kulmanov, Maxat, and Robert Hoehndorf. "DeepPheno: Predicting single gene loss-of-function phenotypes using an ontology-aware hierarchical classifier." PLoS computational biology 16.11 (2020): e1008453.
25. Doğan, Tunca. "HPO2GO: prediction of human phenotype ontology term associations for proteins using cross ontology annotation co-occurrences." PeerJ 6 (2018): e5298.
26. Papatheodorou, Irene, et al. "Expression Atlas: gene and protein expression across multiple studies and organisms." Nucleic acids research 46.D1 (2018): D246-D251.
27. GTEx Consortium, et al. "The Geno-typeTissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans." Science 348.6235 (2015): 648-660.
28. Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." 2017 International Conference on Engineering and Technology (ICET). Ieee, 2017.
29. Graves, Alex, Santiago Fernández, and Jürgen Schmidhuber. "Bidirectional LSTM networks for improved phoneme classification and recognition." International conference on artificial neural networks. Springer, Berlin, Heidelberg, 2005.
30. Xiao, Tianjun, et al. "The application of two-level attention models in deep convolutional neural network for fine-grained image classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
31. Gu, Jiuxiang, et al. "Recent advances in convolutional neural networks." Pattern Recognition 77 (2018): 354377.
32. Malek, Salim, Farid Melgani, and Yakoub Bazi. "One-dimensional convolutional neural networks for spectroscopic signal regression." Journal of Chemometrics 32.5 (2018): e2977.
33. Acharya, Shailesh, Ashok Kumar Pant, and Prashnna Kumar Gyawali. "Deep learning based large scale handwritten Devanagari character recognition." 2015 9th International conference on software, knowledge, information management and applications (SKIMA). IEEE, 2015.
34. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.
35. Zaremba, Wojciech, Ilya Sutskever, and Oriol Vinyals. "Recurrent neural network regularization." arXiv preprint arXiv:1409.2329 (2014).
36. Nwankpa, Chigozie, et al. "Activation functions: Comparison of trends in practice and research for deep learning." arXiv preprint arXiv:1811.03378 (2018).

37. Graves, Alex, Navdeep Jaitly, and Abdel-rahman Mohamed. "Hybrid speech recognition with deep bidirectional LSTM." 2013 IEEE workshop on automatic speech recognition and understanding. IEEE, 2013.

38. Notaro, Marco, et al. "Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods." BMC bioinformatics 18.1 (2017): 1-18.

39. Schietgat, Leander, et al. "Predicting gene function using hierarchical multi-label decision tree ensembles." BMC bioinformatics 11.1 (2010): 1-14.

40. Wang P, Lai WF, Li MJ, et al. : Inference of gene-phenotype associations via protein-protein interaction and orthology. PLoS One. 2013;8(10):e77478. 10.1371/journal.pone.0077478.

Citation: Elhajabdou M, Ehelw AM, Eldib H, Elhabrouk M (2022) Seq_B_LSTM_CNN_HPO: Rare Mendelian Diseases to Genotypes Associations from Multiple Data Sources. Appl Cell Biol, 10(3), 2022 [77-94]