scientific reports

OPEN

Prediction of exosomal miRNA-based biomarkers for liquid biopsy

Akanksha Arora & Gajendra Pal Singh Raghava

In this study, we investigated the properties of exosomal miRNAs to identify potential biomarkers for liquid biopsy. We collected 956 exosomal and 956 non-exosomal miRNA sequences from RNALocate and miRBase to develop predictive models. Our initial analysis reveals that specific nucleotides are preferred at certain positions in miRNAs associated with exosomes. We employed an alignment-based approach, artificial intelligence (AI) models, and ensemble methods for predicting exosomal miRNAs. For the alignment-based approach, we used a motif-based method with MERCI and a similarity-based method with BLAST, achieving high precision but low coverage of about 29%. The AI models, developed using machine learning, deep learning techniques, and pretrained language models, achieved a maximum AUC of 0.707 and an MCC of 0.268 on an independent dataset. Finally, our ensemble method, combining alignment-based and AI-based models, reached a maximum AUC of 0.73 and an MCC of 0.352 on an independent dataset. We have developed a web server, EmiRPred, to assist the scientific community in predicting and designing exosomal miRNAs and identifying associated motifs (https://webs.iiitd.edu.in/raghava/emirpred/).

Keywords Exosomal miRNA, Exosome, Alignment-based approach, Artificial intelligence based models, Liquid biopsy, Biomarkers

Liquid biopsy represents a groundbreaking advancement in diagnostics, enabling the detection and monitoring of various diseases through the analysis of biofluids[1–3]. This minimally invasive technique permits sampling throughout the disease course, circumventing the risks associated with tissue biopsies[4]. Commonly utilized molecules in liquid biopsies include circulating tumor DNA (ctDNA), cell-free RNA (cfRNA), and circulating tumor cells (CTCs). However, these molecules face limitations, such as the lack of surface markers and the low abundance of the desired cell-free nucleic acids[5]. Recently, exosomes have emerged as promising candidates for liquid biopsy applications due to their stability and the critical roles they play in biological processes. Exosomes, small extracellular vesicles released by cells, carry a cargo of proteins, lipids, DNA, mRNA, miRNA, and metabolites[6,7]. Exosomal biomarkers are more stable than cell-free macromolecules and can reflect dynamic changes in the tumor microenvironment[8]. They can be detected earlier in the disease process, making them a more promising tool for early diagnosis and monitoring.

Exosomal miRNAs, which are small non-coding RNA molecules approximately 19–22 nucleotides long, play critical roles in cellular communication and the regulation of gene expression[9,10]. The biogenetic pathway of exosomal miRNAs initiates in the nucleus, where DNA sequences are transcribed by RNA polymerase to form primary miRNAs (pri-miRNAs). These pri-miRNAs adopt hairpin structures of 70–100 nucleotides following initial processing[11]. Exportin 5 mediates the transport of hairpin pri-miRNAs to the cytoplasm, where Dicer facilitates further processing[12,13]. Upon maturation, these double-stranded miRNAs are converted into single-stranded miRNAs and subsequently sorted into exosomes (see Fig. 1). The incorporation of miRNAs into exosomes is a regulated process and is not random[14,15]. In the past, a number of miRNA-based biomarkers have been reported, such as miR-21, miR-1246, and miR-155 for non-small cell lung cancer, miR-17-5p, and miR-92a-3p for colorectal cancer, and miR-200b and miR-200c for ovarian cancer[16–18]. Beyond cancer, exosomal miRNAs can be used as potential biomarkers for other diseases, such as cardiovascular and neurological disorders, and personalized medicine.

However, predicting exosomal miRNA computationally poses several critical challenges. Firstly, miRNAs are short RNA molecules which limits the complexity and informativeness of sequence-based features. This short length leads to significant feature sparsity and sequence similarity, complicating accurate differentiation between exosomal and non-exosomal miRNAs. Secondly, experimentally validated datasets distinguishing exosomal miRNAs from their non-exosomal counterparts remain relatively small and imbalanced, presenting hurdles for developing robust prediction models. Thirdly, the molecular mechanisms guiding miRNA packaging into exosomes are not yet fully understood, making it challenging to computationally capture consistent

Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, Okhla Phase 3, A-302 (R&D Block), New Delhi 110020, India. email: raghava@iiitd.ac.in

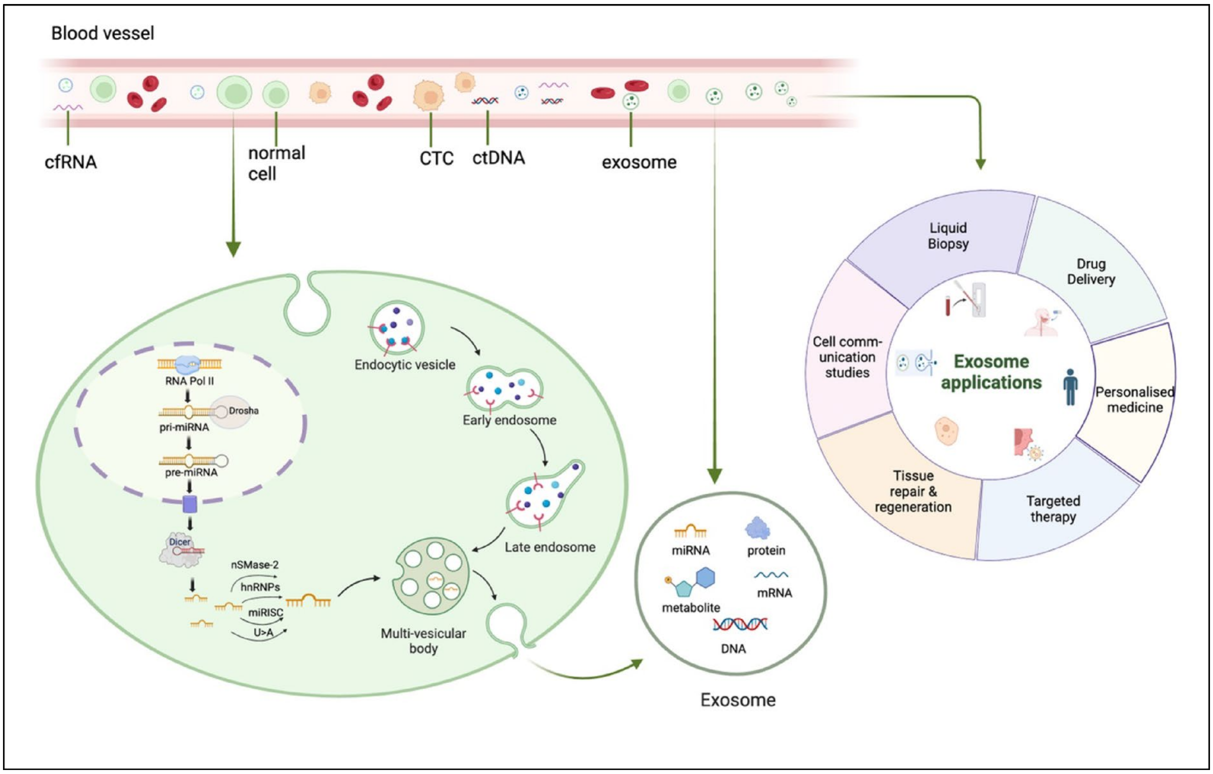

Fig. 1. Shows biomolecules in body fluid commonly used in liquid biopsy, as well as the mechanism of secretion of miRNA from cell to exosomes.

biological signals. Lastly, the current lack of well-characterized exosomal sequence signatures adds complexity to computational predictions, requiring advanced techniques to reliably identify subtle patterns and discriminative features.

In this study, a systematic attempt has been made to identify exosomal miRNAs and their characteristic features to help leverage their potential in diagnostics and therapeutics. Firstly, we created a dataset of experimentally validated exosomal and non-exosomal miRNA sequences. Then, we divided this dataset into training and independent datasets, where the training dataset contains 80% miRNA sequences, and the validation/independent dataset contains the remaining 20% miRNA sequences. We trained and developed our models on the training dataset using five-fold cross-validation. All hyperparameters are tuned on the training dataset, and only the final model is evaluated on the validation dataset. This is important to avoid overoptimization of the models. We have tried various alignment-based and AI-based techniques to develop a method to predict exosomal miRNA with high precision. Initially, we used alignment-based approaches involving motif search and sequence similarity search. These alignment-based approaches are only successful if the query sequence has a known motif or similarity with the annotated sequence but fail in the absence of a motif or similarity. In order to overcome these limitations, we developed artificial intelligence (AI) models. Our AI-based models include Machine Learning (ML), Deep Learning (DL), and Pretrained Language Models (PLM). Finally, we developed an ensemble method that combines the strength of alignment and AI-based models. This comprehensive approach aims to predict exosomal miRNAs accurately, paving the way for their broader application in biomedical research and clinical practice.

Results

In this study, we employed different techniques to predict exosomal miRNA, which can be categorized into three main categories: (i) Alignment-based approaches, (ii) AI-based models, and (iii) Ensemble methods. Alignment-based approaches encompass motif-search using MERCI software and similarity-search using BLAST. AI-based methods involve the development of ML, DL, and PLM-based models. The AI-based models used numerous features and their combinations, such as compositional, binary, structural features, and embeddings. To harness the full potential of alignment-based and AI-based techniques, we devised an ensemble method that integrates the strengths of both approaches. The comprehensive architecture of EmiRPred and the methodologies employed are illustrated in Fig. 2.

Alignment-based classification methods

Motif-search

In our study, we discovered motifs within exosomal miRNA using MERCI software applied with various parameters[19]. For instance, employing a Gap value of 0 led to the discovery of 11 motifs that covered 26

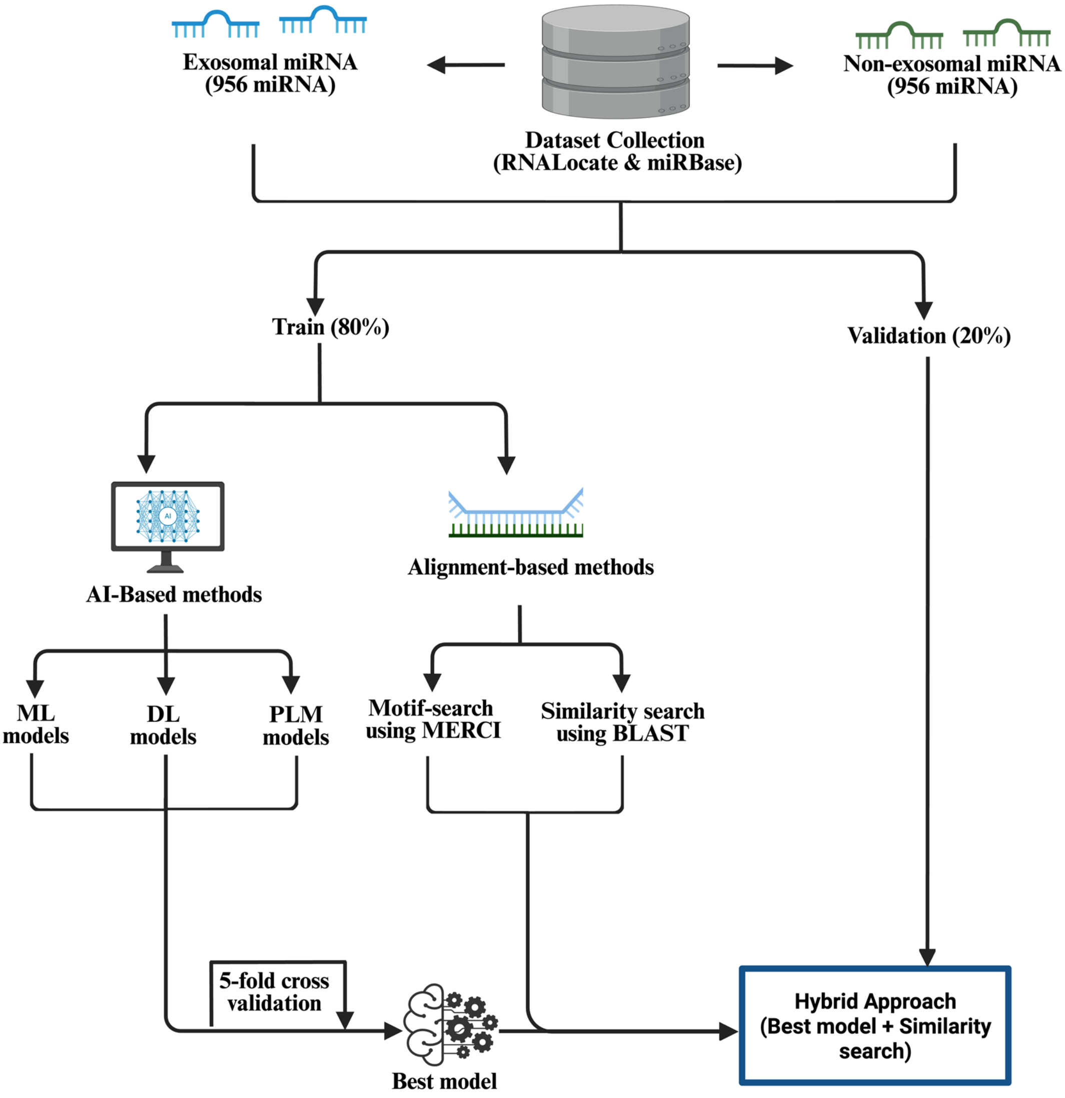

Fig. 2. The complete architecture of algorithm used in EmiRPred.

exosomal sequences in the training dataset and 8 sequences in the validation dataset. Similarly, applying a Gap value of 1 uncovered 11 motifs spanning 30 exosomal sequences in the training dataset and 10 sequences in the validation dataset. The detailed results of motif discovery using MERCI under different settings are presented in Supplementary Table S1, along with the sequences covered in the validation dataset.

Similarity-search using BLAST

In this paper, we applied blastn-short to perform a similarity search against a training dataset of exosomal and non-exosomal miRNA for the e-values ranging from 10^{-6} to 10^{6}[20]. Using this approach, we obtained an optimal performance at e-value 10^{-2}, with 233 correct hits and 85 incorrect hits for exosomal miRNA sequences in the training set; and 66 correct hits and 24 incorrect hits for exosomal miRNA sequences in the validation set. The e-values lesser than 10^{-2} did not show enough coverage for sequences, and the values greater than 10^{-2} showed a higher error rate. The full results for BLAST from e-values 10^{-6} to 10^{6} are shown in Table 1.

AI-based classification methods

ML models

Initially, we computed a wide range of sequence features for miRNA sequences; details are given in the Materials & Methods section. Next, we applied a variety of ML techniques for developing prediction models that includes DT, KNN, XGB, LR, SVC, RF, and ET. These results are given in Supplementary Table S2 and the hyperparameters for each ML model is given in Supplementary Table S3. The performances of ML-based models developed using different classes of features are as follows:

e-value	Training Dataset				Validation set			
	Correct hits	Incorrect hits	No hits	Total	Correct hits	Incorrect hits	No hits	Total
10^{6}	769	758	0	1527	185	198	0	383
10^{5}	769	758	0	1527	185	198	0	383
10^{4}	769	758	0	1527	185	198	0	383
10^{3}	769	758	0	1527	185	198	0	383
10^{2}	769	758	0	1527	185	198	0	383
10^{1}	769	758	0	1527	185	198	0	383
10^{0} (1)	626	559	342	1527	152	146	85	383
10^{-1}	300	143	1084	1527	79	44	260	383
10^{-2}	233	85	1209	1527	66	24	293	383
10^{-3}	175	53	1299	1527	49	15	319	383
10^{-4}	119	34	1374	1527	35	13	335	383
10^{-5}	99	24	1404	1527	30	11	342	383
10^{-6}	50	14	1463	1527	14	6	363	383

Table 1. Number of correct, incorrect, and total hits for exosomal miRNA sequences in training and validation set for e-values ranging from 10^{-6} to 10^{6}.

Composition-based features

We calculated various composition-based features using Nfeature, including nucleotide composition, composition of reverse complement, nucleotide repeat index, distance distribution of nucleotides, and pseudo composition[21]. Our RF-based model developed using the composition of reverse complement RNA sequence for k-mers 3 and 4 (RDK-3 & RDK-4) achieved a maximum AUC of 0.677. Our KNN model developed using TF-IDF achieved a maximum AUC of 0.656.

Binary profile

We computed binary profiles or one hot encoding features to represent the nucleotide sequences as binary vectors. We developed ML-based models using these binary profiles. An SVM-based model performed the best on these features with an AUC of 0.642, on the validation dataset.

Secondary structure-based features

We estimated secondary structure-based features for miRNA sequences using the RNAfold tool from the ViennaRNA package 2.0[22]. These features included minimum free energy, ensemble free energy, centroid free energy, centroid diversity, frequency of the minimum free energy (MFE) structure in the ensemble, and ensemble diversity. A logistic regression model, trained on these secondary structure-based features, achieved the highest AUC of 0.558 on the validation dataset.

PLM embeddings as features

To extract the embeddings from PLM models, we used two pre-trained models—DNABert and BERT-base-uncased and then fine-tuned them on our training dataset[23,24]. The embeddings were then extracted from the fine-tuned models for miRNA sequences which were then used to develop models using ML models. A random forest model achieved the maximum AUC 0.598 and 0.565 using embeddings of DNABERT and BERT-base-uncased, respectively.

Best features

We developed ML models using a combination of best features, including mononucleotide binary profiles, the composition of reverse complementary miRNA sequences (RDK-3, RDK-4), and TF-IDF. This resulted in a total of 382 features, which were normalized using the StandardScaler from the Scikit-learn package[25]. Our Extra Trees model, built using these features, achieved an AUC of 0.707 on the validation dataset. Additionally, a Mann–Whitney test was conducted on the 382 features to identify those that significantly differentiate exosomal from non-exosomal miRNA sequences. We found 75 features with p values less than 0.05, indicating a significant difference between the two classes. The results for the individual best features and the combined best features are provided in Tables 2 and 3, respectively. The results for the Mann–Whitney test performed on these features in given in Supplementary Table S4.

Feature importance

We selected the top 20 features from the best-performing set of features according to their feature importance in the model. It was observed that 17 out of the 20 features were significantly different (p value < 0.05) as calculated by the Mann–Whitney test in Section "Results". In binary profile features, it was observed that C at the 1st position (C_1), U at the 21st position (U_21), and G at the 15th position (G_15) are seen in more exosomal miRNA sequences than non-exosomal miRNA sequences. From reverse complement sequence compositional features (RDKs), it is seen that RDK_CAC is significantly increased in exosomal miRNA sequences than in non-exosomal sequences (p value < 0.0001). In TFIDF features from the range of (1,3) in reverse complement miRNA

		Training set					Validation set				
Model	Thr	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Binary features/One hot encoding—mononucleotides											
DT	0.51	54.09	54.49	54.28	0.553	0.086	58.38	53.03	55.61	0.571	0.114
RF	0.50	61.61	59.63	60.63	0.640	0.212	58.92	63.64	61.36	0.634	0.226
LR	0.50	54.99	54.35	54.68	0.578	0.093	52.43	60.10	56.40	0.560	0.126
XGB	0.50	57.20	57.12	57.16	0.614	0.143	56.76	52.02	54.31	0.571	0.088
KN	0.51	55.51	59.76	57.62	0.613	0.153	58.92	60.10	59.53	0.614	0.190
GNB	0.52	54.73	55.14	54.94	0.569	0.100	54.05	55.55	54.83	0.568	0.100
SVC	0.50	59.40	60.95	60.17	0.634	0.204	61.08	62.12	61.62	0.642	0.232
ET	0.50	56.42	59.50	57.95	0.628	0.159	58.38	65.66	62.14	0.633	0.241
Reverse Complement Sequence Composition (RDK3 + RDK 4)											
DT	0.53	54.86	49.21	52.06	0.529	0.041	58.92	52.02	55.35	0.559	0.110
RF	0.50	58.89	56.99	57.95	0.627	0.159	61.62	63.64	62.66	0.677	0.253
LR	0.50	57.07	56.60	56.84	0.596	0.137	56.22	60.10	58.23	0.602	0.163
XGB	0.50	58.50	56.73	57.62	0.601	0.152	58.38	58.59	58.49	0.620	0.170
KN	0.48	60.31	57.52	58.93	0.624	0.178	65.41	53.54	59.27	0.644	0.191
GNB	0.08	55.90	56.46	56.18	0.594	0.124	52.97	55.05	54.05	0.568	0.080
SVC	0.50	62.26	56.60	59.45	0.628	0.189	63.78	57.58	60.57	0.644	0.214
ET	0.50	59.66	59.37	59.52	0.639	0.190	62.70	59.09	60.84	0.662	0.218
Term Frequency- Inverse Document Frequency (TFIDF) on reverse complementary sequence (kmer range—1,3)											
DT	0.01	55.25	55.28	55.26	0.553	0.105	52.43	60.10	56.40	0.563	0.126
RF	0.51	59.53	57.78	58.67	0.622	0.173	63.78	56.57	60.05	0.632	0.204
LR	0.51	54.86	56.86	55.85	0.594	0.117	51.89	53.03	52.48	0.565	0.049
XGB	0.51	58.75	58.58	58.67	0.605	0.173	62.70	53.03	57.70	0.620	0.158
KN	0.41	56.94	59.37	58.14	0.602	0.163	71.35	58.59	64.75	0.656	0.301
GNB	0.46	56.81	56.73	56.77	0.588	0.135	52.43	51.52	51.96	0.564	0.039
SVC	0.51	57.85	62.01	59.91	0.635	0.199	61.08	61.11	61.10	0.634	0.222
ET	0.51	61.61	62.01	61.81	0.652	0.236	58.92	58.59	58.75	0.629	0.175

Table 2. AI-based methods results for the best performing features.

sequences, it is seen that "U" was found in abundance in exosomal sequences than non-exosomal, dinucleotides "AC" and "UC", and trinucleotides "GGA", "GGC", and "GUC" were also significantly increased in exosomal sequences. The top 20 most important features in distinguishing exosomal and non-exosomal miRNA sequences have been shown in Fig. 4. The results for the importance of all features are given in Supplementary Table S5.

DL models

Sequences

We applied the CNN algorithm to develop a prediction model using the binary profiles of miRNA sequences[26]. The model gave an AUC of 0.611 on the training dataset after five-fold cross-validation and 0.621 on an independent validation dataset. The detailed results for CNN model applied on miRNA sequences are given in Supplementary Table S2 and the hyperparameters for each model is given in Supplementary Table S3.

Structure images

We obtained the secondary structures of miRNA sequences using the RNAfold of the Vienna RNA package[22]. RNAfold predicted secondary structure of miRNA in form of images. These images were used to develop models CNN and ResNet 50[27,28]. Our CNN and ResNet50 achieved AUCs of 0.551 and 0.553, respectively. The results for these models applied on RNA secondary structure images are given in Supplementary Table S2.

PLM models

We used pre-trained PLM models like DNABERT and BERT-base-uncased and fine-tuned them on our data[23,24]. DNABERT was able to differentiate between the exosomal and non-exosomal sequences with an AUC of 0.535 and BERT-base-uncased was able to differential them with an AUC of 0.608. The detailed results for finetuned PLM models are given in Supplementary Table S2 and their hyperparameters are given in Supplementary Table S3.

Hybrid classification method

To develop a model that is able to differentiate between exosomal and non-exosomal miRNA sequence classes with high accuracy, we developed a hybrid model that combined the predictive power of our best performing ML model on the best identified features with motif-search, and similarity search algorithm. The best performing

Model	Thr	Training set					Validation set				
		Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Best features combined—One hot encoding (mononucleotides) + RDK-3 + RDK-4 + Reverse complement TFIDF (1,3)											
DT	0.50	56.16	53.83	55.00	0.550	0.100	56.76	54.55	55.61	0.557	0.113
RF	0.51	57.59	61.74	59.65	0.648	0.193	60.54	59.09	59.79	0.637	0.196
LR	0.50	57.46	56.99	57.23	0.602	0.144	60.54	58.59	59.53	0.620	0.191
XGB	0.55	55.12	63.98	59.52	0.636	0.192	65.41	56.57	60.84	0.640	0.220
KN	0.50	57.98	59.63	58.80	0.621	0.176	63.24	58.08	60.57	0.656	0.213
GNB	0.98	55.25	53.83	54.55	0.558	0.091	38.38	70.71	55.09	0.586	0.096
SVC	0.51	55.77	62.80	59.25	0.644	0.186	64.32	64.65	64.49	0.678	0.290
ET	0.50	62.39	62.01	62.20	0.672	0.244	63.24	63.64	63.45	0.707	0.269
Hybrid model: Best features combined with motif-search and similarity search											
DT	0.51	57.33	53.83	55.59	0.614	0.112	57.84	54.04	55.87	0.622	0.119
RF	0.55	61.22	65.83	63.51	0.694	0.271	64.86	62.63	63.71	0.685	0.275
LR	0.59	60.96	60.16	60.56	0.663	0.211	64.32	61.62	62.92	0.687	0.259
XGB	0.59	61.74	61.21	61.48	0.680	0.230	69.19	53.54	61.10	0.694	0.230
KN	0.54	69.00	54.35	61.74	0.673	0.236	70.81	54.04	62.14	0.690	0.252
GNB	0.93	60.44	50.00	55.26	0.618	0.105	49.73	67.68	59.01	0.646	0.177
SVC	0.54	60.44	62.27	61.35	0.691	0.227	67.03	65.15	66.06	0.711	0.322
ET	0.52	65.63	62.93	64.29	0.703	0.286	67.57	67.68	67.62	0.730	0.352

Table 3. Results for (a) best performing features combined (AI-based methods), and (b) hybrid model: best-performing features (AI-based methods) combined with motif-search and similarity search (Alignment-based methods).

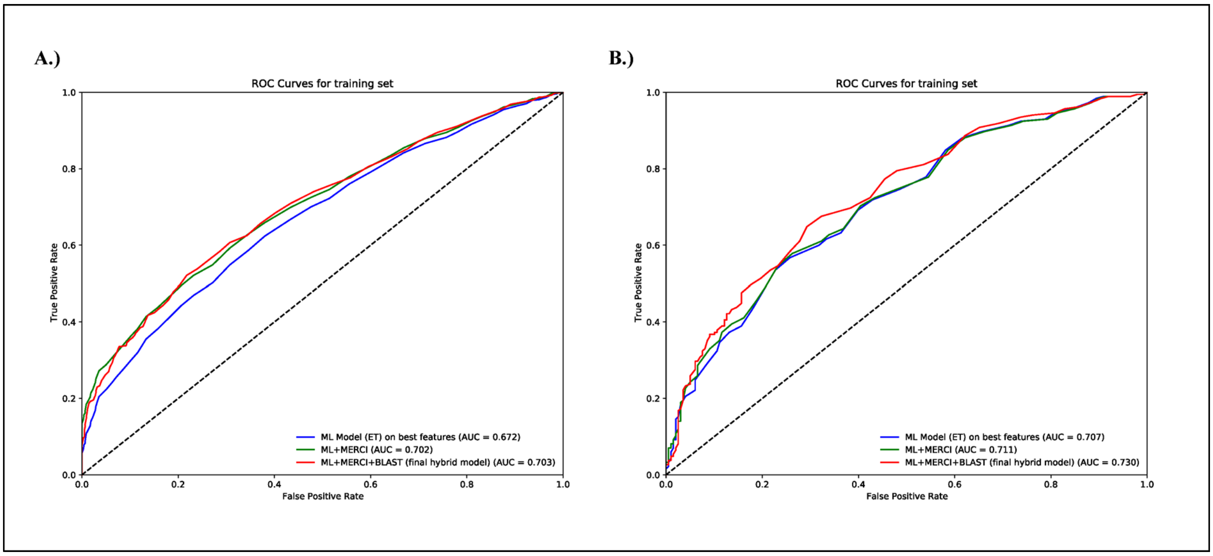

Fig. 3. AUROC graphs for the hybrid model (**A**) Training set and (**B**) Independent Validation set.

prediction model is ET model with an AUC of 0.707 on independent validation dataset, which was improved to 0.712 when motif-search algorithm results were added using the hybrid approach scoring system described in "Methods". The AUC further improved to 0.73 when we added the similarity-search algorithm using BLAST on e-value = 10^{-2} to our model. Along with the AUC of 0.73, this final model also showed an accuracy of 67.62% and an MCC of 0.352 on an independent validation set. The resulting metrics for both training and independent validation set for this hybrid model are given in Table 3. The Area Under the Receiver Operating Curves (AUROC) for the training and validation set for the hybrid model are shown in Figs. 3 and 4.

Comparison with existing methods

Presently, there are a few existing methods that predict the subcellular location of miRNA, with "exosome" being one of the locations. However, there is no tool that solely focuses on exosomal miRNA and predicts it with high accuracy. The tools that take miRNA sequences as input include miRNALoc, and EL-RMLocNet[29,30]. We wanted

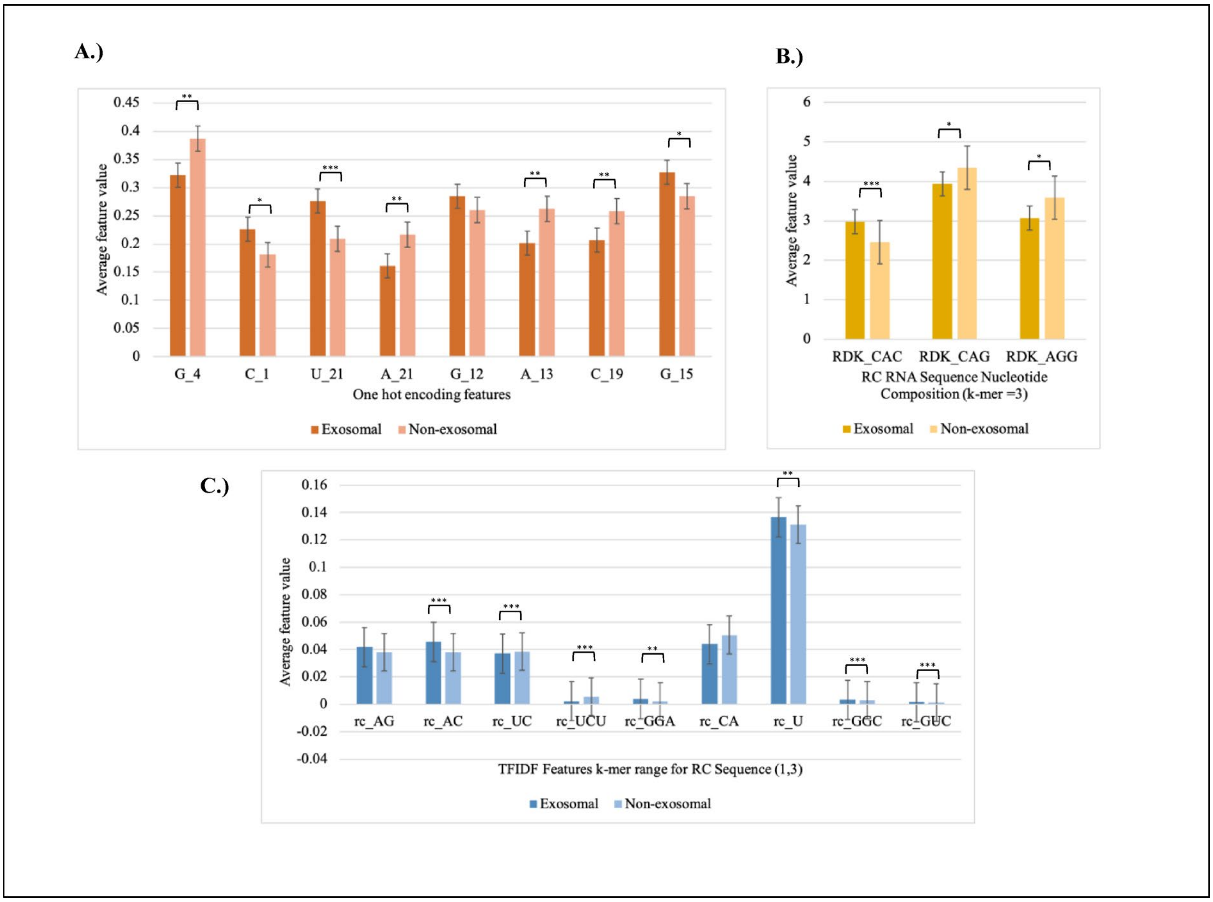

Fig. 4. The 20 most important features for the classification of exosomal and non-exosomal miRNA.

to compare the performances of these existing tools with our prediction tool. The tool miRNALoc reports the highest accuracy of 50% for predicting the subcellular localization of miRNA on an independent validation set, whereas EL-RMLocNet reports the highest AUC of 0.629 for predicting human miRNA on a benchmark dataset. To compare the results, we fed our independent validation set into the prediction servers, where we got an AUC of 0.494, an accuracy of 48.04%, and an MCC of − 0.028 from the miRNALoc web server as compared to EmiRPred which gives an AUC of 0.73, accuracy of 67.62%, and MCC of 0.352.

However, we were not able to get the predictions from EL-RMLocNet as it only takes one sequence at a time for the prediction, making it unfeasible to predict the validation set comprising 383 sequences. To overcome this issue, we instead used data from EL-RMLocNet and predicted exosomal sequences using our web server EmiRPred. We used the human miRNA data from EL-RMLocNet as our model is built specifically for human miRNA for lengths 15–26 nucleotides. The dataset provided on this server only had 305 human miRNA sequences (158 exosomal and 147 non-exosomal). However, about 95 sequences were longer than 26 nucleotides ranging from lengths 61 to 149 nucleotides with most of being non-exosomal (~ 70%) which shows that the authors might have taken pre-miRNAs in addition to the mature miRNA. After removing these sequences, about 210 miRNA sequences were left. Out of these 210 sequences, about 50 sequences contain “T” instead of “U”, and all 50 sequences were non-exosomal. After removing these, we went ahead with the remaining 160 mature sequences (130 exosomal and 30 non-exosomal) which do not contain “T” and fed it into our web server which resulted in an AUC of 0.891 and MCC of 0.639. These comparisons demonstrate that EmiRPred is better than the previous methods for predicting exosomal miRNA sequences.

Webserver and standalone software

One of the objectives of this study is to assist the scientific community working in the field of diagnostics or prognostics, specifically on exosomal miRNA-based biomarkers. This web server includes four modules: Predict, Design, Motif-scan, and BLAST-search. The “Predict” module accepts query sequences as input and predicts exosomal miRNA. Our “Design” module allows users to generate all possible mutant miRNA sequences and predict exosomal mutant miRNA. The "Motif-scan" module enables users to identify or scan known exosomal motifs in the query miRNA sequence. The "BLAST-search" module allows users to perform similarity searches against a database of known exosomal and non-exosomal miRNA sequences. This web server is compatible with smartphones, PCs, iMacs, and tablets. In addition to the web server, we have also created a Python package, a standalone tool, and a GitHub repository for this tool, available at the following links: https://webs.iiitd.edu.in/raghava/emirpred/ and https://github.com/raghavagps/emirpred.

Discussions

Exosomes are small vesicles enclosed by a lipid membrane bilayer, secreted by most cells in the body. They carry various molecules, including proteins, lipids, metabolites, and RNA sequences[6]. In diagnostics, exosomes provide a non-invasive means to detect biomarkers for various conditions, such as cancer, neurodegenerative diseases, and cardiovascular disorders. In recent years a number of exosome-based biomarkers have been identified; for example, exosomal PD-L1 has been highlighted as a biomarker in non-small cell lung cancer[31]. Additionally, exosomal circRNAs have been implicated in colorectal cancer, offering potential as both diagnostic markers and therapeutic targets[32]. Exosomal miR-21 is frequently elevated in breast cancer, while exosomal miR-1246 has been linked to chemotherapy resistance in pancreatic cancer[33,34]. In the realm of therapy, exosomes have been utilized to deliver CRISPR/Cas9 components for gene editing, showing efficacy in correcting genetic mutations in preclinical models of Duchenne muscular dystrophy[35]. Moreover, exosomes engineered to carry small interfering RNA (siRNA) targeting oncogenes like KRAS have demonstrated significant tumor suppression in pancreatic cancer models[36]. These advancements underscore the versatility of exosomes as both biomarkers and therapeutic delivery vehicles, paving the way for more personalized and effective medical interventions.

In this study, we focused on exosomal miRNA, the most abundant molecules found in exosomes, which hold immense potential as diagnostic and prognostic biomarkers. Our objective was to predict exosomal miRNA using various techniques, including motif identification, similarity search, ML, and DL. We identified several motifs and features that differentiate exosomal from non-exosomal miRNA. Initially, we discovered motifs or small patterns frequently found in exosomal miRNA. The most recurring motif was "GgapAAGCAC," which appeared in 4 exosomal and 1 non-exosomal sequences in the validation set. Only a limited number of sequences contained these motifs, covering just 5.2% of miRNA sequences in the validation dataset. This indicates that motif identification alone is insufficient for predicting exosomal miRNA. Next, we employed a similarity search technique commonly used for sequence annotation. As mentioned in the results section, our similarity-based technique demonstrated a high probability of correct prediction but poor coverage. This suggests that alignment-based techniques (like motif and similarity search) are not adequate to predict all exosomal miRNA sequences.

In the last three decades, AI-based techniques have been heavily used to develop classification models in the field of computational biology, particularly in sequence classification. Thus, we utilized AI-based techniques to develop prediction models for this study. One of the advantages of AI-based models over alignment-based models is coverage of the dataset. Here, we systematically applied ML and DL techniques to develop prediction models. We also used all possible features, including compositional features, binary profiles, and embeddings, to develop models. In addition, we also used PLMs in this study to predict exosomal miRNA. Despite our best efforts, we achieved a maximum AUC of 0.707 with an MCC of 0.269 using AI-based models. Therefore, we developed an ensemble/hybrid method that combined the two techniques: Alignment-based models and AI-based models. Our ensemble method utilising these approaches achieved AUC of 0.73 and an MCC of 0.352. The model also achieved a balanced sensitivity and specificity of 67.57% and 67.68%, respectively. This corresponds to a false negative rate (FNR) of approximately 32.43% and a false positive rate (FPR) of approximately 32.32%. These balanced metrics suggest that our model performs comparably in identifying both exosomal and non-exosomal miRNAs, and is not overly biased toward one class.

Although our model achieved moderate predictive performance, it should be noted that predicting exosomal miRNA has its own challenges. Exosomes vary significantly in molecular composition based on their cellular origin, physiological condition, and environment. In addition, the mechanism of sorting of miRNA into exosomes is not fully understood. We have made an attempt to identify important features and motifs that can distinguish between exosomal and non-exosomal miRNA to advance the knowledge of exosomal miRNA sorting mechanism. Our model can potentially aid in the discovery of novel biomarkers for early disease detection. Furthermore, this tool can also support therapeutic applications, such as the design of synthetic exosomal miRNAs for targeted gene regulation or the refinement of exosome-based drug delivery strategies using the design module on our webserver.

It is important to compare newly developed methods with existing ones. To our knowledge, no method has been specifically developed for predicting exosomal miRNA sequences. However, we found miRNA subcellular localization methods like EL-RMLocNet and miRNALoc that can predict the subcellular localization of miRNA sequences. These subcellular localization methods predict multiple locations of miRNA in cells, including exosomal miRNA. Thus, we evaluated the performance of these methods on our independent validation dataset. As described in the results section, our method performed better than existing methods for exosomal miRNA. The limitation of this study is that, although the data is collected from experimentally validated studies, it is purely bioinformatics in nature. The features and motifs identified in this study have been previously reported to be associated with exosomal miRNA; however, in-depth studies with experimental validation are required.

Methods

Data collection and preprocessing

In this study, datasets were created using the data extracted from the databases: RNALocate and miRbase[37,38]. We collected the experimentally validated exosomal miRNA sequences from RNALocate. It contained about 1195 unique exosomal miRNA sequences for humans, with 956 mature miRNA sequences. Similarly, we collected experimentally validated miRNA sequences found in humans from RNALocate that were not detected in exosomes. Additionally, we sourced human miRNAs from miRBase and excluded those present in exosomes. This resulted in 1694 unique non-exosomal miRNA sequences. We randomly selected 956 mature miRNAs from the non-exosomal miRNA to balance the dataset. Our final dataset contains 956 exosomal and 956 non-exosomal miRNAs. The sequence lengths ranged from 16 to 26 nucleotides, with most sequences being between 21 and 24 nucleotides long.

Alignment-based approaches

Motif-search

We used the MERCI (Motif Emerging with Classes Identification) tool to detect motifs found in exosomal miRNA sequences[19]. We identified motifs using the training dataset and then searched those motifs in the independent validation set. MERCI software has the option to select motifs that are exclusively present in the positive class (exosomal) and not present in the negative class (non-exosomal) from the training dataset. Additionally, this software offers a variety of options, such as the frequency of motifs and motifs with or without gaps.

Similarity search

In this study, we used blastn-short to annotate miRNA sequences based on their similarity to exosomal or non-exosomal miRNA sequences[20]. First, we built a database using the "makeblastdb" command for miRNA sequences present in the training dataset. To compute results for the training dataset, we removed self-hits and considered the top hit after removing self-hits. For the independent dataset, we considered the first hit to calculate results at various e-values. This method has been used in the previous studies[39,40].

AI-based classification methods

Feature generation

To develop a prediction model to predict exosomal and non-exosomal miRNA, we applied an array of techniques to extract meaningful features from miRNA sequences. They are discussed here:

Composition-based features

We utilized Nfeature to compute a wide range of composition-based features in miRNA sequences, including nucleotide composition, reverse complementary compositions, and auto-correlation[21]. Additionally, we computed features using Term Frequency-Inverse Document Frequency (TF-IDF), a statistical measure that evaluates the importance of a term in a document relative to a collection of documents. In our study, "terms" refer to k-mers, which are sequences of length k nucleotides, and a "document" refers to a sequence[41]. Term Frequency (TF) is the number of times a k-mer appears in a sequence, normalized by the total number of k-mers in the sequence. Inverse Document Frequency (IDF) is the logarithmically scaled inverse fraction of the sequences that contain the k-mer across the entire dataset. It measures how unique or rare a k-mer is within the whole dataset. By combining TF and IDF, TF-IDF assigns higher weights to k-mers that are frequent in a specific sequence but rare across the dataset, indicating that these k-mers are more informative and characteristic of the sequence's content. The formulas for TF, IDF, and TF-IDF are explained in Eqs. 1, 2, and 3.

$$TF\left(k,s\right)=\frac{Number\ of\ times\ k\text{-}mer\ k\ appears\ in\ sequence\text{``}s\text{''}}{Total\ number\ of\ k\text{-}mers\ in\ sequence\text{``}s\text{''}} \tag{1}$$

$$IDF\left(k,D\right)=\log\frac{Total\ number\ of\ sequences\ in\ dataset\text{``}D\text{''}}{Number\ of\ sequences\ containing\ k\text{-}mer\text{``}k\text{''}} \tag{2}$$

$$TF\text{-}IDF\left(k,s,D\right)=TF\left(k,s\right)\times IDF(k,D) \tag{3}$$

In Eqs. (1), (2), and (3), "k" represents a k-mer, "s" represents a sequence, and "D" represents the dataset of sequences.

The TFIDF features were computed for both the given sequences and reverse complementary sequences for kmer ranges from (1,1), (1,2),…(1,7). The best-performing TFIDF features were then also computed using different types of weighting—Term Frequency Collection (TFC) Weighting, Logarithmic Term Count (LTC) Weighting, and Entropy Weighting.

Binary features

We computed binary profiles, or one-hot encoding features, for sequences to represent nucleotide sequences as binary vectors. To handle variable-length sequences, we use a technique called padding to standardize the length of the sequences. Since the maximum length of miRNA sequences is 26 nucleotides, each sequence is padded to a length of 26 by adding the dummy nucleotide 'X'. For example, a sequence "AUTGTCGGGCUCUCCUAAU CU" of length 21 becomes "AUTGTCGGGCUCUCCUAAUCUXXXXX" of length 26. After padding, one-hot encoding features are computed by converting these sequences into binary profiles.

Structure-based features

The structure-based features of miRNA sequences were computed using the RNAfold tool of Vienna RNA package 2.0, which predicts and identifies the RNA secondary structure based on the lowest possible free energy. It provides feature values for minimum free energy, ensemble free energy, centroid free energy, centroid diversity, the frequency of MFE structure in the ensemble, and ensemble diversity for each sequence in the dataset. In RNA folding, an ensemble typically refers to a collection of possible secondary structures that the RNA molecule can adopt[22]. The structures were predicted for each miRNA sequence in both exosomal and non-exosomal categories. These predicted structures were saved in .jpg image formats to develop further deep-learning classification models based on these structure images.

PLM embeddings as features

In this study, we generated embeddings using two types of BERT PLMs. BERT, developed by Google, is a transformer-based DL model. The following pre-trained models were used in this study.

(a) BERT-Base Uncased: Uncased means that the text used during pre-training and fine-tuning is converted to lowercase, and no distinction is made between uppercase and lowercase letters. For example, "Sequence" and "sequence" would be treated as the same word. We extracted embeddings from this model after fine-tuning it on our training dataset[24].

(b) DNABERT: DNABERT is a variant of BERT designed explicitly for processing DNA sequences. DNA-BERT is pre-trained on a large corpus of DNA sequences using self-supervised learning techniques. To use this model, we pre-processed our data and replaced "U" with "T" to suit the DNABERT model. We extracted the embeddings from this model by first fine-tuning the model on our sequence dataset[23].

Prediction models

ML models

We have used several ML algorithms to differentiate between exosomal and non-exosomal miRNA sequences. These algorithms involve K-Nearest Neighbours (KNN), Gaussian Naïve Bayes (GNB), Decision Tree (DT), Logistic Regression (LR), Extreme Gradient Boosting (XGB), Support Vector classifier (SVC), Extra Tree Classifier (ET) and Random Forest (RF)[42–49]. We have built prediction models using these ML algorithms on different sets of features defined in section "Discussions". We used grid search in sci-kit learn to perform hyperparameter tuning to optimize parameters of these algorithms. Grid search is a hyperparameter tuning technique that tries every possible combination of the given parameters and gives the best performing combination[50,51].

DL models

We applied a DL classifier Convolutional Neural Networks (CNN) to classify exosomal and non-exosomal miRNA sequences. For the sequence classification task, we converted sequences to binary profiles as described in section "Alignment-based classification methods". These binary profile of miRNA is classified using CNN algorithm. Additionally, two DL algorithms, CNN and ResNet (Residual Network), were applied to the miRNA structure images generated as described in section "Structure images". ResNet utilizes shortcut connections to facilitate residual learning, enabling deep networks to train effectively. We used the ResNet50 variant, a 50-layer version of the ResNet architecture, as it balances complexity and efficiency[26–28,52,53]. Given the high computational cost of deep learning models, we employed random search to efficiently tune hyperparameters and optimize model performance. Random search is a hyperparameter tuning technique where hyperparameter values are randomly sampled from specified distributions, rather than exhaustively tested like in grid search. It takes lesser time than grid search as instead of trying every possible combination, it tries a fixed number of random combinations[54].

PLM models

PLMs excel in text classification due to their deep contextual understanding. They are pre-trained on extensive corpora and can be fine-tuned for specific tasks like sequence classification in our study[55]. They adapt well to various domains. The pre-training in PLMs provides efficient feature extraction, enabling accurate classifications even with smaller datasets. We have used two types of PLM models in our study to classify exosomal and non-exosomal miRNA sequences. We have used BERT-based uncased and DNABERT, as mentioned in section "Hybrid classification method", and fine-tuned them according to our sequence length and dataset composition. These fine-tuned models using our specific dataset and adapted to the sequence length of the input. In addition, we performed hyperparameter tuning using random search. The resulting fine-tuned models were then employed to classify sequences as exosomal or non-exosomal.

Cross-validation and performance metrics

The entire dataset, consisting of 1912 sequences, was divided into an 80:20 ratio, with 80% used for training and 20% for validation. A five-fold cross-validation technique was employed on 80% of the training data to evaluate the ML and DL models, keeping the remaining 20% unseen for the models. In five-fold cross-validation, the training data is split into five parts: four parts for training and one part for internal validation. This process is repeated five times, ensuring each fold serves as the test set once. The models were assessed using both threshold-dependent and threshold-independent metrics. Evaluation metrics included sensitivity, specificity, Matthews correlation coefficient (MCC), accuracy, and Area Under the Receiver Operating Characteristics (AUROC). AUROC is threshold-independent, while the other metrics depend on the threshold. Specificity, sensitivity, and MCC were optimized for the best threshold values. These metrics have been used in previous studies to measure the performance of models[56–59]

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FP} + \text{TN}} \times 100 \tag{4}$$

$$\text{Specificity} \ = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \tag{5}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \times 100 \tag{6}$$

$$\mathrm{MCC}=\frac{(\mathrm{TN}\times\mathrm{TP})-(\mathrm{FN}\times\mathrm{FP})}{\sqrt{(\mathrm{FP}+\mathrm{TP})\,(\mathrm{FN}+\mathrm{TP})\,(\mathrm{FP}+\mathrm{TN})\,(\mathrm{FN}+\mathrm{TN})}} \tag{7}$$

where TP, FP, TN, and FN are true positive, false positive, true negative, and false negative, respectively.

Ensemble method

We attempted to improve the prediction of our best-performing prediction model developed on the best set of features by using a combination of various findings. This approach applies a weighted scoring method that combines three techniques: (i) motif-based approach, (ii) similarity search using BLAST, and (iii) ML/DL prediction based methods. In this approach, we assign a score of + 0.5 to an miRNA sequence if it has an exosomal motif and 0 if no exosomal motif is found. We further add + 0.5 if the sequence is found to be similar to an exosomal miRNA sequence using the BLAST algorithm. These scores are then combined with best performing ML/DL model prediction scores, obtained via the predict_proba() function in scikit-learn, which gives the probability of a sequence belonging to a specific class instead of a binary outcome[25]. Together, the motif score, similarity search score, and best ML/DL model score provide a combined score for each sequence, ranging from 0 to 2 as shown in Eqs. (8) and (9). By analyzing these overall scores, the sequences were classified as either exosomal or non-exosomal. Several studies have previously utilized this hybrid/ensemble approach[7,60].

$$E=\begin{cases} E+0.5 & If\, exosomal\, motif\, is\, present\\ E & if\, no\, exosomal\, motif\, present\end{cases} \tag{8}$$

Here, E = Prediction probability score obtained from best performing ML/DL model and E' = Score obtained after adding scores from the motif-based approach

$$E''=\begin{cases} E'+0.5 & If\, BLAST\, hit\, is\, against\, an\, exosomal\, sequence\\ E' & If\, BLAST\, hit\, is\, not\, against\, an\, exosomal\, sequence\, or\, no\, hit\end{cases} \tag{9}$$

Here, E'' = Final score obtained from the best performing ML/DL model, motif-based approach, and BLAST-based approach ranging from 0 to 2.

Data availability

All the datasets generated in this study are available at https://webs.iiitd.edu.in/raghava/emirpred/dataset.php and codes are available at https://github.com/raghavagps/emirpred.

Received: 27 January 2025; Accepted: 11 August 2025
Published online: 25 August 2025

References

1. Shegekar, T., Vodithala, S. & Juganavar, A. The emerging role of liquid biopsies in revolutionising cancer diagnosis and therapy. *Cureus* https://doi.org/10.7759/cureus.43650 (2023).
2. Arora, A. et al. SalivaDB-a comprehensive database for salivary biomarkers in humans. *Database (Oxford)* **2023** (2023).
3. Swarup, N. et al. Multi-faceted attributes of salivary cell-free DNA as liquid biopsy biomarkers for gastric cancer detection. *Biomark. Res.* **11**, 90 (2023).
4. Armakolas, A., Kotsari, M. & Koskinas, J. Liquid biopsies, novel approaches and future directions. *Cancers (Basel)* **15**, 1579 (2023).
5. Yadav, R., Singh, A., Kushwaha, S. & Chauhan, D. Emerging role of exosomes as a liquid biopsy tool for diagnosis, prognosis & monitoring treatment response of communicable & non-communicable diseases. *Indian J. Med. Res.* **159**, 163 (2024).
6. Kalluri, R. & LeBleu, V. S. The biology, function, and biomedical applications of exosomes. *Science (1979)* **367** (2020).
7. Arora, A. et al. A random forest model for predicting exosomal proteins using evolutionary information and motifs. *Proteomics* **24**, e2300231 (2024).
8. Hu, C. et al. Potentiality of exosomal proteins as novel cancer biomarkers for liquid biopsy. *Front. Immunol.* **13** (2022).
9. Zheng, D. et al. The role of exosomes and exosomal microRNA in cardiovascular disease. *Front. Cell Dev. Biol.* **8** (2021).
10. Salehi, M. et al. Exosomal microRNAs in regulation of tumor cells resistance to apoptosis. *Biochem Biophys Rep* **37**, 101644 (2024).
11. Fan, J. et al. Generation of small RNA-modulated exosome mimetics for bone regeneration. *ACS Nano* **14**, 11973–11984 (2020).
12. Ahmed, F., Ansari, H. R. & Raghava, G. P. Prediction of guide strand of microRNAs from its sequence and secondary structure. *BMC Bioinform.* **10**, 105 (2009).
13. Ahmed, F., Kaundal, R. & Raghava, G. P. PHDcleav: a SVM based method for predicting human Dicer cleavage sites using sequence and secondary structure of miRNA precursors. *BMC Bioinform.* **14**, S9 (2013).
14. Li, C. et al. The role of Exosomal miRNAs in cancer. *J. Transl. Med.* **20**, 6 (2022).
15. Wozniak, A. L. et al. The RNA binding protein FMR1 controls selective exosomal miRNA cargo loading during inflammation. *J. Cell Biol.* **219** (2020).
16. Wang, M. et al. Emerging function and clinical values of exosomal microRNAs in cancer. *Mol. Ther. Nucleic Acids* **16**, 791–804 (2019).
17. Bakhsh, T. et al. Molecular detection of exosomal miRNAs of blood serum for prognosis of colorectal cancer. *Sci. Rep.* **14**, 8902 (2024).
18. Liang, X., Wu, Q., Wang, Y. & Li, S. MicroRNAs as early diagnostic biomarkers for non-small cell lung cancer (Review). *Oncol. Rep.* **49**, 8 (2022).
19. Vens, C., Rosso, M.-N. & Danchin, E. G. J. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics* **27**, 1231–1238 (2011).
20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
21. Mathur, M. et al. Nfeature: A platform for computing features of nucleotide sequences. *bioRxiv* 2021.12.14.472723 (2021) https://doi.org/10.1101/2021.12.14.472723.
22. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
23. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

24. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).
25. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python (2012).
26. Kalchbrenner, N., Grefenstette, E. & Blunsom, P. A Convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 655–665 (Association for Computational Linguistics, 2014). https://doi.org/10.3115/v1/P14-1062.
27. Chauhan, R., Ghanshala, K. K. & Joshi, R. C. Convolutional neural network (CNN) for image detection and recognition. In *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)* 278–282 (IEEE, 2018). https://doi.org/10.1109/ICSCCC.2018.8703316.
28. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016). https://doi.org/10.1109/CVPR.2016.90.
29. Asim, M. N. et al. EL-RMLocNet: An explainable LSTM network for RNA-associated multi-compartment localization prediction. *Comput. Struct. Biotechnol. J.* **20**, 3986–4002 (2022).
30. Meher, P. K., Satpathy, S. & Rao, A. R. Publisher Correction: miRNALoc: predicting miRNA subcellular localizations based on principal component scores of physico-chemical properties and pseudo compositions of di-nucleotides. *Sci. Rep.* **11**, 3287 (2021).
31. Chen, G. et al. Exosomal PD-L1 contributes to immunosuppression and is associated with anti-PD-1 response. *Nature* **560**, 382–386 (2018).
32. Vakhshiteh, F., Hassani, S., Momenifar, N. & Pakdaman, F. Exosomal circRNAs: new players in colorectal cancer. *Cancer Cell Int.* **21**, 483 (2021).
33. Hashemi, M. et al. Pre-clinical and clinical importance of miR-21 in human cancers: Tumorigenesis, therapy response, delivery approaches and targeting agents. *Pharmacol Res.* **187**, 106568 (2023).
34. Ghafouri-Fard, S., Khoshbakht, T., Hussen, B. M., Taheri, M. & Samadian, M. A review on the role of miR-1246 in the pathoetiology of different cancers. *Front Mol. Biosci.* **8** (2022).
35. Lu, Y., Godbout, K., Lamothe, G. & Tremblay, J. P. CRISPR-Cas9 delivery strategies with engineered extracellular vesicles. *Mol. Ther. Nucleic Acids* **34**, 102040 (2023).
36. Kamerkar, S. et al. Exosomes facilitate therapeutic targeting of oncogenic KRAS in pancreatic cancer. *Nature* **546**, 498–503 (2017).
37. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).
38. Cui, T. et al. RNALocate v2.0: An updated resource for RNA subcellular localization with increased coverage and annotation. *Nucleic Acids Res.* **50**, D333–D339 (2022).
39. Sharma, N. et al. AlgPred 2.0: An improved method for predicting allergenic proteins and mapping of IgE epitopes. *Brief Bioinform.* **22** (2021).
40. Kaur, D., Arora, A., Vigneshwar, P. & Raghava, G. P. S. Prediction of peptide hormones using an ensemble of machine learning and similarity-based methods. *Proteomics* https://doi.org/10.1002/pmic.202400004 (2024).
41. TF–IDF. *Encyclopedia of Machine Learning* 986–987 (Springer US, 2011). https://doi.org/10.1007/978-0-387-30164-8_832.
42. Joshi, D., Mishra, A. & Anand, S. A naïve Gaussian Bayes classifier for detection of mental activity in gait signature. *Comput. Methods Biomech. Biomed. Eng.* **15**, 411–416 (2012).
43. Wu, Y., Ianakiev, K. & Govindaraju, V. Improved k-nearest neighbor classification. *Pattern Recognit.* https://doi.org/10.1016/S0031-3203(01)00132-7 (2002).
44. Bulac, C. & Bulac, A. Decision trees. In *Advanced Solutions in Power Systems: HVDC, FACTS, and AI Techniques* (2016). https://doi.org/10.1002/9781119175391.ch18.
45. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* https://doi.org/10.1007/s10994-006-6226-1 (2006).
46. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016). https://doi.org/10.1145/2939672.2939785
47. Stoltzfus, J. C. Logistic regression: A brief primer. *Acad. Emerg. Med.* **18**, 1099–1104 (2011).
48. Breiman, L. Random forests. *Mach. Learn.* https://doi.org/10.1023/A:1010933404324 (2001).
49. Cristianini, N. & Ricci, E. Support vector machines. In *Encyclopedia of Algorithms* 928–932 (Springer US, 2008). https://doi.org/10.1007/978-0-387-30162-4_415.
50. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011).
51. Agrawal, T. Hyperparameter optimization using scikit-learn. In *Hyperparameter Optimization in Machine Learning* 31–51 (Apress, 2021). https://doi.org/10.1007/978-1-4842-6579-6_2.
52. Kalchbrenner, N., Grefenstette, E. & Blunsom, P. A convolutional neural network for modelling sentences. (2014).
53. LeCun, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**, 541–551 (1989).
54. Bergstra, J., Ca, J. B. & Ca, Y. B. Random search for hyper-parameter optimization Yoshua Bengio. *J. Mach. Learn. Res.* vol. 13 http://scikit-learn.sourceforge.net. (2012).
55. Rathore, A. S., Arora, A., Choudhury, S., Tijare, P. & Raghava, G. P. S. ToxinPred 3.0: An improved method for predicting the toxicity of peptides. *bioRxiv* 2023.08.11.552911 (2023) https://doi.org/10.1101/2023.08.11.552911.
56. Chaudhary, K., Nagpal, G., Dhanda, S. K. & Raghava, G. P. S. Prediction of Immunomodulatory potential of an RNA sequence for designing non-toxic siRNAs and RNA-based vaccine adjuvants. *Sci. Rep.* **6**, 20678 (2016).
57. Bhalla, S. et al. Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *Sci. Rep.* **7**, 44997 (2017).
58. Dhall, A., Patiyal, S., Sharma, N., Usmani, S. S. & Raghava, G. P. S. Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Brief Bioinform.* **22**, 936–945 (2021).
59. Jarwal, A. et al. A deep learning method for classification of HNSCC and HPV patients using single-cell transcriptomics. *Front. Mol. Biosci.* **11** (2024).
60. Aggarwal, S. et al. An ensemble method for prediction of phage-based therapy against bacterial infections. *Front Microbiol.* **14**, 1148579 (2023).

Acknowledgements

The authors are thankful to the Council of Scientific and Industrial Research (CSIR) for providing the fellowship. The authors are also thankful to the Department of Computational Biology, IIITD, New Delhi for its infrastructure and facilities. We thank the Department of Biotechnology (DBT) for providing an infrastructure grant to the institute (Grant BT/PR40158/BTIS/137/24/2021). We would like to acknowledge that figures were created using BioRender, and English was corrected using Grammarly.

Author contributions

AA collected and processed the data, implemented the algorithms, developed the prediction models, and built the front end and back end of the web server. AA, and GPSR prepared the manuscript. GPSR conceived and coordinated the project. All authors have read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Biorxiv DOI

https://doi.org/10.1101/2024.06.20.599824

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-15814-y.

Correspondence and requests for materials should be addressed to G.P.S.R.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.