Computers in Biology and Medicine

DHUpredET: A Comparative Computational Approach for Identification of Dihydrouridine Modification Sites in RNA Sequence

--Manuscript Draft--

Manuscript Number:	CIBM-D-24-11894
Article Type:	Full Length Article
Keywords:	Dihydrouridine; Extra tree classifier; RNA molecules; Position-specific two nucleotides; Feature extraction
Corresponding Author:	Md. Shazzad Hossain Shaon Oakland University Michigan State, USA, UNITED STATES
First Author:	Md. Fahim Sultan
Order of Authors:	Md. Fahim Sultan
	Tasmin Karim
	Md. Shazzad Hossain Shaon
	Sayed Mehedi Azim
	Md. Yeasin Biplob
	Md. Shoaib Hossain Alshan
	Iman Dehzangi
	Mst Shapna Akter
Abstract:	Dihydrouridine (DHU/D) is a modified nucleoside found in RNA molecules. It is formed by reducing the uridine nucleoside, a process that eliminates two hydrogen atoms. DHU has been associated with a variety of diseases, including cancer, brain tumors, and hormonal disorders. Predicting DHU locations in illness-associated RNAs can help us understand disease processes and identify possible treatment targets. However, laboratory-based detection of D sites is laborious and expensive. In this study, we conducted a comparative computational analysis to identify D sites, employing effective machine learning models and efficient feature encoding methods. Initially, we explored various state-of-the-art feature encoding approaches, evaluating 30 machine learning techniques for each, and favored the top eight models based on their independent testing and cross-validation outcomes. As a result, we introduced the DHUpredET, using the extra tree classifier methods for predicting DHU sites. The DHUpredET model demonstrated balanced performance across all evaluation criteria, outperforming state-of-the-art models by 8% and 14% in terms of accuracy and sensitivity on an independent test set. Further analysis revealed that the model achieved higher accuracy with position-specific two nucleotides (PS2) features, leading us to conclude that PS2 features are best suited for the DHUpredET model. Therefore, our proposed model emerges as the most favorite choice for predicting D sites. In addition, we conducted an in-depth analysis of local features and identified a particularly significant attribute with a feature score of 0.035 for PS2_299 attributes. This tool holds immense promise as an advantageous instrument for accelerating the discovery of D modification sites, which contributes to many targeting therapeutic and understanding RNA structure. The datasets used in the study and the source codes are publicly available at https://github.com/Shazzad-Shaon3404/DHUpredET-DHU-prediction.git.
Suggested Reviewers:	
Opposed Reviewers:	

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

none

30 September 2024

Dear,

Editor-in-Chief

Computers in Biology and Medicine, Elsevier.

I am writing to submit my manuscript entitled "**DHUpredET: A Comparative Computational Approach for Identification of Dihydrouridine Modification Sites in RNA Sequence.**" for consideration for publication in the journal "**Computers in Biology and Medicine**".

In this paper, we present an additional and effective model for detecting Dihydrouridine Modification Sites (D/ DHU) based on nine feature encoding techniques. Our model, called DHUpredET, outperformed existing approaches in terms of accuracy and effectiveness in discovering ACPs. Ultimately, this model aids in detecting DHU site more quickly and precisely, saving time and expense.

We believe that our work will be of interest to the readers of "Computers in Biology and Medicine" and will make a valuable contribution to the field of cancer treatment.

Type of Manuscript: Research article.

*** Correspondence**

Md. Shazzad Hossain Shaon, Department of Computer Science and Informatics, Oakland University, Rochester, MI 48309, USA.

Thank you for considering our submission. We look forward to hearing from you.

Sincerely,
Md Shazzad Hossain Shaon
Department of Computer Science and Informatics
Email: shazzad15-3404@diu.edu.bd

Highlight

1. Employed nine feature encoding approaches: modern NLP features and biologically significant properties.

2. Developed DHUpredET, a machine learning-based method which outperforms earlier approaches.

3. Built a simpler model and evaluated it using numerous evaluation criteria to ensure robust and consistent performance.

4. Determined which characteristics are most relevant for prediction, with (PS2) appearing as the most critical.

DHUpredET: A Comparative Computational Approach for Identification of Dihydrouridine Modification Sites in RNA Sequence

Md. Fahim Sultan [a], Tasmin Karim [a], Md. Shazzad Hossain Shaon [a *], Sayed Mehedi Azim [b], Md. Yeasin Biplob [c], Md. Shoaib Hossain Alshan [d] Iman Dehzangi [b,e,], Mst Shapna Akter [f]

a. Department of Computer Science and Informatics, Oakland University, Rochester, MI 48309, USA.
b. Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA.
c. Department of Computer Science & Engineering, Daffodil International University, Daffodil Smart City, Birulia, Dhaka, 1216, Bangladesh.
d. Department of Computer Science, Rutgers University, Camden, NJ, 08102, USA
e. Department of Computer Science and Engineering, Shanto-Mariam University of Creative Technology, Uttara, Dhaka-1230, Bangladesh.
f. Department of Computer Science and Engineering, Oakland University, Michigan, USA.

Email in order:

fahim15-3416@diu.edu.bd,

tasmin15-2920@diu.edu.bd,

shazzad15-3404@diu.edu.bd,

sayedmehedi.azim@rutgers.edu ,

yeasin15-3055@diu.edu.bd ,

shoaibhossainalshan@gmail.com

i.dehzangi@rutgers.edu ,

akter@oakland.edu

Abstract- Dihydrouridine (DHU/D) is a modified nucleoside found in RNA molecules. It is formed by reducing the uridine nucleoside, a process that eliminates two hydrogen atoms. DHU has been associated with a variety of diseases, including cancer, brain tumors, and hormonal disorders. Predicting DHU locations in illness-associated RNAs can help us understand disease processes and identify possible treatment targets. However, laboratory-based detection of D sites is laborious and expensive. In this study, we conducted a comparative computational analysis to identify D sites, employing effective machine learning models and efficient feature encoding methods. Initially, we explored various state-of-the-art feature encoding approaches, evaluating 30 machine learning techniques for each, and favored the top eight models based on their independent testing and cross-validation outcomes. As a result, we introduced the DHUpredET, using the extra tree classifier methods for predicting DHU sites. The DHUpredET model demonstrated balanced performance across all evaluation criteria, outperforming state-of-the-art models by 8% and 14% in terms of accuracy and sensitivity on an independent test set. Further analysis revealed that the model achieved higher accuracy with position-specific two nucleotides (PS2) features, leading us to conclude that PS2 features are best suited for the DHUpredET model. Therefore, our proposed model emerges as the most favorite choice for predicting D sites. In addition, we conducted an in-depth analysis of local features and identified a particularly significant attribute with a feature score of 0.035 for PS2_299 attributes. This tool holds immense promise as an advantageous instrument for accelerating the discovery of D modification sites, which contributes to many targeting therapeutic and understanding RNA structure. The datasets used in the study and the source codes are publicly available at https://github.com/Shazzad-Shaon3404/DHUpredET-DHU-prediction.git.

Keywords: Dihydrouridine, Extra tree classifier, RNA molecules, Position-specific two nucleotides, Feature extraction

Introduction

RNA Modification refers to a sequence of chemical changes made to initial ribonucleic acid (RNA) transcripts that culminate in the synthesis of mature RNA molecules essential to a variety of biological activities [1]. At present, more than 300 particular types of RNA modifications are discovered [1-5]. Furthermore, the environment of RNA modification is strikingly similar across all three domains of life: eukaryotes, bacteria, and archaea. Despite their great evolutionary deficiencies, these domains contain numerous alterations and modifying enzymes [1, 6, 7].

Dihydrouridine (DHU), denoted by the letter "D" in RNA sequences and having the chemical formula $C_9H_{14}N_2O_6$, is a significant alteration identified in numerous forms of RNA, including transfer RNA (tRNA), messenger RNA (mRNA), and small nucleolar RNA (snoRNA) in different organisms. This alteration has sparked enthusiasm because of its prevalent presence and probable functional importance in RNA molecules [8-10]. The dihydrouridine biosynthesis molecules catalyze the reduction of uridine's (U) C5-C6 double bond, leading to the formation of dihydrouridine [6]. This alteration can be observed at the variable position of the anticodon helix

in tRNA molecules and poses major health consequences, including the development of lung cancer, Alzheimer's disease, and Huntington's disorders [11-13]. Therefore, D sites are crucial for understanding RNA molecules' structure, function, and regulatory activities, as well as their prospective uses in diagnostics and therapies [14].

Experimental methods for recognizing D sites, such as mass spectrometry or antibody-based assays, can be complicated, time-consuming, and expensive. On the other hand, computational approaches provide a high-throughput capabilities alternative, facilitating experts to evaluate massive RNA sequence databases rapidly and effectively. Advanced computational models, especially using machine learning and artificial intelligence, can achieve high levels of accuracy in predicting D sites. They can detect subtle patterns and features in RNA sequences that might be missed by traditional methods. **Figure 1** presents the fishbone diagram of the existing machine learning methods proposed to predict D sites.

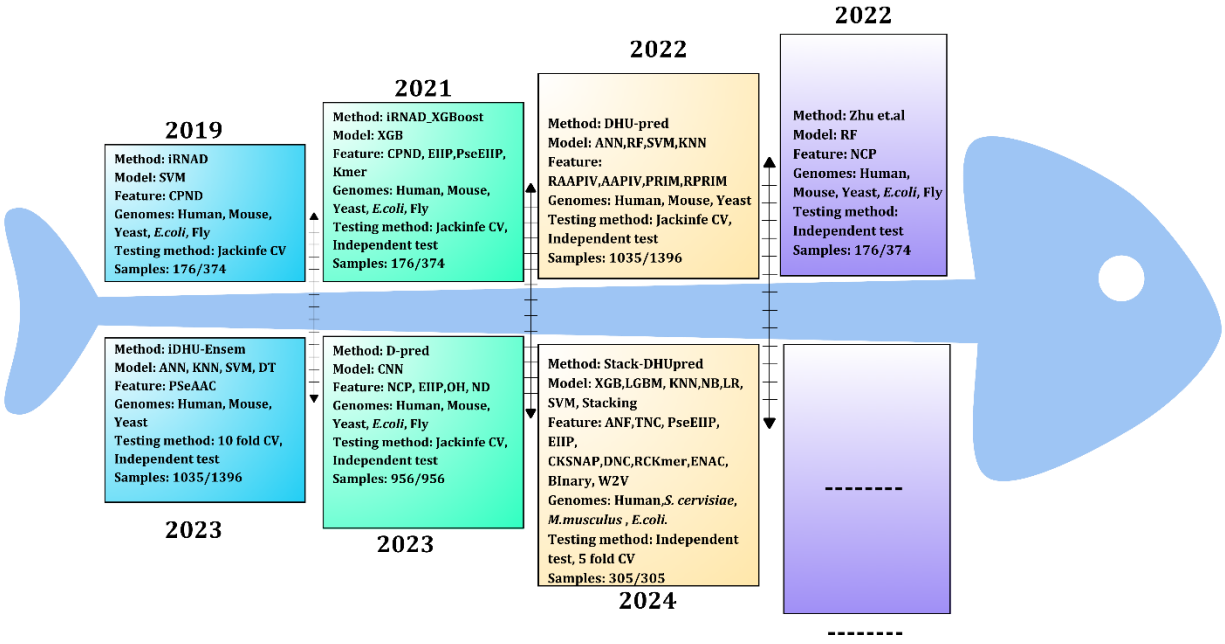

Figure 1. An overview of the existing methods in various years, represented with a fishbone diagram.

During the past few years, several machine learning-based methods proposed to predict D sites. In 2019, Xu et al. proposed an iRNAD framework based on a support vector machine (SVM) model with various features and achieved promising results [14]. Later on, Dou et al. developed iRNAD_XGBOOST with extreme gradient boosting (XGB) on different feature selection approaches [15]. In 2022, Zhu et al. introduced a Random Forest (RF) based machine learning approach for the D sites identification [16]. At the same time, Suleman et al. proposed the DHU-pred method with RF, where the authors used a relatively larger dataset compared to their prior studies to train their model [17]. Later on, the iDHU-Ensem method was proposed, where Suleman et al. included artificial neural network (ANN), K-nearest neighbor (KNN), SVM, and decision tree (DT), and obtained a better result [18]. In the same time, Yu et al. developed a D-pred model with a local self-attention layer and a convolutional neural network (CNN) [9]. Self-attention techniques, particularly local self-attention, might prove computationally expensive, especially

when dealing with complex sequences. Most recently, Roshid et al. developed stacking-based machine learning models called Stack-DHUpred, where the authors used probabilistic values from the baseline models and obtained a better outcome compared to their previous studies [19]. However, there is still room for improvement in the performance.

In general, most of the previous studies used a limited number of feature embedding approaches and utilized small datasets to train their model. As a result, the present study focuses on recently obtained datasets using various feature embedding approaches to determine the best feature extraction method and framework for predicting D sites.

In this study, we executed an in-depth analysis employing various machine learning techniques. We also use nine feature encoding methods, including state-of-the-art methods generated from natural language features, biologically feasible physical and chemical properties, nucleic acid-based compositional features, and residue composition data. Through this extensive analysis, we sought to uncover the most informative features and optimal machine-learning algorithms for predicting D sites in RNA sequences based on various types of evaluation metrics. As a result, we propose DHUpredET, a machine learning-based method than can significantly outperform previous studies found in the literature. Notably, it excels at leveraging position-specific two nucleotides (PS2) features with appropriate execution. DHUpredET identifies the positive class with greater than 85% and specifies the negative class with more than 82% accuracy, which produces a balance outcome. The datasets used in the study and the source code for DHUpredET are publicly available at https://github.com/Shazzad-Shaon3404/DHUpredET-DHU-prediction.git.

2. Materials and methods

2.1. Dataset descriptions

In this study, we use the dataset that was introduced in [19]. This dataset contains a total of 805 active compounds from different species, such as *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae*, and *Escherichia coli* [3,20,10]. Next, they used a 90% cluster database at high identity with tolerance (CD-HIT) method [21] to reduce the redundant data of the positive molecules. As a result, the number of positive samples was reduced to 305 samples. They also collected same number of negative samples (305) from different studies [22-24]. Finally, they used, 244 positive and 244 negative compounds as the training set, while an additional set of 61 positive and 61 negative samples were reserved as the testing set. **Table 1**. Describe the details of the datasets.

Table 1. Dataset description of the train file and test file used in this study.

Mode	Total	Train	Test
D sites (positive)	305 D sites	244 D sites	61 D sites
non-D sites (negative)	305 non-D sites	244 non-D sites	61 non-D sites

2.2. Overall methodologies

This study presents a machine learning-based approach for DHU prediction. **Figure 2** demonstrates the overall methodology of the study. At first, we collected the dataset through evaluation of previous studies, resulting in an extensive variety of RNA sequences from different species. Then, we used multiple feature embedding techniques on the collected data. In order to improve efficiency, we utilized a variety of machine learning models. As a result, the eight greatest

performing models were chosen after a rigorous evaluation of their overall performances. Among them, DHUpredET emerged as the most promising model for predicting D sites and PS2 as the most effective encoding methods.

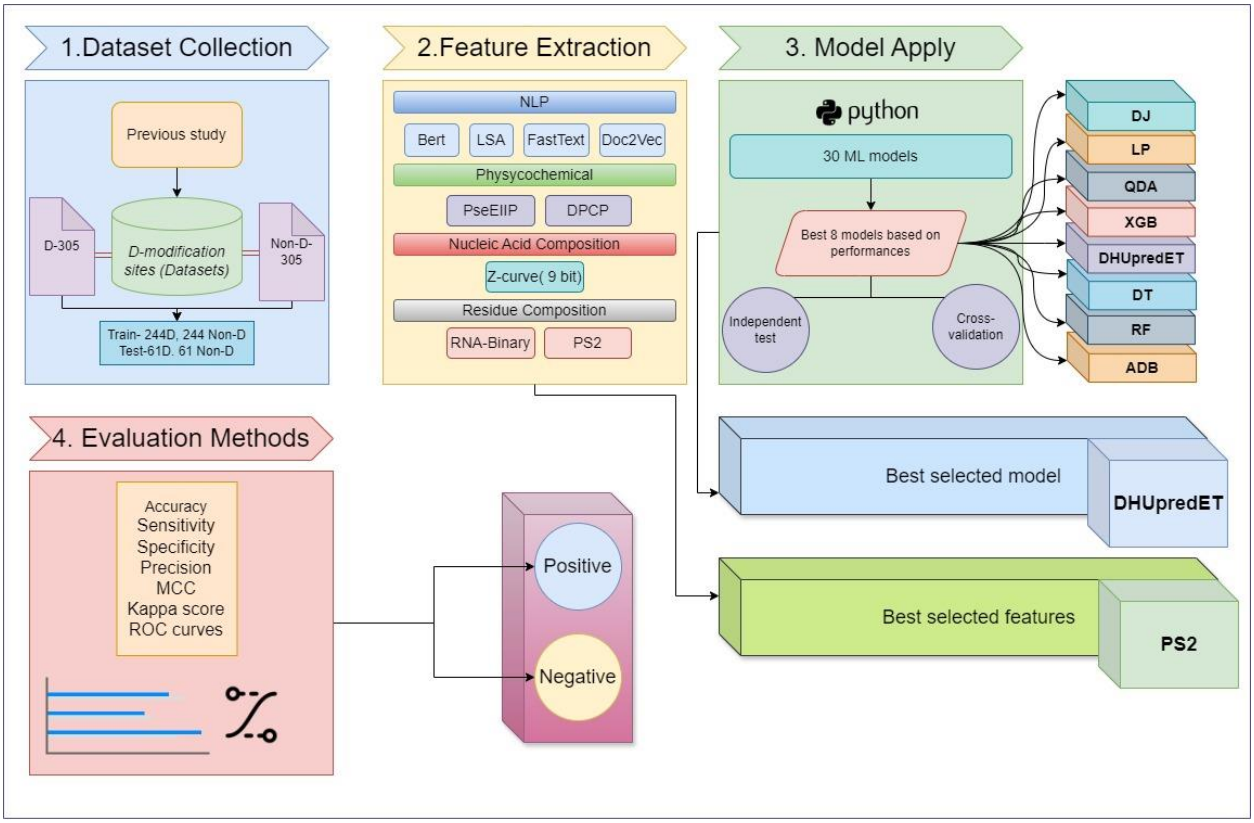

Figure 2. Workflows of the current study. Dataset collection from the previous studies, feature extraction approach of four kinds of descriptors with nine feature encoding methods, application of the various models, and selection process, the overall comparison of the models, and selecting the optimal feature and best-fit models for the study.

2.3. Feature extraction

Feature extraction is vital step for a better performance of machine learning models because it transforms raw data into a more appropriate representation that allows for successful learning and prediction. In this study, we use various feature extraction techniques to determine the most effective attributes for D sites prediction. These techniques included NLP-based methods such as bidirectional encoder representations from transformers (BERT) [25-28], FastText word embedding (FastText) [29,30], latent semantic analysis (LSA) [31-33], and document-to-vector (Doc2Vec) [34-36]. Furthermore, physicochemical characteristics such as electron-ion interaction pseudopotentials of trinucleotide (PseEIIP) [37, 38] and dinucleotide physicochemical properties (DPCP) [39] are also deployed. The study further utilized residue composition-based RNA binary (Binary) [40, 41] and PS2 [42,43], in addition to nucleic acid composition-based features, including Z-curve phase-specific mononucleotides (Z-curve 9-bit) [44]. These strategies are intended to improve the accuracy and efficacy of D site detection. In the following subsections, all these methods are described, in detail.

2.3.1 NLP-based feature extractions

BERT: Bert is well-known in the data mining sector for making encoder transformers. With a pre-trained approach, it has received a lot of attention and recognition. It encodes contextual by analyzing words bidirectionally, indicating that it considers both left and right circumstances. It tokenizes input text, fine-tunes for certain tasks, and supplies complemented word visualizations, allowing it to excel at various text-based activities [25-28]. The formula of the BERT feature can be expressed as:

$$B = \prod_{m \in Ma} \mathrm{P}(m \mid W, A) \tag{1}$$

$$\mathrm{P}(m \mid W, A) = X(W.O_m) \tag{2}$$

In equations 1 and 2, *Ma* is the vector of the masked word in the sentence, *W* is the vector associated with a word present in the input sentence, *A* is the parameter of the approach, O_m signifies the output embedding for the masked word, *m*, $W.O_m$ dot product computed between the word vector and the outcome embedding vector, and $\mathrm{P}(m \mid W, A)$ refers to the probability of the word *m*, conditioned on both *W* and *A*.

FastText: FastText is an NLP model developed by Facebook AI Research that excels at word embedding and text classification tasks. FastText describes languages by organizing symbols across n-grams. As an instance, the pattern of "XYZM" can be split into smaller n-grams such as "XY," "XYZ," "XYM," "YZ," "YZM,", and "ZM". The n-grams preserve mathematical data which allows them to encode sentences with equivalent frequencies [29, 30]. The equation can be stated as:

$$F = -\frac{1}{T}\sum_{n=1}^{T} x_n \log(P(TL_{a_n})) \tag{3}$$

In equation 3, *F* denotes the loss function. It averages a negative logarithm of the anticipated distribution of probabilities *P*, throughout a sequence of n-gram features denoted as x_n, each of which can be represented by word embeddings of a_n. The look-up matrices of word embeddings are denoted by *L*. *T* represents linear output transformation. The SoftMax function's coefficient is applied to the output.

LSA: LSA is the most popular NLP-based embedding system in the data mining fields. It works by initially displaying a collection of text documents as a matrix, with rows representing distinct terms (words) and columns representing documents. Then, it performs singular value decomposition (SVD) on the matrix. LSA decreases the matrix's dimensionality by maintaining just the top *k* singular values and accompanying singular vectors, resulting in a lower-dimensional semantic space. Each phrase and document is represented as a vector in this semantic space, which captures latent semantic associations [31-33]. The mathematical term of the LSA feature denotes:

$$LSA = U \sum V_N \tag{4}$$

Where, U is the left singular vectors, representing the relationships between terms in the reduced-dimensional semantic space. V_N is the transpose of matrix V, representing associations among texts in a reduced-dimensional conceptual environment.

Doc2Vec: Doc2Vec is a sophisticated tool for creating document embeddings that capture the semantic content of documents, rendering it suitable for various text analysis purposes. It was introduced by researchers at Google and is based on the distributed memory (DM) and distributed bag of words (DBOW) architectures. It operates by learning a neural network to determine the wider context of a document, whereas word-to-vector analyzes the surrounding words of a particular word. Each document is allocated a distinctive vector in a space with high dimensions that captures the semantic meaning [34-36]. The formula can be stated as:

$$D = N1 \sum i = 1N(a + V) \tag{5}$$

Here, D is the main vector, N is the number of words, a represents the vectors of each word in the document, V refers to the particular vector, and $N1$ is the pooling operation.

2.3.2 Physicochemical-based feature extractions

DPCP: This is the structural and chemical properties of neighboring nucleotide pairs inside RNA sequences. This approach sheds light on RNA's stability, flexibility, and functional qualities by considering different aspects, including base stacking, hydrogen bonding, and nucleotide interactions. [39]. The formula is referred to as:

$$D_{ij} = f(P_i, Q_j) \tag{6}$$

where D_{ij} is the combined property of nucleotides, P_i represents the physicochemical values of the i^{th} nucleotide base in the dinucleotide pair, Q_j is the property value of the j^{th} base, and f is a function that combines the properties of two bases for the overall property.

PseEIIP: This feature introduces attributes of modifications to the original EIIP values to improve their efficacy in identifying nucleotide sequences. The actual formula for determining PseEIIP values varies based on the individual alterations and characteristics used in the computation. PseEIIP values are often calculated by combining EIIP values (A =1.1260, T/U=0.1335, G=0.0806, C=0.1340). These values are based on each base's electron-ion interaction potential. EIIP values are useful for a variety of statistical inquiries, such as sequence alignment, motif finding, and identifying functional components in nucleic acid sequences [37-38]. The formula is represented as:

$$EIIP = \frac{\sum_{i=1}^{n} E_i.N_i}{\sum_{i=1}^{n} N_i} \tag{7}$$

$$PseEIIP = \sum_{i=1}^{n} (I_i.E_i) \tag{8}$$

In equation 7, n is the total number of nucleotides, E_i is the electron-ion values of the i^{th} position, and N_i is the frequency of the i^{th}.

In equation 8, I_i represents weights assigned to each nucleotide base or feature in the sequence, E_i is the electron-ion values of the i^{th} position, and n is the total number of nucleotides.

2.3.3 Residue composition-based feature extractions

Binary Profile: For this feature, each amino acid in the sequence is assigned to a 20-dimensional array with just one non-zero component. This mapping generates a matching matrix, where each row represents an amino acid and each column represents a specific property [40, 41].

PS2: The PS2 matrix is commonly 16×16, with each row and column representing one of the available dinucleotides (16 in total: AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, and UU). The value in each cell of the matrix represents the frequency or likelihood of the associated dinucleotide pair occurring at that location in the RNA sequence [42. 43]. The equation can be represented as:

$$M_{ij} = \frac{count(d_{ij})}{N-1} \quad (9)$$

Where, M_{ij} represents the frequency, N-1 is the total number of positions, and $count(d_{ij})$ is the number of dinucleotide pairs of d_{ij} at position i.

2.3.4 Nucleic acid composition-based feature extractions

Z-curve 9-bit: Each nucleotide base (A = 100000000, C = 010000000, G = 001000000, and U = 000100000) in the Z-curve 9-bit representation is translated to a 9-bit binary vector that encodes features like local stacking energy, hydrogen bonding, and base-pairing preferences. These features are determined using the positional connections of neighboring nucleotides [44].

2.4 Machine Learning Models Construction

In this study, different machine-learning models are used to predict D sites. After extensive testing with different hyperparameters, eight models stood out for their outstanding performance. These include AdaBoost (ADB), label propagation (LP), quadratic discriminant analysis (QDA), extreme gradient boosting (XGB), Decision Tre (DT), Random Forest (RF), Decision Jungle (DJ), and an extra tree classifier-based approach (ET), referred to as DHUpredET for PS2 feature extraction procedure as it performed better[45-54]. These models were chosen based on their ability to reliably predict D sites, assessed by an exhaustive analysis and juxtaposition of their effectiveness indicators. In **Table 2**, we illustrate all the hyperparameters used for these classifiers.

Table 2. Hyperparameters description of the study

Models	**Best parameters**
ADB	estimator=None, *, n_estimators=50, learning_rate=1.0, algorithm='SAMME.R', random_state=None

LP	kernel='knn',max_iter=1000
QDA	*, priors=None, reg_param=0.0, store_covariance=False, tol=0.0001
XGB	n_estimators=100, max_depth=100, learning_rate=0.1, subsample=1.0, colsample_bytree=1.0, reg_alpha=30, reg_lambda=30, gamma=0, min_child_weight=1
DT	max_depth=10, criterion='entropy', min_samples_split=10, min_samples_leaf=1, max_features=15 , random_state=42
RF	n_estimators=100,criterion='entropy',max_features="sqrt",random_state=100
DHUpredET	n_estimators=100, random_state=10, max_depth=None, min_samples_split=4, min_samples_leaf=1, max_features='auto', bootstrap=False, class_weight=None, criterion='gini', n_jobs=None

2.4.1 DHUpredET model development procedure

This study used a wide range of algorithms, including ten classical approaches, ten voting-based meta-learning methods, and ten stack-based meta-learning strategies, which are shared in supplementary materials and provided in the GitHub repository We chose eight models for their reasonable scores over numerous assessment measures. It was observed that while other models exhibited some improvement in accuracy, they displayed a tendency towards bias in positive class identification. As a result, we selected the Extra Trees Classifier which outperformed all other classifiers using PS2 features.

The extra tree classifier is a decision tree-based ensemble learning approach that integrates many trees to increase forecast accuracy and resilience. In this design, the model is configured to employ 100 decision trees (n_estimators=100), which ensures a varied range of classifiers and reduces overfitting. The random_state option is set to 10, guaranteeing that the findings are reproducible over several runs. With max_depth=None, the decision trees can grow until all leaves are pure or there are less than four samples (min_samples_split=4) at a node, with at least one sample required at each leaf (min_samples_leaf=1). The max_features option is set to 'auto', indicating that all features are evaluated for determining the optimal split at each node. Bootstrap sampling is deactivated (bootstrap=False), indicating that the full dataset is used to build each tree. The class_weight contention is set to None, which means that all classes are considered equally. The criterion option is set to 'gini', which specifies the criteria for dividing nodes. Finally, setting n_jobs=None indicates that the model will use all available processors for parallel processing. **Figure 3.** Demonstrates the overall procedure of the DHUpredET model.

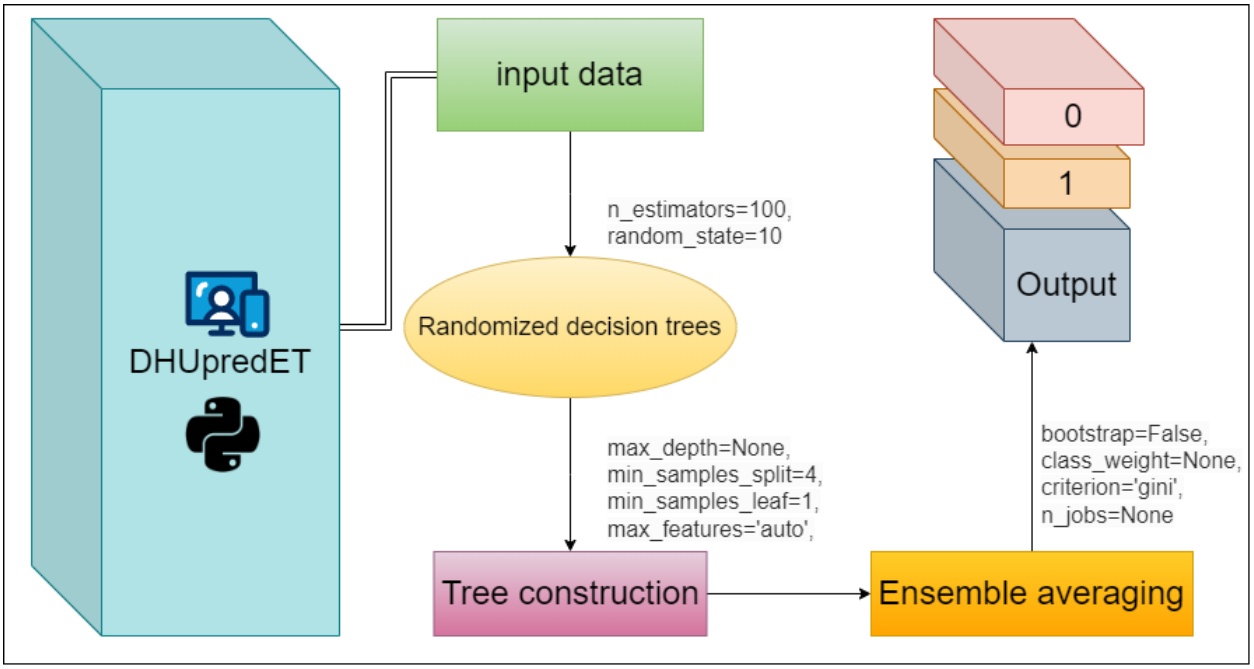

Figure 3. Overall strategies of the DHUpredET model.

2.5 Evaluation Metrics

The study used various evaluation metrics to properly evaluate the models' performance. The metrics used were the Matthews Correlation Coefficient (MCC), sensitivity (Sen), specificity (Spe), precision (Pre), recall (Rec), F1 score (F1s), kappa score (Kpp), and accuracy (Acc) [56-60]. These metrics are formulated as follows:

$$\begin{cases} Acc = \frac{(TP+TN)}{(TP+TN+FP+FN)} \\ Spe = \frac{TN}{(TN+FP)} \\ Sen = \frac{TP}{(TP+FN)} \\ MCC = \frac{TP*TN-FP*FN}{\sqrt{TP+FP*(TP+FN)*(TN+FP)}} \\ Pre = \frac{TP}{(TP+FP)} \\ Rec = \frac{TP}{(TP+FN)} \\ F1s = \frac{2*(Pre*Rec)}{((Pre+Rec))} \\ Kpp = \frac{2*(TP*TN-FP*FN)}{(TP+FP)*(FP+TN)+(TP+FN)*(FN+TN)} \end{cases} \quad (10)$$

Where TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negative.

3. Experimental results

In order to identify the best model for D site prediction, we conducted a thorough analysis, evaluating a wide range of models. Our goal was to identify the best model and features suitable for the prediction of D site. **Tables 3** provides a complete performance evaluation of our models. The corresponding results for 5-fold cross-validation is provided as the supplementary material in **Table_S1**.

Table 3. Independent test performances of eight models

Descriptor	**Classifier**	**Acc (present in percentage with 1 decimal point)**	**MCC (Decimal number with 2 decimal point prediction)**	**Kpp (Decimal number with 2 decimal point prediction)**	**Pre (present in percentage with 1 decimal point)**	**Rec present in percentage with 1 decimal point**	**F1s (Decimal number with 2 decimal point prediction)**	**Sen (present in percentage with 1 decimal point)**	**Spe (present in percentage with 1 decimal point)**
DPCP	RF	65.6	0.31	0.31	66.1	63.9	0.65	67.2	0 63.9
	ADB	69.6	0.39	0.39	66.1	77.0	0.71	62.3	77.0
	DT	54.9	0.09	0.09	54.5	50.9	0.56	50.8	59.0
	XGB	66.3	0.32	0.32	65.6	68.8	0.67	63.9	68.8
	QDA	59.8	0.19	0.19	61.1	54.1	0.57	65.5	54.1
	DJ	66.3	0.32	0.32	65.1	70.4	0.67	62.3	70.4
	LP	66.3	0.32	0.32	66.6	65.5	0.66	67.2	65.5
	ET	69.7	0.39	0.39	70.0	68.8	0.69	70.4	68.8
Z curve 9 bit	RF	65.5	0.31	0.31	63.7	72.1	0.67	59.0	72.1
	ADB	63.1	0.26	0.26	62.5	65.5	0.64	60.6	65.5
	DT	51.6	0.03	0.03	51.7	49.1	0.50	54.1	49.1
	XGB	59.0	0.19	0.18	65.7	37.7	0.47	80.3	37.7
	QDA	65.5	0.31	0.31	64.6	68.8	0.66	62.3	68.8
	DJ	59.8	0.20	0.19	62.0	50.8	0.55	68.8	50.8
	LP	63.1	0.26	0.26	62.1	67.2	0.64	59.0	67.2
	ET	**65.5**	0.31	0.31	65.0	67.2	0.66	63.9	67.2
Binary	RF	72.1	0.45	0.44	77.5	62.3	0.69	81.9	62.3
	ADB	72.1	0.44	0.44	71.4	73.7	0.72	70.4	73.7
	DT	68.8	0.37	0.37	70.9	63.9	0.67	73.7	63.9
	XGB	72.9	0.49	0.45	86.8	54.1	0.66	91.8	54.1
	QDA	69.6	0.48	0.39	96.1	40.9	0.57	98.3	40.9
	DJ	73.7	0.47	0.47	73.7	73.7	0.73	73.7	73.7
	LP	66.3	0.36	0.32	61.3	88.5	0.72	44.2	88.5
	ET	74.5	0.50	0.49	80.0	65.5	0.72	83.6	65.5
PS2	RF	78.6	0.57	0.57	79.6	77.0	0.78	80.3	77.0
	ADB	73.7	0.47	0.47	73.7	75.4	0.74	72.1	75.4
	DT	59.8	0.20	0.19	65.0	42.6	0.51	77.0	42.6
	XGB	67.2	0.38	0.34	81.8	44.2	0.57	90.1	44.2
	QDA	50.8	0.02	0.01	53.8	11.4	0.18	90.1	11.4
	DJ	71.3	0.42	0.42	69.7	75.4	0.72	67.2	75.4
	LP	69.6	0.43	0.39	63.6	91.8	0.75	47.5	91.8
	ET (DHUpredET)	85.2	**0.70**	**0.70**	**86.4**	**83.6**	**0.85**	**86.8**	**83.6**

LSA	RF	50.0	0.00	0.00	50.0	59.0	0.54	40.9	59.0
	ADB	**52.4**	0.05	0.04	51.9	67.2	0.58	37.7	67.2
	DT	49.1	0.01	0.01	49.3	60.6	0.54	37.7	60.6
	XGB	50.0	0.00	0.00	0.00	0.00	0.00	100.0	0.00
	QDA	52.4	0.05	0.04	51.9	65.5	0.57	39.3	65.5
	DJ	56.5	0.14	0.13	54.7	75.4	0.63	37.7	75.4
	LP	50.0	0.00	0.00	50.0	68.8	0.57	31.1	68.8
	ET	47.5	0.05	0.04	48.0	60.6	0.53	34.4	60.6
BERT	RF	67.6	0.34	0.34	67.8	65.5	0.66	68.8	65.5
	ADB	58.2	0.16	0.16	57.3	63.9	0.60	52.4	63.9
	DT	54.1	0.08	0.08	53.8	57.3	0.55	50.8	57.3
	XGB	63.9	0.27	0.27	63.9	63.9	0.63	63.9	63.9
	QDA	50.8	0.09	0.01	1.00	1.64	0.03	100.0	0.01
	DJ	60.6	0.21	0.21	60.6	60.6	0.60	60.6	60.6
	LP	67.2	0.34	0.34	64.7	75.4	0.69	59.0	75.4
	ET	**70.4**	0.41	0.40	71.1	68.8	0.70	72.1	68.8
FastText	**RF**	54.9	0.09	0.09	55.6	49.1	0.52	60.6	49.1
	ADB	42.6	0.14	0.14	43.2	47.5	0.45	37.7	47.5
	DT	52.4	0.05	0.04	51.7	72.1	0.60	32.7	72.1
	XGB	50.0	0.00	0.00	0.00	0.00	0.00	100.0	0.00
	QDA	44.2	0.11	0.11	44.6	47.5	0.46	40.9	47.5
	DJ	53.2	0.07	0.06	52.2	75.4	0.61	31.1	75.4
	LP	50.0	0.00	0.00	50.0	40.9	0.45	59.0	40.9
	ET	46.7	0.06	0.06	46.7	47.5	0.47	45.9	47.5
Doc2Vec	RF	70.6	0.41	0.41	70.9	71.1	0.71	70.4	70.1
	ADB	63.1	0.26	0.26	61.7	68.8	0.65	57.3	68.8
	DT	60.6	0.21	0.21	61.0	59.0	0.60	62.3	59.0
	XGB	50.0	0.00	0.00	0.00	0.00	0.00	100.0	0.00
	QDA	50.0	0.00	0.00	0.00	0.00	0.00	100.0	0.00
	DJ	60.6	0.21	0.21	58.9	70.4	0.64	50.8	70.4
	LP	68.8	0.37	0.37	67.1	73.7	0.70	63.9	73.7
	ET	71.3	0.42	0.42	70.9	72.1	0.71	70.4	72.1
PseEIIP	RF	73.7	0.47	0.47	73.0	75.4	0.74	72.1	75.4
	ADB	65.5	0.31	0.31	64.1	70.4	0.67	60.6	70.4
	DT	59.8	0.19	0.19	60.3	57.3	0.58	62.3	57.3
	XGB	64.6	0.29	0.29	64.0	67.2	0.65	62.3	67.2
	QDA	63.9	0.27	0.27	63.0	67.2	0.65	60.6	67.2
	DJ	66.3	0.33	0.32	64.7	72.1	0.68	60.6	72.1
	LP	72.1	0.44	0.44	69.0	80.0	0.74	63.9	80.3
	ET	**73.7**	0.47	0.47	73.0	75.0	0.74	72.1	75.4

As seen from **Table 3**, most of the applied models have delivered promising results from various extracted features. Among them, ET using PS2 achieved the best results. Therefore, ET as a classifier and PS2 as the feature encoding method used to build DHUpredET model. DHUpredET obtains 85.3 rate of accuracy with excellent performances in other evaluation criteria. DHUpredET enhances the prediction accuracy by 10% compared to other physicochemical-based encoding methods. In NLP-based embedding, ADB and RF models acquired optimal results using LSA and FastText features. Overall, evaluations, the physicochemical-based PS2 features would be the most optimal features.

Figure 4 shows the results of our comparison study using both the independent test method and the 5-fold cross-validation procedure. Our analysis is focused on comparing F1 scores and accuracy metrics. As shown in this figure, the results achieved for the independent test set and 5-fold cross-validation are consistent which demonstrates the generality and the robustness of our proposed method. DHUpredET consistently achieves excellent F1s and Acc across both assessment methodologies. This implies that DHUpredET has been effective at reducing both false positives and false negatives, indicating its effectiveness for its intended application. In the process of obtaining excellent scores in both measures across many evaluation techniques (independent test and cross-validation), DHUpredET demonstrates stability and durability in performance evaluation.

Figure 4. Overall comparison of Accuracy and F1 score between independent test and 5-fold-cross-validation approach, where **(A)** Accuracy (Acc) **(B)** F1 Score (F1s)

The receiver operating characteristic (ROC) curve and precision-recall (PR) curve are two commonly used evaluation methods in the data science fields [61, 62]. The ROC curve illustrates the relationship between Sen and Spe, while the PR curve provides information about a model's efficiency, particularly when dealing with unbalanced datasets in which the number of negative cases significantly exceeds the positive ones. **Figure 5** shows the ROC and PR curves for our models using the independent test. Notably, the DHUpredET model performs adequately in all the subplots (A-I). In Subplot A (PS2, containing ROC and PR curves), DHUpredET receives noteworthy scores of 91.1% in ROC and 91.5% in PR curves. These results highlight that the model appropriately identifies both positive and negative categories, excelling at detecting positive examples while remaining specific for negative ones.

On the other hand, the PR curves demonstrates the model's outstanding performance. It also demonstrates the usefulness of the model in the presence of biased data sets. As shown in subplot D (Binary, containing ROC and PR curves), our model performed better with 92.3% of PR curves. Based on the findings of the present study, DHUpredET reliably performed the best across all subplots (A-I) compared to other models.

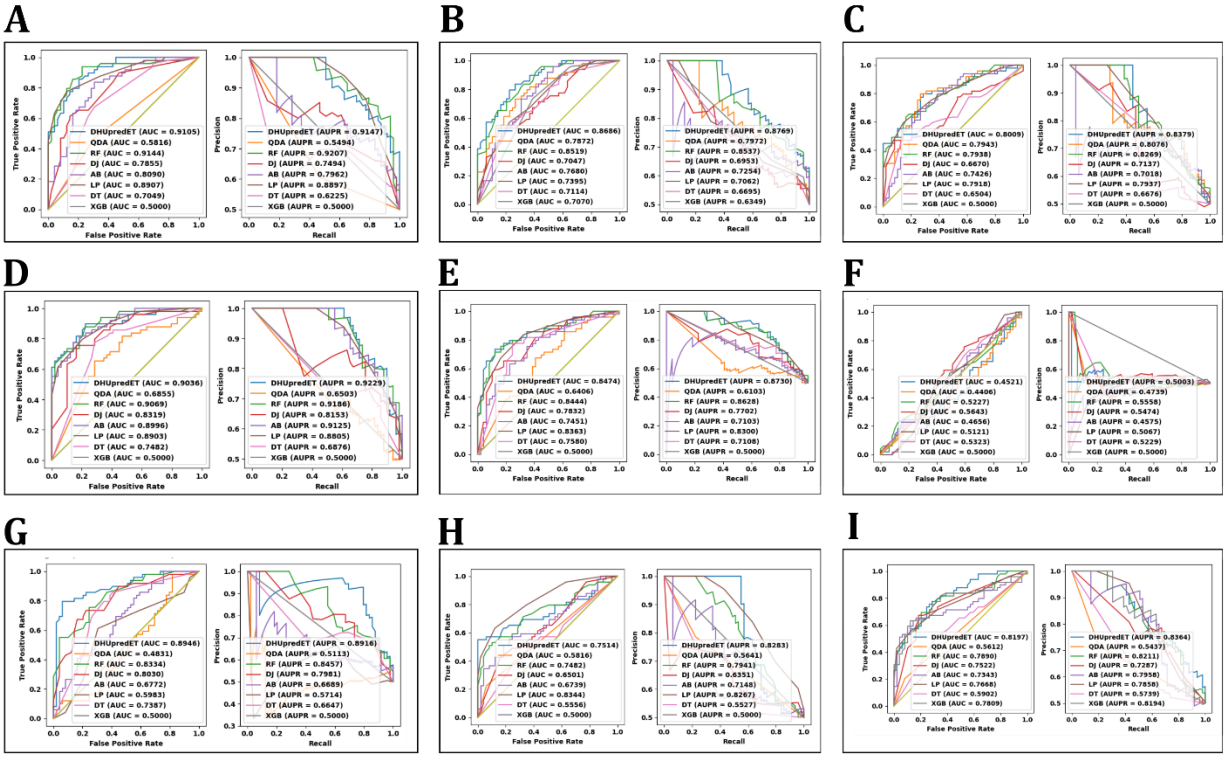

Figure 5. ROC curves and PR curves visualization based on independent test method with all feature extraction methods. **(A)** PS2, containing ROC and PR curves **(B)** PseEIIP, containing ROC and PR curves **(C)** DPCP, containing ROC and PR curves **(D)** Binary, containing ROC and PR curves **(E)** Z curve 9 bit, containing ROC and PR curves **(F)** LSA, containing ROC and PR curves **(G)** FastText, containing ROC and PR curves **(H)** Doc2Vec, containing ROC and PR curves **(I)** BERT, containing ROC and PR curves.

3.1 Feature analysis

As our study found PS2 features have the most optimal attributes, therefore in **Figure 6**, We analyzed the contribution of each feature in our model's decision-making in PS2 feature extractions. Based on the results, we can conclude that the PS2_299 feature is the most important in our model, as indicated by its significant value achieved of 0.035. The PS2_471 feature follows closely after, with a notable significance score of 0.030, suggesting that it contributes considerably to the model's predictive capabilities. Furthermore, the PS2_316 feature appears as a significant contributor, instead with a significantly lower relevance score of 0.025. These feature significance ratings provide valuable details about our model's explanatory capacity. It is apparent that the three features—PS2_299, PS2_471, and PS2_316—play vital parts in the model's decision-making process. Their prominence emphasizes their importance in capturing the underlying patterns involved with D site identification.

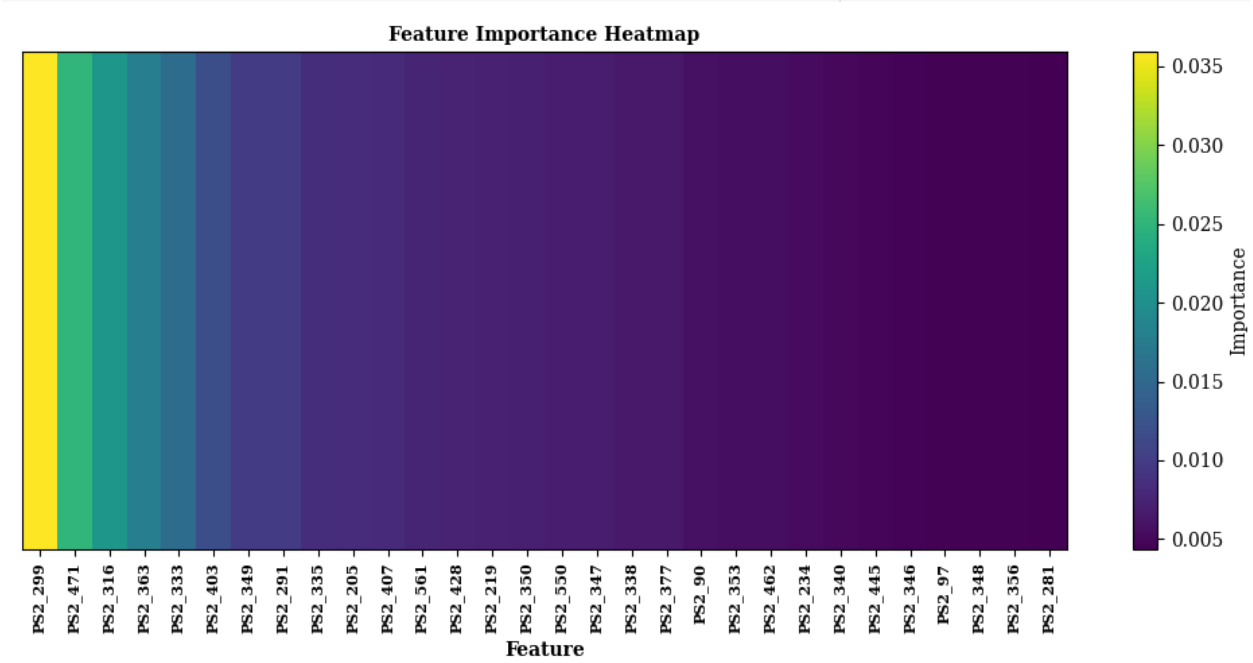

Figure 6. A heatmap of the DHUpredET model is based on PS2 features, and the study exhibits the top 30 important features.

3.2 Compariosn of DHUpredET with the previous models

In **Table 4**, we compare DHUpredET to the state-of-the-art methods, evaluating their performance across key metrics including Acc, Sen, AUC, MCC, PR, and Spe.

Table 4. Comparative analysis of the most recent methods based on the independent dataset with DHUpredET model.

Model	**Acc**	**MCC**	**Sen**	**Spe**	**AUC**	**PR**
Dpred [9]	0.68	0.384	0.508	0.853	0.703	0.706
Stack-DHUpred [19]	0.778	0.567	0.725	0.830	0.871	0.893
iDHU-Ensem [18]	0.574	0.152	0.689	0.459	0.574	0.541
iRNAD [14]	0.721	0.489	0.508	0.934	0.721	0.696
DHUpredET	**0.8525**	**0.7053**	**0.8689**	0.8361	**0.9105**	**0.9147**

As shown in Table 4, DHUpredET achives an accuracy of 85.3%, MCC of 70.53%, Sen of 56.9%, AUC of 91.1%, and PR of 91.5%. These results greatly outperform previous methodologies. DHUpredET model improved accuracy by more than 5%compared to all other models on the independent test dataset. Despite the greatest increase in accuracy, our model performed adequately across various evaluation metrics. The considerable improvement in MCC of more than 0.2 and Sen of more than 10% demonstrates that our model efficiently identifies positive classes while maintaining a balanced Spe for negative classes. Additionally, the AUC score increased by more than 3%, while the PR score boosted by 1%. The above results indicate that our model outperforms state-of-the-art approaches and is well-suited to properly discriminating between positive and negative classes in DHU prediction.

The comparison diagram in **Figure 8** visually illustrates the performance gap between the DHUpredET model and other methods. DHUpredET occupies a larger space, indicating its superior performance across multiple evaluation metrics compared to other methods..

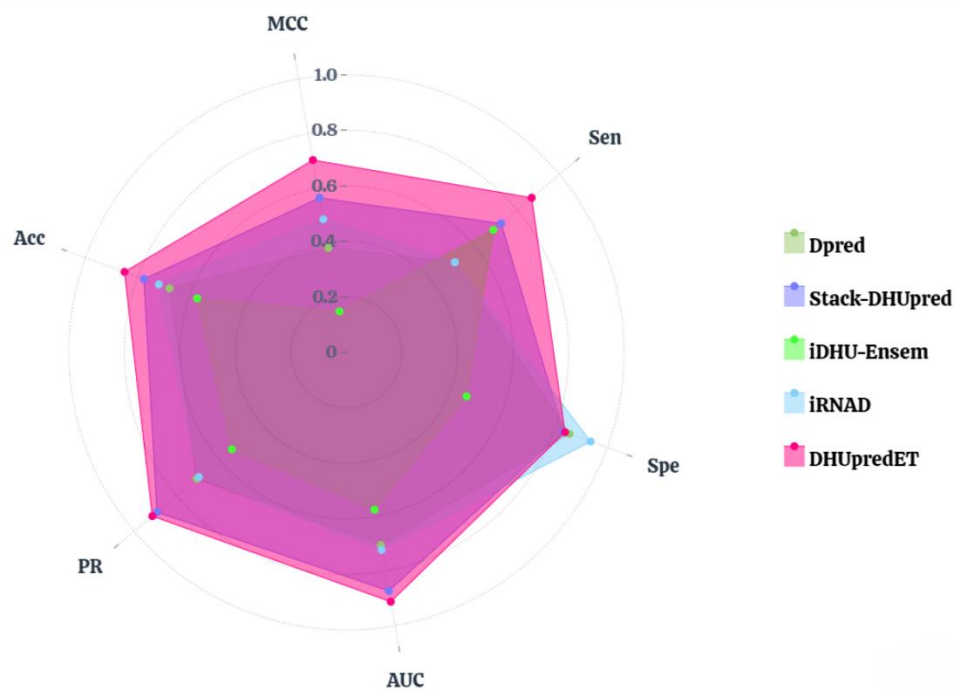

Figure 7. Comparison of state-of-the-art models with DHUpredET based on accuracy, MCC, AUC, PR, specificity, and sensitivity.

4. Discussion

Dysregulation of RNA modifications has been associated with a variety of diseases, including malignancy and neurological disorders. Identifying DHU alteration sites may provide insight into disease causes and possible treatment targets. Experimental approaches for detecting RNA modifications, such as DHU, can be time-consuming and costly, with varying degrees of sensitivity and specificity.

Computational techniques provide an alternative way for predicting modification sites with adequate accuracy, improving the prioritizing of experimental validation efforts. These approaches are important because they allow for the systematic analysis of massive datasets that would be hard to manage manually and provide a scalable answer to comprehending complicated biological processes by allowing for the full identification of D sites across a wide range of RNA molecules. The computer-based approach is especially useful when a preliminary high-throughput screening is required to focus future experimental efforts on the most promising targets. These predictions can thus accelerate the pace of discovery by directing empirical research towards the most relevant areas, enhancing the efficiency and effectiveness of biological study. In our extensive exploration of various studies, we observed a prevalent rhyme among authors, which presented different

techniques for detecting D sites. These representations generally employ various feature extraction techniques and machine learning strategies. To conduct an exhaustive comparative analysis, we utilized a variety of feature extraction approaches and analyzed 30 algorithms for each feature set. After a comprehensive analysis, we evaluated the final models based on their performance and selected the best features and methods. We explored integrating several feature sets, such as physicochemical features only, compositional features only, and all features paired together, then observed that PS2 features outperformed all of them. As a result, we adjusted our attention to using specific factors, favoring both best-fit models and features. The aim was to develop a cost-effective model without jeopardizing functionality. After rigorous testing and hyperparameter optimization, we determined that the ET model was the best choice. Using PS2 features, which emphasize local interdependence between nearby nucleotides, enables the model to improve its prediction effectiveness by exploiting perplexing nucleotide interactions.

Conclusion

Dihydrouridine is a RNA modification that influences both the structural and functional patterns of RNA, playing an essential role in manufacturing proteins and cellular adaption processes. Here we developed a new machine learning technique, called DHUpredET, for predicting DHU using extra tree as a classifier and PS2 feature encoding. Our proposed framework achieved 85.3% in terms of prediction accuracy, significantly outperforming the state-of-the-art models found in the literature by over 5%. Her ewe also comprehensively study variables that are strongly connected to the identification of D sites. The DHUpredET model improves our ability to find dihydrouridine sites with excellent accuracy and enhances our understanding of the underlying processes that control RNA modifications from the computational perspective. The datasets used in the study and the source code for DHUpredET are publicly available at https://github.com/Shazzad-Shaon3404/DHUpredET-DHU-prediction.git.

Data and code availability

The datasets of the study and the codes are available at this GitHub link:
https://github.com/Shazzad-Shaon3404/DHUpredET-DHU-prediction.git

Contributions

Conceptualization: Md. Fahim Sultan, Tasmin Karim, Md. Shazzad Hossain Shaon; **Data curation:** Tasmin Karim, Md. Shazzad Hossain Shaon, Md. Fahim Sultan, Md, Yeasin Biplob, Md. Shoaib Hossain Alshan; **Formal analysis:** Md. Fahim Sultan, Tasmin Karim, Md. Shazzad Hossain Shaon, Mst Shapna Akter; **Investigation:** Sayed Mehedi Azim, Md. Fahim Sultan, Tasmin Karim, Md. Shazzad Hossain Shaon, Mst Shapna Akter; **Methodology:** Md. Shazzad Hossain Shaon, Md. Fahim Sultan, Tasmin Karim, Sayed Mehedi Azim, Md. Shoaib Hossain Alshan, Mst Shapna Akter; **Project administration:** Md. Shazzad Hossain Shaon, Md. Fahim Sultan, Tasmin Karim, Sayed Mehedi Azim, Mst Shapna Akter, Iman Dehezangi; **Resources:** Md. Shazzad Hossain Shaon, Tasmin Karim, Sayed Mehedi Azim, Md. Yeasin Biplob, Md. Shoaib Hossain Alshan; **Supervision:** Iman Dehzangi ,Sayed Mehedi Azim , Md. Shazzad Hossain Shaon, Mst Shapna Akter; **Validation:** Iman Dehzangi ,Sayed Mehedi Azim , Md. Shazzad Hossain Shaon, Mst Shapna Akter; **Visualization:** Tasmin Karim, Md. Shazzad Hossain Shaon, Md. Shoaib Hossain Alshan; **Writing—original draft:** Md. Fahim Sultan, Tasmin Karim, Md. Shazzad Hossain Shaon, Md. Yeasin Biplob, Md. Shoaib Hossain Alshan; **Writing—review editing:** Iman Dehzangi, Sayed Mehedi Azim, Md. Shazzad Hossain Shaon, Mst Shapna Akter, Tasmin Karim; **Software:** Tasmin Karim, Md. Fahim Sultan, Md. Shazzad Hossain Shaon, Iman Dehezangi, Mst Shapna Akter;

The final version of the manuscript has been read and approved by all authors.

References

1. Post-translational modification https://en.wikipedia.org/wiki/Post-translational_modification. (accessed on 10 February 2024)
2. Arzumanian, V.A., Dolgalev, G.V., Kurbatov, I.Y., Kiseleva, O.I. and Poverennaya, E.V., 2022. Epitranscriptome: review of top 25 most-studied RNA modifications. International Journal of Molecular Sciences, 23(22), p.13851. https://doi.org/10.3390/ijms232213851
3. Boccaletto, P., Stefaniak, F., Ray, A., Cappannini, A., Mukherjee, S., Purta, E., Kurkowska, M., Shirvanizadeh, N., Destefanis, E., Groza, P. and Avşar, G., 2022. MODOMICS: a database of RNA modification pathways. 2021 update. Nucleic acids research, 50(D1), pp.D231-D235. https://doi.org/10.1093/nar/gkj084
4. Cappannini, A., Ray, A., Purta, E., Mukherjee, S., Boccaletto, P., Moafinejad, S.N., Lechner, A., Barchet, C., Klaholz, B.P., Stefaniak, F. and Bujnicki, J.M., 2024. MODOMICS: a database of RNA modifications and related information. 2023 update. Nucleic Acids Research, 52(D1), pp.D239-D244. https://doi.org/10.1093/nar/gkad1083
5. Shi, H., Wei, J. and He, C., 2019. Where, when, and how: context-dependent functions of RNA methylation writers, readers, and erasers. *Molecular cell*, *74*(4), pp.640-650. https://doi.org/10.1016/j.molcel.2019.04.025

6. Foltan, J.S., 2008. tRNA genes and the genetic code. Journal of theoretical biology, 253(3), pp.469-482. https://doi.org/10.1016/j.jtbi.2008.03.006
7. Goodenbour, J.M. and Pan, T., 2006. Diversity of tRNA genes in eukaryotes. Nucleic acids research, 34(21), pp.6137-6146. https://doi.org/10.1093/nar/gkl725
8. Edmonds, C.G., Crain, P.F., Gupta, R., Hashizume, T., Hocart, C.H., Kowalak, J.A., Pomerantz, S.C., Stetter, K.O. and McCloskey, J.A., 1991. Posttranscriptional modification of tRNA in thermophilic archaea (Archaebacteria). Journal of bacteriology, 173(10), pp.3138-3148. https://doi.org/10.1128/jb.173.10.3138-3148.1991
9. Yu, F., Tanaka, Y., Yamashita, K., Suzuki, T., Nakamura, A., Hirano, N., Suzuki, T., Yao, M. and Tanaka, I., 2011. Molecular basis of dihydrouridine formation on tRNA. Proceedings of the National Academy of Sciences, 108(49), pp.19593-19598. https://doi.org/10.1073/pnas.1112352108
10. Draycott, A.S., Schaening-Burgos, C., Rojas-Duran, M.F., Wilson, L., Schärfen, L., Neugebauer, K.M., Nachtergaele, S. and Gilbert, W.V., 2022. Transcriptome-wide mapping reveals a diverse dihydrouridine landscape including mRNA. PLoS Biology, 20(5), p.e3001622. https://doi.org/10.1371/journal.pbio.3001622
11. Kato, T., Daigo, Y., Hayama, S., Ishikawa, N., Yamabuki, T., Ito, T., Miyamoto, M., Kondo, S. and Nakamura, Y., 2005. A novel human tRNA-dihydrouridine synthase involved in pulmonary carcinogenesis. Cancer research, 65(13), pp.5638-5646. https://doi.org/10.1158/0008-5472.CAN-05-0600
12. Mendez, M.F., 2017. Early-onset Alzheimer disease. Neurologic clinics, 35(2), pp.263-281. https://doi.org/10.1016/j.ncl.2017.01.005
13. Durr, A., Gargiulo, M. and Feingold, J., 2012. The presymptomatic phase of Huntington disease. Revue neurologique, 168(11), pp.806-808. https://doi.org/10.1016/j.neurol.2012.07.003
14. Xu, Z.C., Feng, P.M., Yang, H., Qiu, W.R., Chen, W. and Lin, H., 2019. iRNAD: a computational tool for identifying D modification sites in RNA sequence. Bioinformatics, 35(23), pp.4922-4929. https://doi.org/10.1093/bioinformatics/btz358
15. Dou, L., Zhou, W., Zhang, L., Xu, L. and Han, K., 2021. Accurate identification of RNA D modification using multiple features. RNA biology, 18(12), pp.2236-2246. https://doi.org/10.1080/15476286.2021.1898160
16. Zhu, H., Ao, C.Y., Ding, Y.J., Hao, H.X. and Yu, L., 2022. Identification of D Modification sites using a random forest model based on nucleotide chemical properties. International Journal of Molecular Sciences, 23(6), p.3044. https://doi.org/10.3390/ijms23063044
17. Suleman, M.T., Alkhalifah, T., Alturise, F. and Khan, Y.D., 2022. DHU-Pred: accurate prediction of dihydrouridine sites using position and composition variant features on diverse classifiers. PeerJ, 10, p.e14104. https://doi.org/10.7717/peerj.14104
18. Suleman, M.T., Alturise, F., Alkhalifah, T. and Khan, Y.D., 2023. iDHU-Ensem: Identification of dihydrouridine sites through ensemble learning models. Digital Health, 9, p.20552076231165963. https://doi.org/10.1177/20552076231165963
19. Harun-Or-Roshid, M., Maeda, K., Manavalan, B. and Kurata, H., 2024. Stack-DHUpred: Advancing the accuracy of dihydrouridine modification sites detection via stacking approach. Computers in Biology and Medicine, 169, p.107848. https://doi.org/10.1016/j.compbiomed.2023.107848
20. Sun, W.J., Li, J.H., Liu, S., Wu, J., Zhou, H., Qu, L.H. and Yang, J.H., 2016. RMBase: a resource for decoding the landscape of RNA modifications from high-throughput

sequencing data. Nucleic acids research, 44(D1), pp.D259-D265. https://doi.org/10.1093/nar/gkv1036

21. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28(23), pp.3150-3152. https://doi.org/10.1093/bioinformatics/bts565
22. Manavalan, B., Hasan, M.M., Basith, S., Gosu, V., Shin, T.H. and Lee, G., 2020. Empirical comparison and analysis of web-based DNA N4-methylcytosine site prediction tools. Molecular Therapy-Nucleic Acids, 22, pp.406-420. https://doi.org/10.1016/j.omtn.2020.09.010
23. Hasan, M.M., Tsukiyama, S., Cho, J.Y., Kurata, H., Alam, M.A., Liu, X., Manavalan, B. and Deng, H.W., 2022. Deepm5C: a deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy. Molecular Therapy, 30(8), pp.2856-2867. https://doi.org/10.1016/j.ymthe.2022.05.001
24. Hasan, M.M., Basith, S., Khatun, M.S., Lee, G., Manavalan, B. and Kurata, H., 2021. Meta-i6mA: an interspecies predictor for identifying DNA N 6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. Briefings in Bioinformatics, 22(3), p.bbaa202. https://doi.org/10.1093/bib/bbaa202
25. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692
26. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805
27. Huo, H. and Iwaihara, M., 2020. Utilizing BERT pretrained models with various fine-tune methods for subjectivity detection. In Web and Big Data: 4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, September 18-20, 2020, Proceedings, Part II 4 (pp. 270-284). Springer International Publishing. https://doi.org/10.1007/978-3-030-60290-1_21
28. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F. and Dai, J., 2019. Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530. https://doi.org/10.48550/arXiv.1908.08530
29. Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, pp.135-146. https://doi.org/10.1162/tacl_a_00051
30. Choi, J. and Lee, S.W., 2020. Improving FastText with inverse document frequency of subwords. Pattern Recognition Letters, 133, pp.165-172. https://doi.org/10.1016/j.patrec.2020.03.003
31. Kwon, H., Kim, J. and Park, Y., 2017. Applying LSA text mining technique in envisioning social impacts of emerging technologies: The case of drone technology. Technovation, 60, pp.15-28. https://doi.org/10.1016/j.technovation.2017.01.001
32. Yu, B., Xu, Z.B. and Li, C.H., 2008. Latent semantic analysis for text categorization using neural network. Knowledge-Based Systems, 21(8), pp.900-904. https://doi.org/10.1016/j.knosys.2008.03.045

33. Chen, W.K., Chen, L.S. and Pan, Y.T., 2021. A text mining-based framework to discover the important factors in text reviews for predicting the views of live streaming. Applied Soft Computing, 111, p.107704. https://doi.org/10.1016/j.asoc.2021.107704
34. Le, Q. and Mikolov, T., 2014, June. Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.
35. Chen, Q. and Sokolova, M., 2021. Specialists, scientists, and sentiments: Word2Vec and Doc2Vec in analysis of scientific and medical texts. SN Computer Science, 2, pp.1-11. https://doi.org/10.1007/s42979-021-00807-1
36. Mishra, S., Pappu, A. and Bhamidipati, N., 2019, May. Inferring advertiser sentiment in online articles using wikipedia footnotes. In Companion proceedings of the 2019 world wide web conference (pp. 1224-1231). https://doi.org/10.1145/3308560.3316752
37. nidia, Robson P., et al. "Feature extraction approaches for biological sequences: a comparative study of mathematical features." Briefings in Bioinformatics 22.5 (2021): bbab011. https://doi.org/10.1093/bib/bbab011
38. Han, Siyu, et al. "LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property." Briefings in bioinformatics 20.6 (2019): 2009-2027. https://doi.org/10.1093/bib/bby065
39. Manavalan, B., Basith, S., Shin, T.H., Lee, D.Y., Wei, L. and Lee, G., 2019. 4mCpred-EL: an ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome. Cells, 8(11), p.1332. https://doi.org/10.3390/cells8111332
40. Chen, Z., Chen, Y.Z., Wang, X.F., Wang, C., Yan, R.X. and Zhang, Z., 2011. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. PloS one, 6(7), p.e22930. https://doi.org/10.1371/journal.pone.0022930
41. Chen, Z., Zhou, Y., Song, J. and Zhang, Z., 2013. hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1834(8), pp.1461-1467. https://doi.org/10.1016/j.bbapap.2013.04.006
42. Liu, B., Gao, X. and Zhang, H., 2019. BioSeq-Analysis2. 0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic acids research, 47(20), pp.e127-e127. https://doi.org/10.1093/nar/gkz740
43. Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R. and Virgin, H.W., 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature biotechnology, 34(2), pp.184-191. https://www.nature.com/articles/nbt.3437#citeas
44. Gao, F. and Zhang, C.T., 2004. Comparison of various algorithms for recognizing short coding sequences of human genes. Bioinformatics, 20(5), pp.673-681. https://doi.org/10.1093/bioinformatics/btg467
45. Ying, C., Qi-Guang, M., Jia-Chen, L. and Lin, G., 2013. Advance and prospects of AdaBoost algorithm. Acta Automatica Sinica, 39(6), pp.745-758. https://doi.org/10.1016/S1874-1029(13)60052-X
46. Garza, S.E. and Schaeffer, S.E., 2019. Community detection with the label propagation algorithm: a survey. Physica A: Statistical Mechanics and its Applications, 534, p.122058. https://doi.org/10.1016/j.physa.2019.122058

47. Tharwat, A., 2016. Linear vs. quadratic discriminant analysis classifier: a tutorial. International Journal of Applied Pattern Recognition, 3(2), pp.145-180. https://doi.org/10.1504/IJAPR.2016.079050
48. Bose, S., Pal, A., SahaRay, R. and Nayak, J., 2015. Generalized quadratic discriminant analysis. Pattern Recognition, 48(8), pp.2676-2684. https://doi.org/10.1016/j.patcog.2015.02.016
49. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I. and Zhou, T., 2015. Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), pp.1-4.
50. Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785
51. Suthaharan, S. and Suthaharan, S., 2016. Decision tree learning. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, pp.237-269. https://doi.org/10.1007/978-1-4899-7641-3_10
52. Rigatti, S.J., 2017. Random forest. Journal of Insurance Medicine, 47(1), pp.31-39. https://doi.org/10.17849/insm-47-01-31-39.1

Qi, Y., 2012. Random forest for bioinformatics. Ensemble machine learning: Methods and applications, pp.307-323. https://doi.org/10.1007/978-1-4419-9326-7_11

54. Shotton, J., Sharp, T., Kohli, P., Nowozin, S., Winn, J. and Criminisi, A., 2013. Decision jungles: Compact and rich models for classification. Advances in neural information processing systems, 26.
55. Sharaff, A. and Gupta, H., 2019. Extra-tree classifier with metaheuristics approach for email classification. In Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018 (pp. 189-197). Springer Singapore. https://doi.org/10.1007/978-981-13-6861-5_17
56. Oostwal, E., Straat, M. and Biehl, M., 2021. Hidden unit specialization in layered neural networks: ReLU vs. sigmoidal activation. Physica A: Statistical Mechanics and its Applications, 564, p.125517. https://doi.org/10.1016/j.physa.2020.125517
57. Umakantha, N., 2016. A new approach to probability theory with reference to statistics and statistical physics. Journal of Modern Physics, 7(09), p.989. http://dx.doi.org/10.4236/jmp.2016.79090
58. Kraemer HC. Kappa coefficient. Wiley StatsRef: statistics reference online. 2014 Apr 14:1-4. https://doi.org/10.1002/9781118445112.stat00365.pub2
59. Satu, M.S., Akter, T. and Uddin, M.J., 2017, February. Performance analysis of classifying localization sites of protein using data mining techniques and artificial neural networks. In 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 860-865). IEEE. https://doi.org/10.1109/ECACE.2017.7913023
60. Thabtah, F., 2019. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics for Health and Social Care, 44(3), pp.278-297. https://doi.org/10.1080/17538157.2017.1399132
61. Hoo, Z.H., Candlish, J. and Teare, D., 2017. What is an ROC curve?. *Emergency Medicine Journal*, *34*(6), pp.357-359. https://doi.org/10.1136/emermed-2017-206735

686 62. Davis, J. and Goadrich, M., 2006, June. The relationship between Precision-Recall and
687 ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp.
688 233-240). https://doi.org/10.1145/1143844.1143874
689

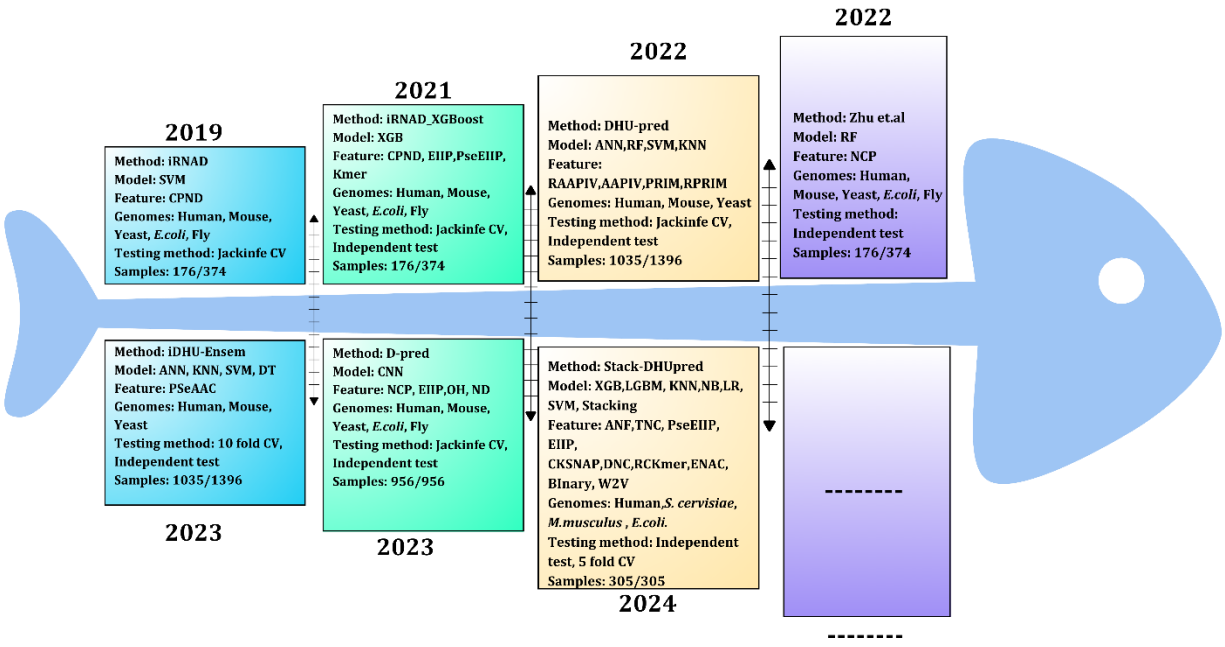

Figure 1. An overview of the existing methods in various years, represented with a fishbone diagram.

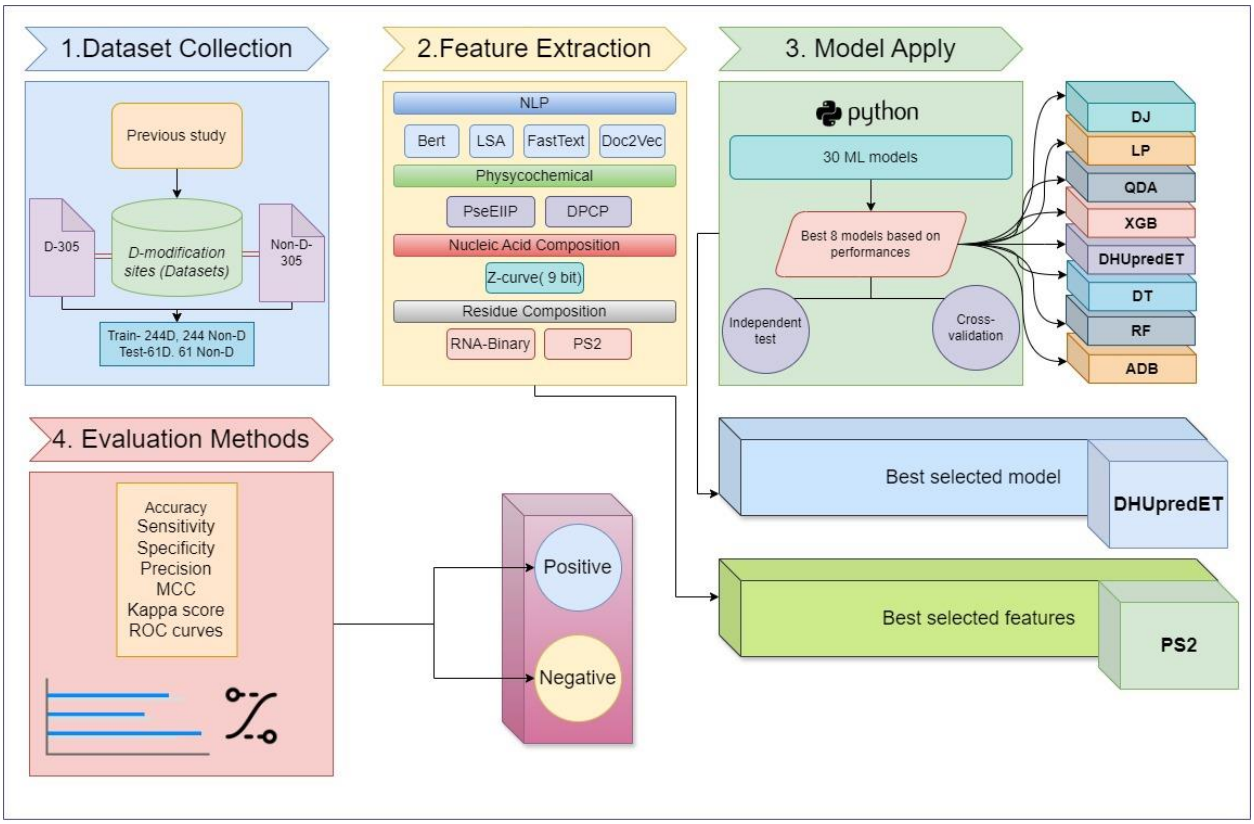

Figure 2. Workflows of the current study. Dataset collection from the previous studies, feature extraction approach of four kinds of descriptors with nine feature encoding methods, application of the various models, and selection process, the overall comparison of the models, and selecting the optimal feature and best-fit models for the study.

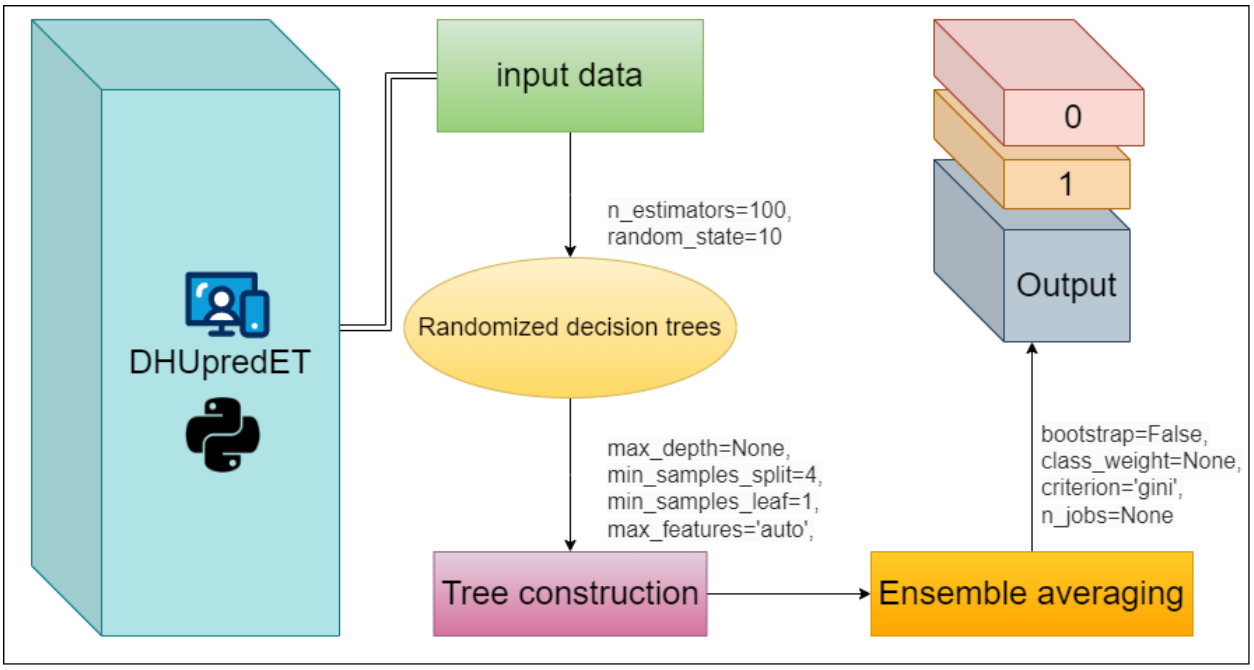

Figure 3. Overall strategies of the DHUpredET model.

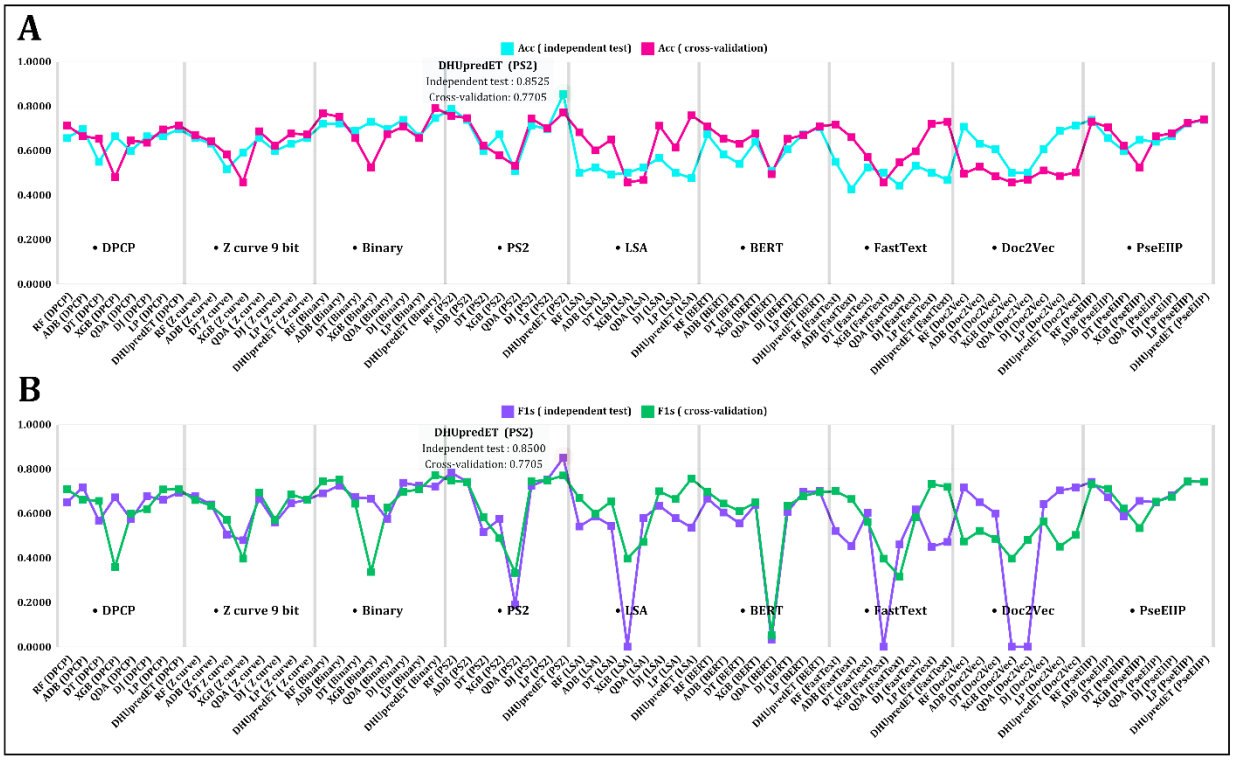

Figure 4. Overall comparison of Accuracy and F1 score between independent test and 5-fold-cross-validation approach, where **(A)** Accuracy (Acc) **(B)** F1 Score (F1s)

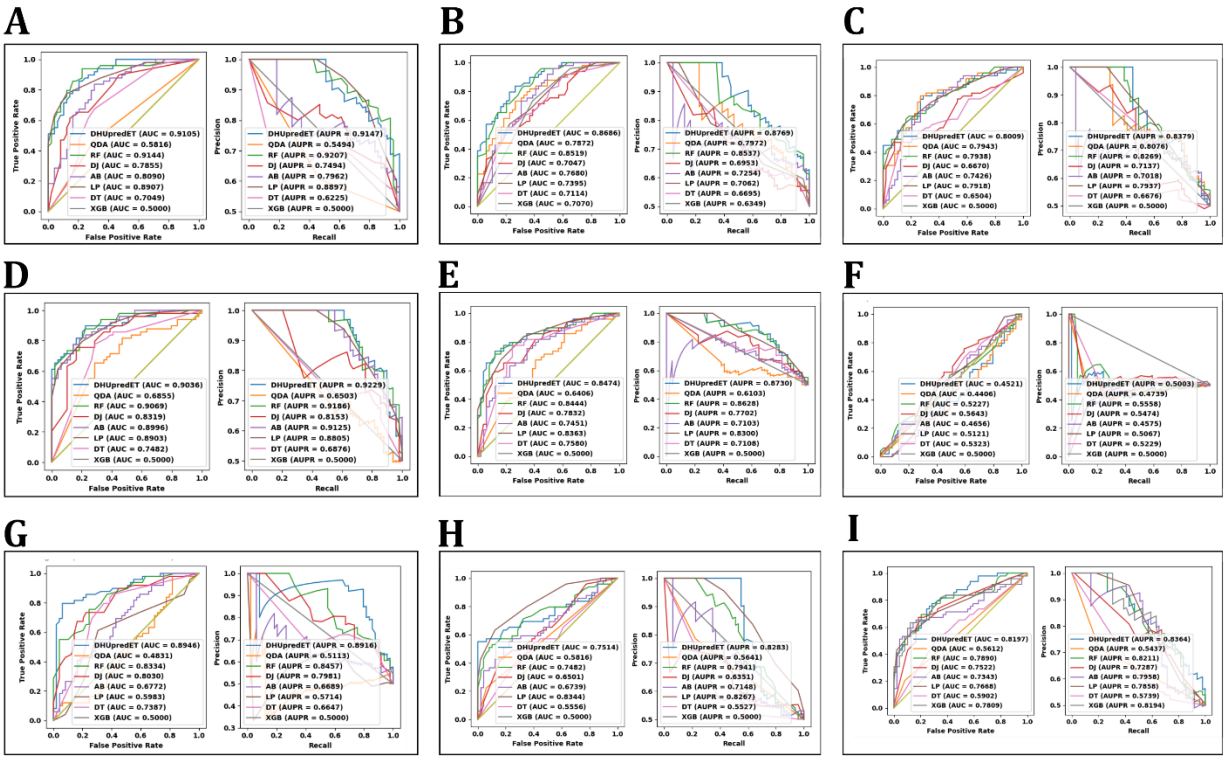

Figure 5. ROC curves and PR curves visualization based on independent test method with all feature extraction methods. **(A)** PS2, containing ROC and PR curves **(B)** PseEIIP, containing ROC and PR curves **(C)** DPCP, containing ROC and PR curves **(D)** Binary, containing ROC and PR curves **(E)** Z curve 9 bit, containing ROC and PR curves **(F)** LSA, containing ROC and PR curves **(G)** FastText, containing ROC and PR curves **(H)** Doc2Vec, containing ROC and PR curves **(I)** BERT, containing ROC and PR curves.

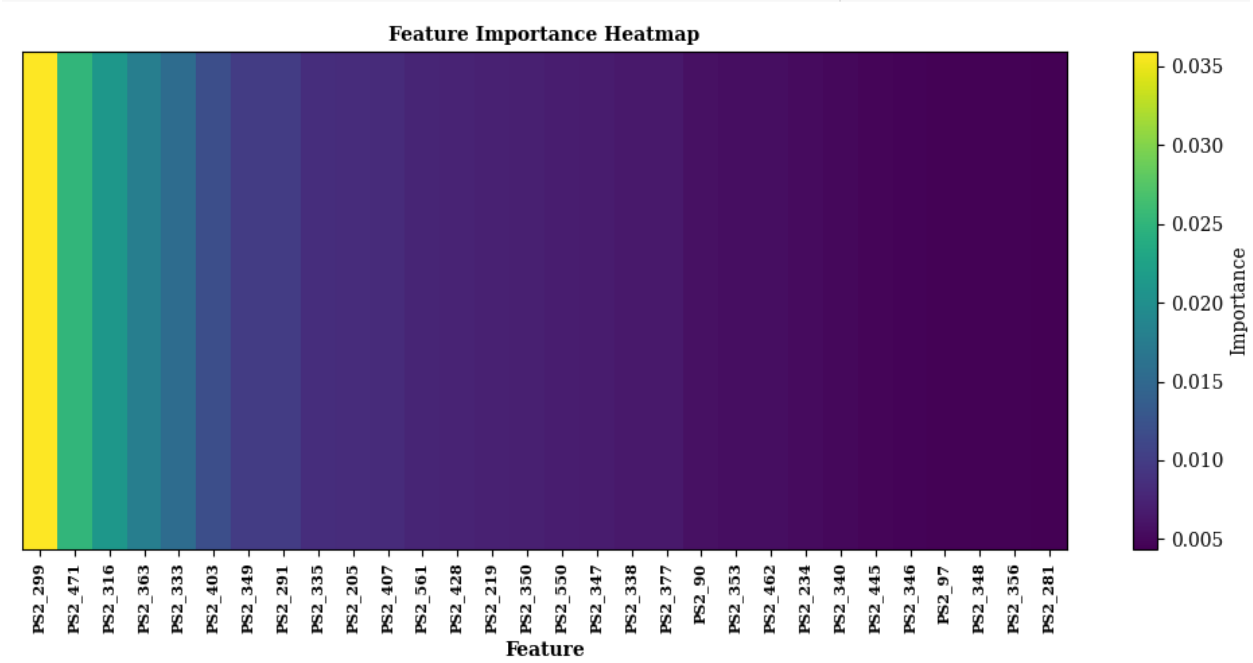

Figure 6. A heatmap of the DHUpredET model is based on PS2 features, and the study exhibits the top 30 important features.

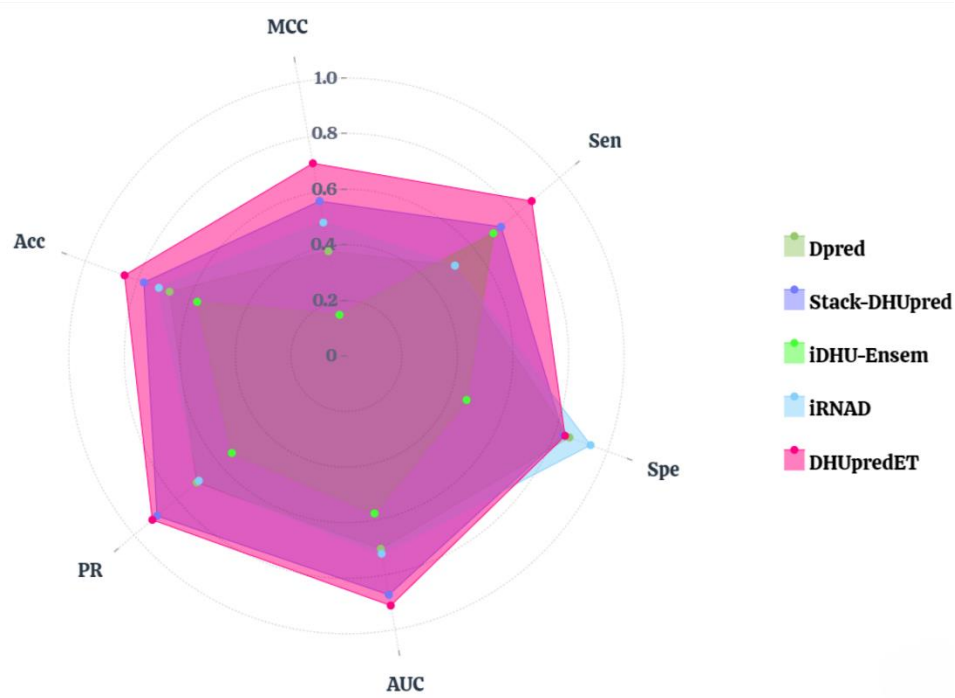

Figure 7. Comparison of state-of-the-art models with DHUpredET based on accuracy, MCC, AUC, PR, specificity, and sensitivity.

Click here to access/download
Supplementary Material
suppliments files (1).xlsx