



Hybrid Deep Neural Network for Handling Data Imbalance in Precursor MicroRNA

Elakkiya R.¹, Deepak Kumar Jain², Ketan Kotecha^{3*}, Sharnil Pandya⁴, Sai Siddhartha Reddy¹, Rajalakshmi E.¹, Vijayakumar Varadarajan⁵, Aniket Mahanti⁶ and Subramaniaswamy V¹

¹ School of Computing, SASTRA Deemed University, Thanjavur, India, ² College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China, ³ Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune, India, ⁴ Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India, ⁵ School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, Australia, ⁶ The University of Auckland, Auckland, New Zealand

OPEN ACCESS

Edited by:

Thippa Reddy Gadekallu,
VIT University, India

Reviewed by:

Praveen Kumar,
VIT University, India
Prabaharan Sengodan,
Malla Reddy Institute of Engineering &
Technology, India
Devi Mani,
King Khalid University, Saudi Arabia

*Correspondence:

Ketan Kotecha
head@scaai.siu.edu.in

Specialty section:

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: 24 November 2021

Accepted: 03 December 2021

Published: 23 December 2021

Citation:

R. E, Jain DK, Kotecha K, Pandya S, Reddy SS, E. R, Varadarajan V, Mahanti A and V S (2021) Hybrid Deep Neural Network for Handling Data Imbalance in Precursor MicroRNA. *Front. Public Health* 9:821410. doi: 10.3389/fpubh.2021.821410

Over the last decade, the field of bioinformatics has been increasing rapidly. Robust bioinformatics tools are going to play a vital role in future progress. Scientists working in the field of bioinformatics conduct a large number of researches to extract knowledge from the biological data available. Several bioinformatics issues have evolved as a result of the creation of massive amounts of unbalanced data. The classification of precursor microRNA (pre miRNA) from the imbalanced RNA genome data is one such problem. The examinations proved that pre miRNAs (precursor microRNAs) could serve as oncogene or tumor suppressors in various cancer types. This paper introduces a Hybrid Deep Neural Network framework (H-DNN) for the classification of pre miRNA in imbalanced data. The proposed H-DNN framework is an integration of Deep Artificial Neural Networks (Deep ANN) and Deep Decision Tree Classifiers. The Deep ANN in the proposed H-DNN helps to extract the meaningful features and the Deep Decision Tree Classifier helps to classify the pre miRNA accurately. Experimentation of H-DNN was done with genomes of animals, plants, humans, and Arabidopsis with an imbalance ratio up to 1:5000 and virus with a ratio of 1:400. Experimental results showed an accuracy of more than 99% in all the cases and the time complexity of the proposed H-DNN is also very less when compared with the other existing approaches.

Keywords: bioinformatics, precursor microRNA, hybrid deep neural network, deep artificial neural network, deep decision tree classifier

INTRODUCTION

The problem of imbalanced data has often been recognized as a significant challenge in Machine and Deep Learning. The ML classifiers tend to get influenced by the majority classes when it comes to imbalanced data. This results in a poor classification of the minority classes. Whenever the class sizes are significantly different, the classification methods perform much better in the larger class, achieving accurate results, whereas the minority class seems to have a very poor recall. The classification of one class being considerably neglected in comparison to another is still one of the most difficult problems to solve in the creation of new classification models today. Bioinformatics is one of the major streams that provide a large number of highly imbalanced data. These enormous

datasets require a great solution that can handle such problems and are appropriate for such imbalance ratios. One of the major data imbalance problems in bioinformatics is the identification of the precursor micro RNAs (pre-miRNA) from the genome. The miRNA is a special type of non-coding RNA. They have a length of 21 nucleotides and they play a key role in the posttranscriptional regulation of gene expression. The miRNA is a minute, endogenous RNA having a hairpin-like structure. Numerous key biological processes, such as disease genesis and progression, are regulated by miRNAs. They can act as an oncogene or tumor suppressor in a variety of cancers assisting in prognosis prediction, therapeutic assessment, and disease diagnosis (1). A promising therapy for breast cancer is delivered by the latest miRNA-based drugs (2).

The genome data consists of hundreds of thousands of sequences out of which there will be only a few miRNAs. Moreover, there are several other similar RNAs which makes it even more difficult to differentiate them. Further, the minority class usually contains a small number of high-variability instances, making it more challenging for a classifier the generalization of the new dataset. So, to find them in the highly imbalanced genome data and classify them is a great challenge and hence there is a need to find an optimal solution for identifying them with a simple procedure. Although the computational methods have been necessary tools in miRNA gene finding and functional studies (3). The Machine Learning (ML) techniques have made it possible to classify them to a good extent (4). Machine Learning and Deep Learning (DL) have taken control over every field and are providing adequate solutions for every complex problem. There are few algorithms proposed for which the True Positive Rate (TPR) is very less and complexity is more. For the genome datasets, there are many negative samples and very few positive samples of miRNA sequences. Generally, this type of problem is related to Imbalanced data. To handle such Imbalanced data there are a few solutions out of which Sampling is the most famous and efficient way. There are two variants of sampling: Oversampling and Undersampling. Oversampling is the technique in which the fewer samples in the dataset will be upscale (in other words upscaling the minorities) to make the data to be balanced. In this technique, many dummies for fewer samples are generated. When the dataset is with a smaller number of samples (both positive and negative) this technique will help to build a model with an adequate number of samples. Under Sampling is the technique in which the greater number of samples in the dataset will be downscaled (in other words downscaling the majorities) to make the data to be balanced. In this technique, the Selection method or Random sample method is chosen. In general, when the dataset is very large, the model will take a huge time to execute, this technique will help us to overcome this issue. In one of the proposed models (5), they chose an Imbalance Ratio (IR) up to 1:2000 (meaning one positive sample and 2,000 negative samples), and different ratios like 1:1, 1:100, 1:500, 1:1000, 1:1500, 1:2000. To improve the model and follow the consistency, different IRs like 1:2500, 1:3000, 1:3500, 1:4000, 1:4500, 1:5000 have been added. A model was proposed where, in addition to the Animal and Plant dataset, the Human and Arabidopsis dataset with IR up to 1:5000 and

virus dataset with IR up to 1:400 was considered to enhance the performance (6).

The existing systems like Multi-layer perceptron (MLP), Support Vector Machine (SVM), Deep Self-Organizing Map (DeepSOM), Deep Elastic SOM(DeSOM), Deep Elastic Ensemble SOM (DeeSOM), Deep Belief Neural Network (DeepBN) can solve this problem to some extent only. The Deep Architectures which work without SMOTE (Synthetic Minority Oversampling Technique) (7) take more time and accuracy is also poor, the other systems stated above also produce low accuracy if the class imbalance ratio is gradually increased. Most of the existing works so far have not considered multiclass pre-miRNA classification having an imbalanced dataset, even if they have discussed their performance is not up to the mark and hence end up having low accuracy. The accuracy of the existing work also goes down as the imbalance ratio is increased as it is given in the real-time scenarios. So, to address the problems like high run time and low accuracy faced in these existing systems, we propose a Hybrid Deep Neural Network (H-DNN) architecture, which is implemented by integrating the Deep Artificial Neural Networks (ANN) with a Deep Decision Tree classifier (Deep DT). Deep ANN is capable of optimizing a wide range of coefficients. As a result, Deep ANN can tolerate a much larger variation than traditional models. With enough training, Deep ANN will be prepared for any circumstance. It is a type of machine learning technology with a large memory capacity since it is capable of remembering every call. On imbalanced data, DT usually performs effectively. They operate by understanding a hierarchy of if/else scenarios, which forces them to address both majorities as well as minority classes. The main contributions of the proposed work can be listed as follows:

- Preprocessing of multiclass genome dataset and selection of features using Select-K-Best and Recursive Feature Elimination.
- Building proposed H-DNN framework by integrating Deep Artificial Neural Network and Deep Decision Tree for classification of miRNA from a multiclass genome dataset.
- Apply the H-DNN architecture with different imbalance ratios of multiclass genome dataset consisting of Animal, Plant, Arabidopsis, Human, and Virus genomes.

The rest of the paper is organized as follows: Section Related Work discusses the related work. Section Proposed Work discusses the proposed H-DNN architecture and its components. Section Experimental Work details the experiments and presents the results of H-DNN and section Results and Discussions summarizes the proposed approach and its future enhancements.

RELATED WORK

Due to the imbalanced class distributions in the data, handling such class imbalance problems and drawing a large amount of interest toward it, has become the major issue. To deal with these class imbalance problems many ensemble methods are proposed but most of them focus on the two-class imbalance problems, leaving many issues unsolved in the multiclass imbalance

problems. To solve those multi-class problems efficiently an algorithm called AdaBoost.NC was proposed. It produced better results when compared with other popular ensemble Techniques (8). As the availability is expanding in all kinds of fields it becomes difficult to perform analysis and progress from raw data to decision-making processes. Even though new data engineering techniques are showing great success in solving real-world problems, learning from imbalanced data has emerged as a new challenge. To enable efficient learning from imbalanced data it is required to have new approaches, understandings, algorithms, principles, and tools, because of its complex characteristics (9, 10). Class imbalances make supervised learning problems difficult because some classes have more examples than others. The existing methods have completely focused on binary-class cases. A dynamic sampling method for Multilayer Perceptron (MLP) was proposed to solve the multi-class imbalance problems. To train the MLP, DyS dynamically selects informative data. This Dynamic sampling technique has performed well when compared with other techniques (11). There are several other preprocessing techniques for the data, classification algorithms, and evaluating the model involved in modeling methods (12). The important regulators of gene expression are MicroRNAs (miRNAs). The genetic loss of the tumor suppressor miRNAs or Amplification and the overexpression of one's 'oncomiRs' is related to human cancer and is enough to cause tumorigenesis

in mouse models. Besides, the depletion of global miRNA by genetics and in the components of miRNA biogenesis machinery epigenetic alterations is oncogenic. All these together with novel miRNA regulatory pathways and factors show how important miRNA dysregulation is in cancer (13). Even after identifying hundreds of miRNAs in different species, a lot of others remain unknown. So, in understanding miRNA-mediated posttranscriptional regulation mechanisms the discovery of new miRNA genes is an important step. The biological approaches in identifying miRNA genes might be limited. So, to overcome those limitations in the biological approaches sophisticated computational approaches are followed to identify the possible miRNA (14). Some of the work deals with the miRNAs in terms of diagnosing dreadful diseases in the medical field. Several web-based bioinformatics tools were proposed for the analysis of miRNA such as expression, detection, etc. with an easy-to-use model that doesn't require any previous knowledge of utilization (15). A Support Vector Machine (SVM) based approach was proposed to classify the real and pseudo miRNA. It achieved an accuracy of up to 90% (16). The model was exclusively for the human dataset but the remaining datasets could also be used. In this work, then genomic information was not required. Later, software that classified the RNA of mammals, nematodes, and also urochordates was proposed. A new methodology was proposed based on Support Vector Machine (SVM) and named

TABLE 1 | Comparison of various state-of-the-art methodologies.

Author	Methodology used	Metrics	Dataset
Park (25)	Recurrent Neural Network (RNN).	Various Metrics were used. Experimental results yielded an F1-score of 0.93 for the Human dataset, 0.94 for the Cross-Species dataset, and 0.93 for the New pre-miRNA dataset.	Human, Cross-Species and New miRNA
Thomas et al. (26)	Deep Learning with a restricted Boltzmann machine (RBM) for classification and Modified sampling technique is applied to address the class imbalance problem.	Various Metrics were used. The model gave an accuracy of 0.968	Human Dataset
Stegmayer et al. (27)	Deep Self-organizing maps (DeepSOM) with Clustering	Various evaluation metrics were used. The model provided almost 95% of Gm and 97% of accuracy for the most imbalanced data.	H. sapien and A. thaliana, Animal, and Plant
Bugnon et al. (5)	Hybrid deep learning method with CNN and LSTM.	Various Metrics were used for different imbalance ratios. Experimental results yielded an F1-score of more than 40% for Animal and more than 33% for the Plant dataset.	Animal and Plant
Tang et al. (28)	CNN with different types of feature learning and encoding methods.	The results gave an accuracy of 99.25% and an F1-score of 99.25% for Rfam-300	Rfam-300, Rfam-120, Rfam-60, Rfam-30
Shi et al. (29)	Localized multiple kernel learning model with a nonlinear synthetic kernel (LMKL-D)	The model gave an accuracy of 98.3%, sensitivity of 93.06 Specificity of 99.27 and mean of 96.11.	pre-microRNAs and pseudo pre-microRNA
Yones et al. (30)	Convolutional deep residual neural network	The model provided a precision of 0.93 for A. thaliana full genome, 0.67 for A. gambiae, and 0.50 for H. Sapiens	A. thaliana, C. elegans, A. gambiae, and H. Sapiens
Proposed HDNN Methodology	Artificial Neural Network embedded with Decision Tree	Various metrics were used for evaluation for different imbalance ratios. The results yielded F1 Score of more than 0.50 (0.50–0.93) for Animal, more than 0.58 (0.58–0.95) for Plant, more than 0.60 (0.60–0.94) for Human, more than 0.50 (0.50–0.95) for Arabidopsis and more than 0.50 (0.50–0.94) for Virus.	Animal, Plant, Human, Arabidopsis, and Virus.

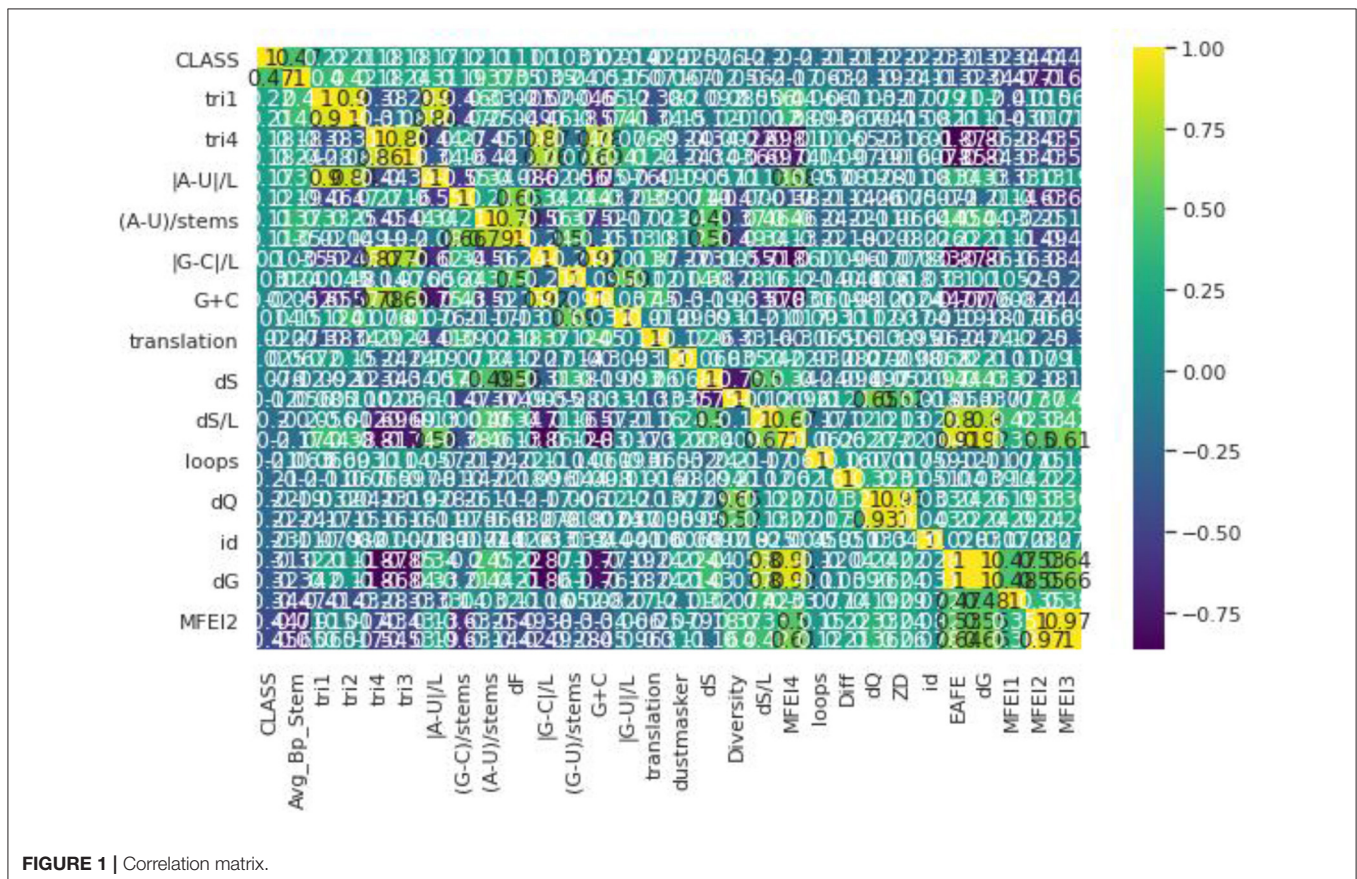
as Microprocessor SVM, which predicted with 50% accuracy (17). To improve the accuracy of this model, the model was trained with another classifier and ensemble both to get better results than acquire 90% in total. Another work dealt with the miRNA sequences of the mouse and human using the SVM approach (18). The model worked with around 377 mouse data and 475 human data and with the candidates, the dataset had 3,441 humans and 3,476 mouse samples in total. The classification was done according to the structure of the miRNAs, the model names, and the mirCoS. One of the works also dealt with the use of ab initio Prediction model clustered microRNA identification (19).

An application for the classification of pre-miRNAs was proposed which dealt with the classification of the mammalian miRNAs such as Human, Rat, and Mouse (20). This approach analyzed the genome properly using the model named ab initio which gives an accuracy of up to 60% when they undergo cross-validation. To analyze the structure of miRNA, a tool named MiRFinder was proposed which was built with the SVM model (21). The structure of the miRNA is a hairpin-like structure, the prediction mainly focused on the structure and later the features were used to predict the miRNA. This tool was mainly used for the genome-wide search of the miRNAs. The dataset which has 697 real miRNA and 5,428 pseudos from different sources was collected and experimented with. The tool gained sensitivity up to 93%. Another model named miRenSVM was

proposed that gave good results for the data in miRBase 13.0 animal and other species (22). For animal data in miRBase 13.0, it gave 100% sensitivity. A novel RNA effective classification methodology namely microPred was proposed which classified the types of miRNA using the Machine Learning approach (23). It used a classifier namely microPred, which also could handle the data imbalance issue as well as new feature extraction relevant biologically. It achieved a specificity of 97.28% and a sensitivity of 90.02%. The model gave the prediction rate of 92.71% for humans and 94.24% for the virus. Another new model was developed that used a method named YamiPred, which is an ensemble of Genetic Algorithms (GA) and Support Vector Machine (SVM) for the prediction of miRNAs (24). The model mainly used a human dataset to predict the human miRNAs. The model appeared to be robust and efficient with different performance characteristics. **Table 1** illustrates the comparison of various state-of-the-art methodologies used for the classification of miRNA.

PROPOSED WORK

In this section, we discuss the proposed Hybrid Deep Neural Network approach for the classification of miRNA from genome data. Different genome datasets have been considered in our proposed work such as Animal, Plant, Human, Arabidopsis, and Virus datasets. First, the dataset is preprocessed and all



the important features are selected and extracted and the non-useful features are omitted. For feature selection and extraction feature selection Algorithms like Select-K-Best and Recursive Feature Elimination are used. After preprocessing the dataset, the proposed H-DNN algorithm is implemented. The proposed H-DNN framework is the integration of Deep ANN with the Deep DT. Then the implemented model has experimented with the different genome datasets (31).

Preprocessing

Pre-processing refers to Feature Selection or Feature Extraction. Different Feature Selection algorithms have been applied with the dataset (32). To have a clear idea of feature selection, the dataset has to be analyzed first. Different genome datasets have been considered in our proposed work such as Animal, Plant, Human, Arabidopsis, and Virus datasets. Even though these datasets have a different number of samples in each, they have the same number of features (33). The dataset used here has a total of 30 features in which one is a serial number, hence we omit that attribute. The remaining 29 features are taken into consideration and passed to the Feature Selection Algorithms (34). **Figure 1** illustrates the correlation matrix plotted for all the features and a heat map is produced out of that as shown. The correlation matrix helps us to characterize and sum up the data for our understanding. Correlation, here, helps us to compute the association between two variables by testing the effectiveness of their linear relationship. After estimating the correlation matrix, the Feature selection Algorithms like Select-K-Best and Recursive Feature Elimination are used. In the Select-K-Best algorithm, the features are assigned with particular scores. The Selection is generally made based on the scores which are greater than the mean of all the scores in this Feature Selection (35). In the Recursive Feature Elimination, the rank is assigned to each feature and the ranks may be the same for more than one feature. The feature selection is made based on the ranks that are assigned. Generally, the selection is done based on the first few ranks like 10, 12, etc., based on several features (36). The Recursive Feature Elimination with Cross-Validation is used. This gives scores between 0 and 1. The 0, being the least important and one is the most important feature to build any model. This algorithm is applied to the dataset and the results state that all the features are nearer to the one (37). Hence, all the features are important. This proves that there is no need for removal or elimination of any feature from the data so all 29 features are used to build the proposed model (38).

Proposed H-DNN Architecture

In this section, we discuss our proposed work. We have developed a Hybrid Deep Neural Network (H-DNN) wherein we have integrated Deep Artificial Neural Network and Deep Decision Tree (DT) classification technique for classifying the miRNAs in the genome data. The integration of feature extraction using machine learning techniques and feature analysis and classification using Deep Neural Network has provided better results. The experimental results have been carried out using Animal, Plant, Human, Arabidopsis and, Virus genome datasets. The results are then analyzed and compared

for the model using Deep ANN and Deep DT. Deep ANN here helps in analyzing the features by optimizing a wide range of coefficients. As a result, Deep ANN can tolerate a much larger variation than traditional models. Deep ANN gets well prepared for any circumstance with sufficient training. It is a type of machine learning technology with a large memory capacity since it is capable of remembering every call. DT usually performs effectively in case of imbalanced data by understanding a hierarchy of making decisions for different scenarios, which forces them to address both majorities as well as minority classes.

The structure of ANN consists of a neuron, a propagation function, and a bias. Consider a neuron k having an input $I_k(t)$ received from the previously connected neurons. The neuron will have an activation function $a_k(t)$, an optional threshold θ_k , an activation function f for computing new activation for time $t + 1$ as given in Equation 1, and a function f_{out} for computing the activation output which is formulated as given in Equation 2.

$$a_k(t+1) = f(a_k(t), p_k(t), \theta_k) \quad (1)$$

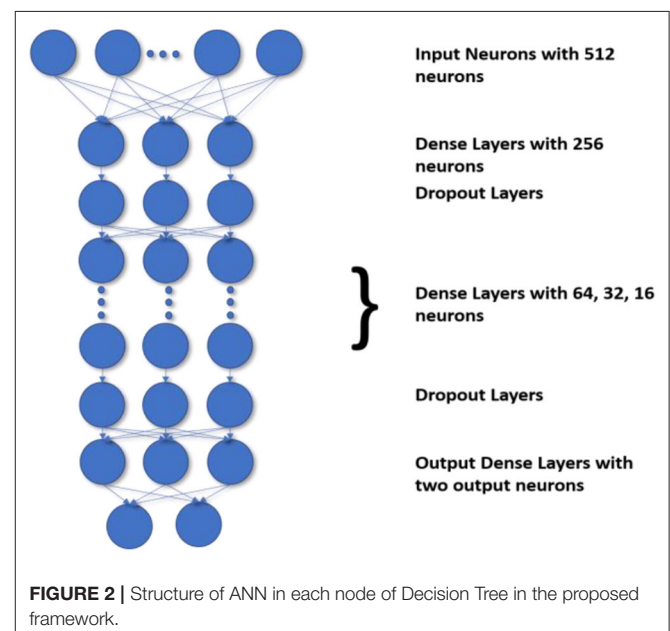
$$O_k = f_{out}(a_k(t)) \quad (2)$$

The input, $I_m(t)$ to the neuron m from the output of k , is computed using the propagation function defined as given by Equation 3. After adding the bias, w_{0m} , we formulate the input as given in Equation 4.

$$I_m(t) = \sum_k O_k(t) w_{km} \quad (3)$$

$$I_m(t) = \sum_k O_k(t) w_{km} + w_{0m} \quad (4)$$

Artificial Neural Networks (ANNs) have several different coefficients, which the ANNs can optimize. So, when compared with the traditional models, ANNs can handle much more



variability (39). This makes ANN a stronger model when it comes to memorization (40). A decision tree is a supervised learning Algorithm. Both regression and classification problems can be solved using Decision Trees. In the tree representation, each leaf node represents a class label and the internal node of the tree represents the attributes. Decision trees include three main units that are node, branch, and leaf. The node is the decision. The branch is the potential decision, and the leaf is the potential outcome of each decision. The DT illustrates the long-term consequences of certain decisions. The trade-offs and probabilities can also be expressed using decision trees. In a decision tree, after every level, we have a split taking place which in turn results in the next level. Therefore, the major challenge in Decision trees is to identify the attribute for the root node in each level. This is called attribute selection. There are various popular attribute selection measures in decision trees we have used Gini Index for attribute selection in our framework. Gini index can be formulated as shown in Equation 5.

$$\text{Gini Index} = 1 - \sum_j p_j^2 \quad (5)$$

here p_j denoted probability for class j . Gini index is used here as a cost function for the split evaluation in feature selection. When compared with other algorithms, the decision tree needs less effort for data preparation during preprocessing. It does not require scaling of data or normalization of the data. In addition to this transparency and ease of use makes the decision tree a better model to handle large amounts of data.

The most challenging task is to handle the imbalanced data. The current state-of-the-art methods do not define the decision process take more time and accuracy is also poor. Generally, the state-of-the-art frameworks produce low accuracy if the class imbalance ratio is gradually increased. Before DNN, decision trees were the standard method used to improve accuracy and performance. Though as compared to DNN the accuracy gained by using a decision tree for classification are less the DT preserves the interpretability. So to preserve the DT's interpretability and match the DNN's performance, we use the H-DNN architecture wherein we integrate the Decision Tree and the ANN. The H-DNN model consists of a DT wherein each node of the DT will have an ANN model. The DT is used to make the classification of miRNA in an H-DNN, maintaining high-level interpretability. Every node in a decision tree, on the other hand, is a neural network that makes low-level decisions. First, a hierarchy for DT is constructed and then the sample is passed through the ANN, which is a fully connected neural network, in each decision node. Then the model computes the analysis by executing the final fully connected layer as a series of integrated decision rules. The final prediction is a culmination of these decisions. The ANN structure proposed in this paper deals with a dense layer as input layer with 512 nodes and ReLu activation function followed by another dense layer with 256 nodes and a dropout layer, then three more dense layers with 64, 32, 16 nodes respectively with ReLu activation function and a dropout layer and finally one dense layer with two nodes as output layer. This ANN structure, as illustrated in **Figure 2**, is embedded into each node of the

Algorithm 1 Proposed H-DNN.

Input: Genome data, D , with attributes, a
Output: Decision Tree with classes

1. $tree = \{ \}$
2. $minVal = 0$
3. **for each** $a \in D$ **do** :
 //Considering Input Layer of ANN
 4. **for each node** $n \in a$ **and input** x_n **do** :
 5. $out_n = x_n$
 //Considering Hidden Layer of ANN
 6. **for each hidden node** h **do** :
 7. $out_h = \sum_n \varnothing(w_{hn} \cdot out_n)$
 8. $g = \text{GiniIndex}(out_h, d)$
 9. **for each output node** k **do** :
 10. $out_k = \sum_h \varnothing(w_{kh} \cdot out_h)$
 11. *Train ANN using Backpropagation*(out_h)
 12. **if** $g < minVal$ **then**
 13. $minVal = g$
 14. $tree' = \{a\}$
 15. *partition*($tree, tree'$)
 16. **repeat** till all the partitions are processed
 17. **return** tree

decision tree for making the decision rules. In the fully connected layer of each node of the decision tree, the inner product of the matrix-vector is calculated and an inner product that constitutes the maximum value is chosen for the class decision rule as shown in Equation 6.

$$\begin{aligned} [w_1 \ w_2 \ \dots \ w_n] [x \ \dots \ x] &= [(x, w_1) \ (x, w_2) \ \dots \ (x, w_n)] \\ &= [y_1 \ y_2 \ \dots \ y_n] \rightarrow \text{argmax}(y) \end{aligned} \quad (6)$$

Algorithm 1 illustrates the H-DNN algorithm that uses the Deep Decision Tree having Artificial Neural Network embedded in each root node. The ANN framework in the proposed model has a single input layer, a few hidden layers, and a single output layer. In the final layer of the hidden layer, the decision rule for splitting the tree has been incorporated using Gini Index. Line 1 and 2 in **Algorithm 1** initialize the tree and the minimum criteria to be considered for making the decision rule. The Deep ANN is embedded into each node of the Deep DT as illustrated from lines 3 to 10. For a given input node and input, line 4 and 5 illustrates the implementation of the first layer, i.e., the input layer of the Deep ANN. Using the output of the input layer the hidden layer is calculated using lines 6 and 7. In line 8 the Gini Index is calculated to determine the decision rule for splitting the node or tree further. Using the output of the hidden layers the output layer is constructed in lines 9 and 10. In line 11 the Backpropagation algorithm is called to adjust the weight according to the errors. In lines 12 to 15, the decision rule has been implemented. Once the decision rule is applied the nodes are split accordingly as shown in line 16. This process is then implemented for each node until all the nodes are processed and no further nodes can be split.

Algorithm 2 illustrates the implementation of the backpropagation algorithm used for generalizing the Neural Network (NN) by fine-tuning the weights based on the previous iteration's error rate. After applying the forward propagation, the backpropagation algorithm is called to fine-tune the NN. From lines 1 to 6 in **Algorithm 2** all the required parameters are initialized. From lines 7 to 13 the backpropagation process is applied to the neural network.

The proposed architecture can be divided into four modules as illustrated in **Figure 3**. Module one is the Imbalanced genome dataset which has the collection of Plant, Animal, Human, Arabidopsis, and Virus genome data. This data is preprocessed

in module two wherein the feature variables are correlated to find the relationship between the features. In this module, the required features are extracted using the Select-K-Best Method and Recursive Feature Elimination Method, and the rest of the unwanted attributes are discarded. After the feature selection and extraction, the preprocessed data is passed to the third module. The third module is the proposed H-DNN architecture. The H-DNN architecture is the integration of the Decision Tree and the ANN framework. The ANN framework used here helps to avoid overfitting and thereby increasing the performance. The Decision tree helps to make a decision and examine the benefits and cons of each final option. Every node in a decision tree, on the other hand, is a neural network that makes low-level decisions. The inputs of each node of the Decision Tree are passed to ANN for making the decision and analyzing the probability values of the class. The final evaluation of the model is done in the fourth Module, the Model Evaluation module, using Accuracy, Specificity, Precision, and F1-score as the evaluation metrics.

Algorithm 2 Modified Backpropagation(a).

1. $d \leftarrow$ training dataset $m \times n$ dimension
2. $y \leftarrow$ class labels
3. $w \leftarrow$ weights
4. $l \leftarrow$ number of layers $1, \dots, z$
5. $error_{ij}^l \leftarrow$ error for each i, j, l
6. $b_{ij}^l \leftarrow 0$ for each i, j, l
7. **for** $i = 1$ to m **do** :
8. $d^l \leftarrow a(z) - y(i)$
9. $b_{ij}^l \leftarrow b_{ij}^l + a_j^l \cdot b_i^{l+1}$
10. **if** $j \neq 0$ **then**
11. $error_{ij}^l \leftarrow \frac{1}{m} b_{ij}^l + \lambda w_{ij}^l$
12. **else**
13. $error_{ij}^l \leftarrow \frac{1}{m} b_{ij}^l$

EXPERIMENTAL RESULTS

Experimental Work

In this section, the dataset and the experimental results will be explained in detail by each dataset (41). In the Animal dataset, there are 7,053 positive and 218,154 negative samples and in total there are 225,207 animal sequences. One lakh seventeen thousands one hundred and one plant sequences are taken in which 2,154 positive and 114,929 negative samples are present. One thousand four hundred six positives and 81,228 negatives

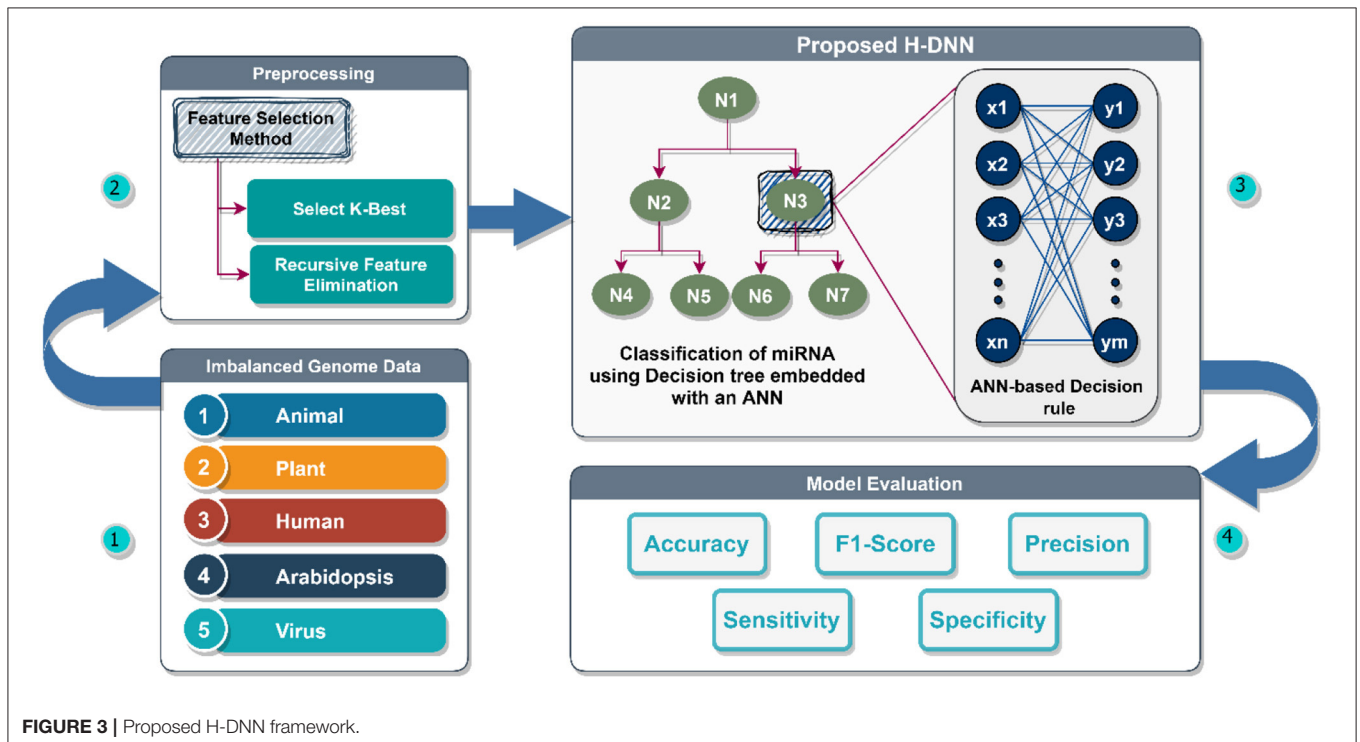


FIGURE 3 | Proposed H-DNN framework.

TABLE 2 | The no. of positive and negative samples for different Imbalance Ratios.

IR	Animals		Plants		Human		Arabidopsis	
	No. of positives	No. of negatives	No. of positives	No. of negatives	No. of positives	No. of negatives	No. of positives	No. of negatives
1:1	7,053	7,053	2,172	2,172	1,406	1,406	231	231
1:100	2,182	218,154	1,149	114,929	812	81,228	231	23,100
1:500	436	218,154	230	114,929	162	81,228	57	28,359
1:1000	218	218,154	115	114,929	81	81,228	28	28,359
1:1500	145	218,154	77	114,929	54	81,228	19	28,359
1:2000	109	218,154	57	114,929	41	81,228	14	28,359
1:2500	87	218,154	46	114,929	32	81,228	11	28,359
1:3000	73	218,154	38	114,929	27	81,228	9	28,359
1:3500	62	218,154	33	114,929	23	81,228	8	28,359
1:4000	55	218,154	29	114,929	20	81,228	7	28,359
1:4500	48	218,154	26	114,929	18	81,228	6	28,359
1:5000	44	218,154	23	114,929	16	81,228	6	28,359

TABLE 3 | The number of positive and negative samples for different IRs for virus.

IR	Virus	
	No. of positives	No. of negatives
1:1	237	237
1:50	17	839
1:100	8	839
1:200	4	839
1:300	3	839
1:400	2	839

and in total 82,634 human samples are given in this dataset. In Arabidopsis 231 positive samples and 28,359 negative samples in total, and there are 28,590 sequences. The virus samples are very few so it has only 1,076 in a total of which 237 are positive and 839 are negative samples (26). This is treated as a separate one and the maximum IR is only 1:400 (27). For the Animal, Plant, Human, and Arabidopsis datasets, the number of positive and negative samples for different IRs are tabulated in **Table 2**. For the virus, the number of positive and negative samples for different IRs is listed in **Table 3**. The initial target is to find the specificity and other metrics till IR up to 1:2000. Later, to improve the model it was extended to 1:5000. For the classification, here two Algorithms are used. They are ANN and Decision Tree Classifiers (10, 25, 28, 39, 40, 42–44).

RESULTS AND DISCUSSIONS

As discussed in section Proposed Work the ANN has been integrated with the DT in our proposed work and the results are represented graphically. Graphs are plotted for clear visualization of the results. As explained in the previous sections, the model is improvised by using different datasets like Animals,

Plants, Humans, Arabidopsis, and Viruses genomes. The work is extended in terms of the Imbalance Ratios also it is extended up to 1:5000. The model gives some predictions even for the 1:5000 ratio also, this can be visualized individually. **Figure 4** illustrates the graphical representation of experimental results done using the proposed H-DNN framework on Animal, Plant, Human, Arabidopsis and, Virus genome datasets. The x-axis in the graph indicates the IR and the y-axis in the graph plotted illustrates the metric values. The result analysis of the proposed H-DNN Algorithm was done based on the evaluation metrics: Accuracy (Acc), Specificity (SP), Sensitivity (SE), F1-score (F1), and Precision (Prec).

First, the ANN and Decision Tree Classifier results were analyzed individually. Then their results were compared with the proposed H-DNN framework. The Experimental analysis was done by implementing the state-of-the-art ANN and DT separately and then by implementing the proposed H-DNN approach using the above-said dataset. The dataset was trained batch-wise. **Table 4** illustrates the result comparison of the proposed H-DNN Algorithm with the ANN and the DT algorithms for the Animal dataset. **Table 5** illustrates the result comparison of the proposed H-DNN Algorithm with the ANN and the DT algorithms for the Plant dataset.

Table 6 illustrates the Experimental results for the Human dataset. **Table 7** illustrates the Experimental results for the Arabidopsis datasets. **Table 8** illustrates the Experimental results for the Virus datasets.

By analyzing the results, we observed that the ANN gives better results and gives a good True Positive Ratio (TPR) with 2 epochs and up to IR 1:2000. If the model is trained further for more epochs, then we are getting an over-fitted model. The time required is also in milliseconds which is very optimal whereas the existing system takes hours to train one epoch. Furthermore, when the model needs to be trained with IR up to 1:5000 Decision Tree Classifier gives better results with good TPR. But when we integrate both the DT and

TABLE 4 | Experimental results using animal dataset.

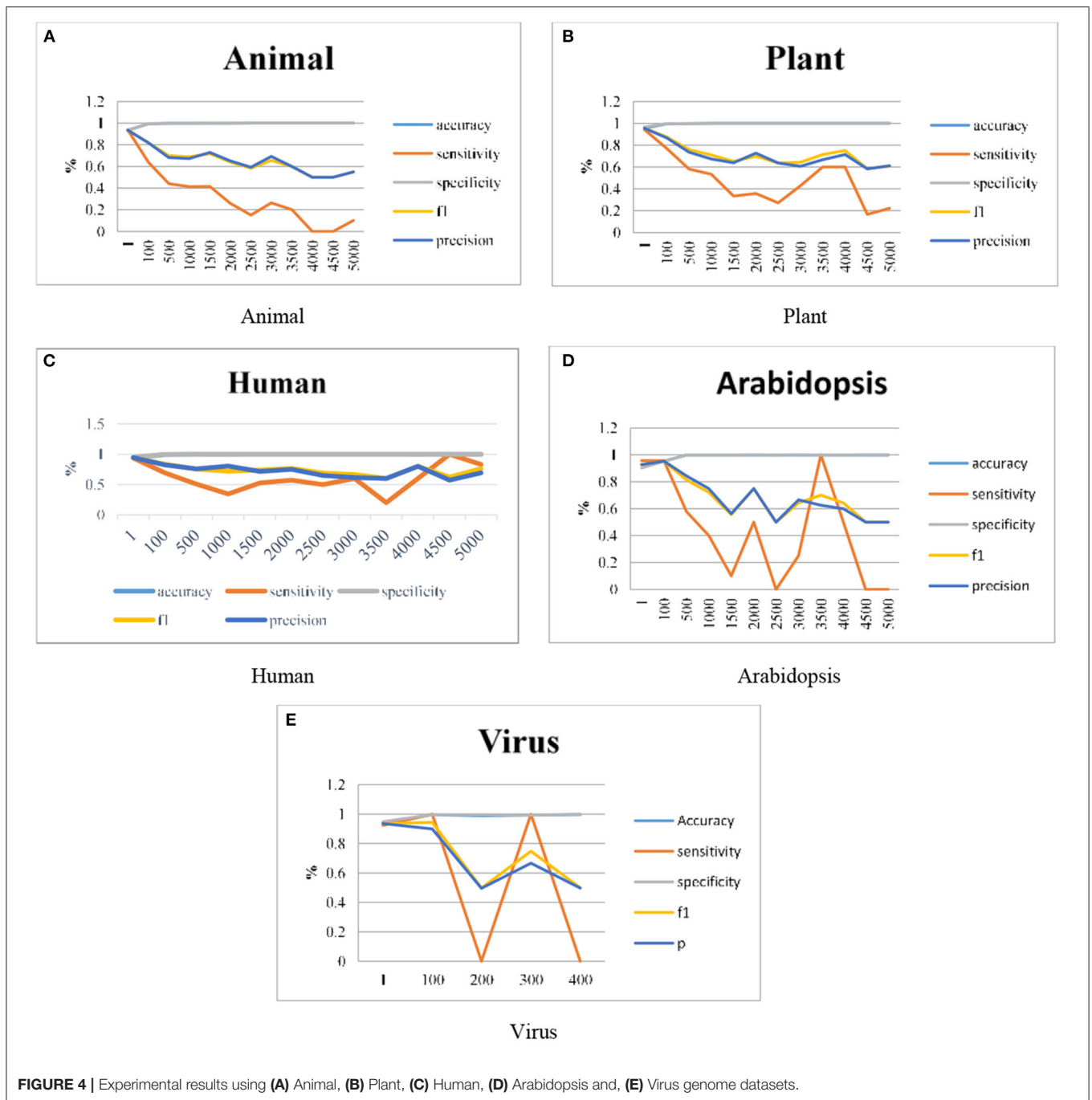
IR	ANN Algorithm					DT Algorithm					Proposed H-DNN Algorithm				
	Acc	SP	SE	F1	Prec	Acc	SE	SP	F1	Prec	Acc	SE	SP	F1	Prec
1	0.51	0.03	1.00	0.37	0.73	0.57	0.49	0.96	0.65	0.83	0.93	0.94	0.93	0.93	0.93
100	0.99	0.99	0.66	0.74	0.69	0.99	0.81	0.83	0.78	0.76	0.99	0.64	1.00	0.82	0.82
500	1.00	1.00	0.53	0.50	0.50	1.00	0.72	0.77	0.60	0.59	1.00	0.44	1.00	0.70	0.68
1,000	1.00	1.00	0.67	0.50	0.50	1.00	0.71	0.83	0.59	0.59	1.00	0.41	1.00	0.69	0.67
1,500	1.00	1.00	0.60	0.50	0.50	1.00	0.71	0.80	0.61	0.61	1.00	0.41	1.00	0.72	0.73
2,000	1.00	1.00	0.44	0.50	0.50	1.00	0.63	0.72	0.57	0.58	1.00	0.26	1.00	0.64	0.65
2,500	1.00	1.00	0.45	0.50	0.50	1.00	0.58	0.73	0.54	0.55	1.00	0.15	1.00	0.58	0.59
3,000	1.00	1.00	0.33	0.50	0.50	1.00	0.63	0.67	0.58	0.60	1.00	0.26	1.00	0.66	0.69
3,500	1.00	1.00	0.13	0.50	0.50	1.00	0.60	0.56	0.55	0.55	1.00	0.20	1.00	0.60	0.60
4,000	1.00	1.00	0.00	0.50	0.50	1.00	0.50	0.50	0.50	0.50	1.00	0.00	1.00	0.50	0.50
4,500	1.00	1.00	0.00	0.50	0.50	1.00	0.50	0.50	0.50	0.50	1.00	0.00	1.00	0.50	0.50
5,000	1.00	1.00	0.00	0.50	0.50	1.00	0.55	0.50	0.52	0.52	1.00	0.10	1.00	0.55	0.55

TABLE 5 | Experimental results using plant dataset.

IR	ANN Algorithm					DT Algorithm					Proposed H-DNN Algorithm				
	Acc	SP	SE	F1	Prec	Acc	SE	SP	F1	Prec	Acc	SE	SP	F1	Prec
1	0.51	1.00	0.96	0.99	0.98	0.73	0.97	0.96	0.97	0.97	0.95	0.94	0.97	0.95	0.95
100	0.99	1.00	0.55	0.76	0.49	0.99	0.88	0.77	0.82	0.68	1.00	0.77	1.00	0.87	0.87
500	1.00	1.00	0.36	0.50	0.50	1.00	0.79	0.68	0.63	0.62	1.00	0.58	1.00	0.76	0.74
1,000	1.00	1.00	0.45	0.50	0.50	1.00	0.77	0.73	0.61	0.59	1.00	0.53	1.00	0.71	0.67
1,500	1.00	1.00	0.54	0.50	0.50	1.00	0.67	0.77	0.58	0.57	1.00	0.33	1.00	0.65	0.64
2,000	1.00	1.00	0.50	0.50	0.50	1.00	0.68	0.75	0.60	0.61	1.00	0.36	1.00	0.70	0.73
2,500	1.00	1.00	0.33	0.50	0.50	1.00	0.64	0.67	0.57	0.57	1.00	0.27	1.00	0.64	0.64
3,000	1.00	1.00	0.11	0.50	0.50	1.00	0.71	0.56	0.57	0.55	1.00	0.43	1.00	0.64	0.61
3,500	1.00	1.00	0.00	0.50	0.50	1.00	0.80	0.50	0.61	0.58	1.00	0.60	1.00	0.71	0.67
4,000	1.00	1.00	0.00	0.50	0.50	1.00	0.80	0.50	0.62	0.61	1.00	0.60	1.00	0.75	0.71
4,500	1.00	1.00	0.00	0.50	0.50	1.00	0.58	0.50	0.54	0.54	1.00	0.17	1.00	0.58	0.58
5,000	1.00	1.00	0.00	0.50	0.50	1.00	0.61	0.50	0.56	0.56	1.00	0.22	1.00	0.61	0.61

TABLE 6 | ANN and DT results using human dataset.

IR	ANN Algorithm					DT Algorithm					Proposed H-DNN Algorithm				
	Acc	SP	SE	F1	Prec	Acc	SE	SP	F1	Prec	Acc	SE	SE	F1	Prec
1	0.50	1.00	0.99	0.86	0.85	0.72	0.97	0.97	0.90	0.90	0.94	0.94	0.95	0.94	0.94
100	0.99	1.00	0.79	0.50	0.50	0.99	0.85	0.89	0.67	0.66	0.99	0.69	1.00	0.83	0.83
500	1.00	1.00	0.56	0.50	0.50	1.00	0.75	0.78	0.63	0.63	1.00	0.51	1.00	0.76	0.76
1,000	1.00	1.00	0.46	0.50	0.50	1.00	0.67	0.73	0.61	0.65	1.00	0.34	1.00	0.72	0.81
1,500	1.00	1.00	0.41	0.50	0.50	1.00	0.76	0.71	0.62	0.61	1.00	0.53	1.00	0.74	0.72
2,000	1.00	1.00	0.45	0.50	0.50	1.00	0.79	0.72	0.63	0.62	1.00	0.57	1.00	0.77	0.75
2,500	1.00	1.00	0.24	0.50	0.50	1.00	0.75	0.62	0.59	0.57	1.00	0.50	1.00	0.69	0.65
3,000	1.00	1.00	0.13	0.50	0.50	1.00	0.80	0.56	0.58	0.56	1.00	0.60	1.00	0.67	0.62
3,500	1.00	1.00	0.00	0.50	0.50	1.00	0.60	0.50	0.55	0.55	1.00	0.20	1.00	0.60	0.60
4,000	1.00	1.00	0.00	0.50	0.50	1.00	0.80	0.50	0.65	0.65	1.00	0.60	1.00	0.80	0.80
4,500	1.00	1.00	0.00	0.50	0.50	1.00	1.00	0.50	0.56	0.54	1.00	1.00	1.00	0.62	0.57
5,000	1.00	1.00	0.00	0.50	0.50	1.00	0.92	0.50	0.63	0.60	1.00	0.83	1.00	0.76	0.69



the ANN we get a better result for almost all the IRs up to 1:5000 as compared to the performance gained by DT and ANN separately.

CONCLUSION

The production of imbalanced data in large quantities in Bioinformatics provides a great scope for work in the AI world. These kinds of problems require modern solutions for which many questions will arise that have to be justified and have

to satisfy all the complex conditions too. One such problem is the classification of these miRNA sequences which are very few in genome data and are similar to other RNA structures. The miRNA seems to help detect and diagnose cancer disease. Hence classifying this miRNA is a big challenge that has to be faced. So, a novel solution is required with minimal complexity that provides better results in terms of the TPR, time, and accurate classification of miRNA from the imbalanced genome dataset. In this paper, we have proposed a Hybrid Neural Network model in which we have integrated Deep ANN and Deep Decision Tree

TABLE 7 | ANN and DT results using arabidopsis dataset.

IR	ANN Algorithm					DT Algorithm					Proposed H-DNN Algorithm				
	Acc	SP	SE	F1	Prec	Acc	SE	SP	F1	Prec	Acc	SE	SE	F1	Prec
1	0.54	1.00	0.85	0.35	0.27	0.74	0.98	0.87	0.64	0.60	0.93	0.96	0.90	0.93	0.93
100	0.54	1.00	0.75	0.35	0.27	0.75	0.98	0.85	0.65	0.61	0.95	0.96	0.95	0.95	0.95
500	1.00	1.00	0.12	0.50	0.50	1.00	0.79	0.56	0.66	0.67	1.00	0.58	1.00	0.81	0.84
1,000	1.00	1.00	0.25	0.50	0.50	1.00	0.70	0.63	0.61	0.62	1.00	0.40	1.00	0.72	0.75
1,500	1.00	1.00	0.55	0.50	0.50	1.00	0.55	0.77	0.53	0.53	1.00	0.10	1.00	0.56	0.56
2,000	1.00	1.00	0.24	0.50	0.50	1.00	0.75	0.62	0.62	0.62	1.00	0.50	1.00	0.75	0.75
2,500	1.00	1.00	0.00	0.50	0.50	1.00	0.50	0.50	0.50	0.50	1.00	0.00	1.00	0.50	0.50
3,000	1.00	1.00	0.00	0.50	0.50	1.00	0.63	0.50	0.57	0.58	1.00	0.25	1.00	0.64	0.67
3,500	1.00	1.00	0.00	0.50	0.50	1.00	1.00	0.50	0.60	0.56	1.00	1.00	1.00	0.70	0.63
4,000	1.00	1.00	0.00	0.50	0.50	1.00	0.75	0.50	0.57	0.55	1.00	0.50	1.00	0.64	0.60
4,500	1.00	1.00	0.00	0.50	0.50	1.00	0.50	0.50	0.50	0.50	1.00	0.00	1.00	0.50	0.50
5,000	1.00	1.00	0.00	0.50	0.50	1.00	0.50	0.50	0.50	0.50	1.00	0.00	1.00	0.50	0.50

TABLE 8 | ANN and DT results using virus dataset.

IR	ANN Algorithm					DT Algorithm					Proposed H-DNN Algorithm				
	Acc	SP	SE	F1	Prec	Acc	SE	SP	F1	Prec	Acc	SE	SP	F1	Prec
1	0.52	0.48	0.99	0.52	0.52	0.86	0.94	0.92	0.86	0.92	0.94	0.92	0.95	0.94	0.94
100	0.99	1.00	0.32	0.50	0.49	0.99	1.00	0.92	0.90	0.93	1.00	1.00	1.00	0.94	0.90
200	0.99	1.00	0.16	0.50	0.50	0.99	1.00	0.96	0.48	0.44	0.99	0.00	1.00	0.50	0.50
300	1.00	1.00	0.00	0.50	0.50	0.99	0.00	0.98	0.76	0.65	0.99	1.00	0.99	0.75	0.67
400	1.00	1.00	0.00	0.50	0.50	1.00	0.00	1.00	0.41	0.49	1.00	0.00	1.00	0.50	0.50

Classifiers. The Deep ANN gives better results when compared with the existing result and good TPR with 2 epochs and up to IR 1:2000. Furthermore, when the model needs to be trained with IR up to 1:5000, Deep Decision Tree Classifier gives better results with good TPR than ANN. When we integrate both Deep ANN and Deep DT we get better performance as compared to both taken individually. In our future work, we would like to further extend by increasing the IR and reducing the time of convergence. We would also like to check the performance of the proposed model on other large genome datasets. The main aim is that the model should not follow any sort of sampling techniques and predict the miRNAs with good TPR.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

ER, SR, and RE contributed in problem formulation, implementation and results. KK, SP, SV, VV, and AM contributed in methodology identification and techniques to project results. All authors contributed to the article and approved the submitted version.

REFERENCES

- Peng Y, Croce CM. The role of microRNAs in human cancer. *Signal Transduct Target Ther.* (2016) 1:2857–67. doi: 10.1038/sigtrans.2015.4
- Bertoli G, Cava C, Castiglioni I. MicroRNAs: new biomarkers for diagnosis, prognosis, therapy prediction, and therapeutic tools for breast cancer. *Theranostics.* (2015) 10:1122–43. doi: 10.7150/thno.11543
- Li L, Xu J, Yang D, Tan X, Wang, H. Computational approaches for microRNA studies: A review. *Mamm Genome.* (2010) 21:1–12. doi: 10.1007/s00335-009-9241-2
- Allmer J, Yousef M. Computational methods for ab initio detection of micrnas. *Front Genet.* (2012) 3:209–12. doi: 10.3389/fgene.2012.00209
- Bugnon LA, Yones C, Milone DH, Stegmayer G. Deep neural architectures for highly imbalanced data in bioinformatics. *IEEE Trans Neural Netw Learn Syst.* (2019) 31:2857–67. doi: 10.1109/TNNLS.2019.2914471
- Gudy's, Szczeaniak MW, Sikora M, Makiłowska I. HuntMi: An efficient and taxon-specific approach in pre-miRNA identification. *BMC Bioinf.* (2013) 14:83. doi: 10.1186/1471-2105-14-83
- Zheng X, Xu S, Zhang Y, Huang X. Nucleotide-level convolutional neural networks for pre-mirna classification. *Sci Rep.* (2019) 9:1–6. doi: 10.1038/s41598-018-36946-4
- Wang S, Yao, X. Multiclass imbalance problems: Analysis and potential solutions. *IEEE Trans Syst Man Cybern B Cybern.* (2012) 42:1119–30. doi: 10.1109/TSMCB.2012.2187280

9. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng.* (2009) 21:1263–84. doi: 10.1109/TKDE.2008.239
10. Alazab M, Khan S, Krishnan SSR, Pham QV, Reddy MPK. A multidirectional LSTM model for predicting the stability of a smart grid. *IEEE Access.* (2020) 8:85454–63. doi: 10.1109/ACCESS.2020.2991067
11. Lin M, Tang K, Yao X. Dynamic sampling approach to training neural networks for multiclass imbalance classification. *IEEE Trans Neural Netw Learn Syst.* (2013) 24:647–60. doi: 10.1109/TNNLS.2012.2228231
12. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst Appl.* (2017) 73:220–39. doi: 10.1016/j.eswa.2016.12.035
13. Lin S, Gregory RI. MicroRNA biogenesis pathways in cancer. *Nature Rev Cancer.* (2015) 15:321–33. doi: 10.1038/nrc3932
14. Saçar MD, Allmer J. Machine learning methods for microRNA gene prediction. *Methods Mol Bio.* (2014) 1107:177–87. doi: 10.1007/978-1-62703-748-8_10
15. Shukla V, Varghese VK, Kabekkodu SP, Mallya S, Satyamoorthy K. A compilation of Web-based research tools for miRNA analysis. *Brief Funct Genomics.* (2017) 16:249–73. doi: 10.1093/bfpg/ew042
16. Xue C, Li F, He T, Liu, G-P, Li Y, Zhang X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinf.* (2005) 6:310. doi: 10.1186/1471-2105-6-310
17. Helvik SA, Snøve O, Sætrom P. Reliable prediction of Drosha processing sites improves microRNA gene prediction. *Bioinformatics.* (2007) 23:142–9. doi: 10.1093/bioinformatics/btl570
18. Sheng Y, Engström PG, Lenhard B. Mammalian MicroRNA prediction through a support vector machine model of sequence and structure. *PLoS ONE.* (2007) 2:e946. doi: 10.1371/journal.pone.0000946
19. Hertel J, Stadler PF. Hairpins in a Haystack: Recognizing microRNA precursors in comparative genomics data. *Bioinformatics.* (2006) 22:e197–202. doi: 10.1093/bioinformatics/btl257
20. Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, Brownstein MJ, et al. Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinf.* (2005) 6:267. doi: 10.1186/1471-2105-6-267
21. Huang T-H, Fan B, Rothschild M, Hu ZL, Li K, et al. MiRFinder: An improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinf.* (2007) 8:341. doi: 10.1186/1471-2105-8-341
22. Ding J, Zhou S, Guan J. MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multiloop features. *BMC Bioinf.* 11:S11. doi: 10.1186/1471-2105-11-S11-S11
23. Batuwita R, Palade V. microPred: Effective classification of premiRNAs for human miRNA gene prediction. *Bioinformatics.* (2009) 25:989–95. doi: 10.1093/bioinformatics/btp107
24. Klefogiannis D, Theofilatos K, Likothanassis S, Mavroudi, S. YamiPred: a novel evolutionary method for predicting pre-miRNAs and selecting relevant features. *IEEE/ACM Trans Comput Biol Bioinf.* (2015) 12:1183–92. doi: 10.1109/TCBB.2014.2388227
25. Pandya S, Ghayvat H. Ambient acoustic event assistive framework for identification, detection, and recognition of unknown acoustic events of a residence. *Adv Eng Inform.* (2021) 47:1012. doi: 10.1016/j.aei.2020.101238
26. Stegmayer G, Yones C, Kamenetzky L, Milone DH. High class-imbalance in pre-miRNA prediction: A novel approach based on deepSOM. *IEEE/ACM Trans Comput Biol Bioinf.* (2017) 14:1316–26. doi: 10.1109/TCBB.2016.2576459
27. Fischer A, Igel C. An introduction to restricted Boltzmann machines. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (Lecture Notes in Computer Science)*. Berlin, Germany: Springer (2012).
28. Ayachit G, Shaikh I, Pandya H, Das J. Salient features data and algorithms for microRNA screening from plants: a review on the gains and pitfalls of machine learning techniques. *Curr Bioinform.* (2020) 15:1091–103. doi: 10.2174/1574893615999200601121756
29. Reddy GT, Reddy MPK, Lakshmana K, Kaluri R, Rajput DS, Srivastava G. Analysis of dimensionality reduction techniques on big data. *IEEE Access.* (2020) 8:54776–88. doi: 10.1109/ACCESS.2020.2980942
30. Koumaki L. Deep learning models in genomics; are we there yet? *Comput Struct Biotechnol J.* (2020) 18:1466–73. doi: 10.1016/j.csbj.2020.06.017
31. Shi H, Wang D, Wu P, Cao Y, Chen, Y. Deep Multiple Kernel Learning for Prediction of MicroRNA Precursors. *Sci Program.* (2021) 2021. doi: 10.1155/2021/9969282
32. Peace RJ, Biggar KK, Storey KB, Green JR. A framework for improving microRNA prediction in non-human genomes. *Nucleic Acids Res.* (2015) 43:e138. doi: 10.1093/nar/gkv698
33. Ghayvat H, Pandya SN, Bhattacharya P, Zuhair M, Rashid M, Hakak S, et al. CP-BDHC: blockchain-based confidentiality-privacy preserving big data scheme for healthcare clouds and applications. *IEEE J Biomed Health Inform.* (2021) 6:1–1. doi: 10.1109/JBHI.2021.3097237
34. Liu B, Fang L, Chen J, Liu F, Wang X. miRNA-dis: MicroRNA precursor identification based on distance structure status pairs. *Molecular BioSyst.* (2015) 11:1194–204. doi: 10.1039/C5MB00050E
35. Wei L, Liao M, Gao Y, Ji R, He Z, Zou Q. Improved and promising identification of human microRNAs by incorporating a highquality negative set. *IEEE/ACM Trans Comput Biol Bioinf.* (2014) 11:192–201. doi: 10.1109/TCBB.2013.146
36. Ghayvat H, Awais M, Pandya S, Ren H, Akbarzadeh S, Chandra Mukhopadhyay S, et al. Smart aging system: uncovering the hidden wellness parameter for well-being monitoring and anomaly detection. *Sensors.* (2019) 19:766. doi: 10.3390/s19040766
37. Mahmud, Kaiser MS, Hussain A, Vassanelli S. Applications of deep learning and reinforcement learning to biological data. *IEEE Trans Neural Netw Learn Syst.* (2018) 29:2063–79.
38. Si Z, Yu H, Ma Z. Learning deep features for DNA methylation data analysis. *IEEE Access.* (2016) 4:2732–7. doi: 10.1109/ACCESS.2016.2576598
39. Stegmayer G, Gerard M, Milone D. Data mining over biological datasets: An integrated approach based on computational intelligence. *IEEE Comput Intell Mag.* (2012) 7:22–34. doi: 10.1109/MCI.2012.2215122
40. LeRoux N, Bengio Y. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Comput.* 20:1631–1649. doi: 10.1162/neco.2008.04-07-510
41. Thomas J, Thomas S, Sael, L. DP-miRNA: an improved prediction of precursor microRNA using deep learning model. *In Proc IEEE Int Conf Big Data Smart Comput.* (2017) 1:96–9. doi: 10.1109/BIGCOMP.2017.7881722
42. Lertampaiporn S, Thammarongtham C, Nukoolkit C, Kaewkamnerdpong B, Ruengjitchachawalya M. Heterogeneous ensemble approach with discriminative features and modified-SMOTEBagging for pre-miRNA classification. *Nucleic Acids Res.* (2013) 41:e21. doi: 10.1093/nar/gks878
43. Bugnon LA, Yones C, Milone DH, Stegmayer G. Genome-wide discovery of pre-miRNAs: comparison of recent approaches based on machine learning. *Brief Bioinformatics.* (2021) 22:bbaa184. doi: 10.1093/bib/bbaa184
44. Tasdelen A, Sen B. A hybrid CNN-LSTM model for pre-miRNA classification. *Sci Rep.* (2021) 11:1–9. doi: 10.1038/s41598-021-93656-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 R., Jain, Kotecha, Pandya, Reddy, E., Varadarajan, Mahanti and V. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.