Received: 29 May 2023 | Revised: 20 July 2023 | Accepted: 21 July 2023

DOI: 10.1002/pmic.202300231

Proteomics
Proteomics and Systems Biology

RESEARCH ARTICLE

A random forest model for predicting exosomal proteins using evolutionary information and motifs

Akanksha Arora | Sumeet Patiyal | Neelam Sharma | Naorem Leimarembi Devi | Dashleen Kaur | Gajendra P. S. Raghava

Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

Correspondence
Gajendra P. S. Raghava, Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Industrial Estate, Phase III, New Delhi 110020, India.
Email: raghava@iiitd.ac.in

Funding information
Department of Biotechnology, Ministry of Science and Technology, India, Grant/Award Number: BT/PR40158/BTIS/137/24/2021

Abstract

Non-invasive diagnostics and therapies are crucial to prevent patients from undergoing painful procedures. Exosomal proteins can serve as important biomarkers for such advancements. In this study, we attempted to build a model to predict exosomal proteins. All models are trained, tested, and evaluated on a non-redundant dataset comprising 2831 exosomal and 2831 non-exosomal proteins, where no two proteins have more than 40% similarity. Initially, the standard similarity-based method Basic Local Alignment Search Tool (BLAST) was used to predict exosomal proteins, which failed due to low-level similarity in the dataset. To overcome this challenge, machine learning (ML) based models were developed using compositional and evolutionary features of proteins achieving an area under the receiver operating characteristics (AUROC) of 0.73. Our analysis also indicated that exosomal proteins have a variety of sequence-based motifs which can be used to predict exosomal proteins. Hence, we developed a hybrid method combining motif-based and ML-based approaches for predicting exosomal proteins, achieving a maximum AUROC of 0.85 and MCC of 0.56 on an independent dataset. This hybrid model performs better than presently available methods when assessed on an independent dataset. A web server and a standalone software ExoProPred (https://webs.iiitd.edu.in/raghava/exopropred/) have been created to help scientists predict and discover exosomal proteins and find functional motifs present in them.

KEYWORDS
exosomal proteins, exosomes, extracellular vesicles, machine learning, motifs, PSSM profile

1 | INTRODUCTION

Protein secretion is crucial for a wide range of functions, including communication among cells [1]. The majority of secreted proteins in eukaryotes go along the (ER)-Golgi pathway [2]. This pathway is guided via a signal peptide present on the N-terminus of the protein, also known as the leader sequence. It helps deliver the nascent proteins from ER to the Golgi apparatus, which are then transported to the cell surface via vesicles [3]. Apart from the classical pathway, that is, the ER-Golgi pathway, some proteins are also secreted

Abbreviations: AUC, area under the curve; AAI, amino acid index; AAC, amino acid composition; ATC, atom composition; BLAST, basic local alignment search tool; DT, decision tree; GNB, Gaussian naive bayes; KNN, K-nearest neighbors; LR, logistic regression; ML, machine learning; MCC, Matthews correlation coefficient; MERCI, Motif emerging and with classes-identification; PSSM, position-specific scoring matrix; PAAC, pseudo amino acid composition; QSO, quasi sequence order; RFE, recursive feature selection; RF, random forest; SEP, shannon entropy; SER, shannon entropy of residue level; SVC, support vector classifier; XGB, extreme gradient boosting.

through unconventional pathways that are able to secrete the leaderless proteins. Unconventional pathways involve both non-vesicular and vesicular transport. In non-vesicular transport, proteins are secreted into the extracellular space, whereas in vesicular transport, proteins are secreted via vesicles. These vesicular structures comprise a variety of classes, and among these classes, exosomes stand out [4, 5].

Exosomes belong to a class of extracellular vesicles with endosomal origin are derived from cells, and range from size 30 to 150 nm [6]. They facilitate interactions with the cellular environment and are extensively found in bodily fluids like urine, saliva, blood, cerebrospinal fluid, bile, breast milk, amniotic fluid, semen, epididymal fluid, and sputum [7]. They are produced in the cytosol as a result of inward budding on late endosomes to form intraluminal vesicles (ILVs) inside a large multivesicular body (MVB) [8]. When MVB merges with the plasma membrane, ILVs are secreted as exosomes into the extracellular environment [9]. Exosomes encompass a compound cargo of contents arising from the original cell, including lipids, DNA, proteins, miRNA, and mRNA (Figure 1) [10]. The content carried by exosomes can change in diseased conditions making it a useful entity for biomarker detection [11]. Exosome-based diagnostics are more specific and sensitive than liquid biopsy or conventional biopsy biomarkers due to their high stability in body fluids [12, 13]. In addition, exosomal markers are readily available from most biofluids which makes exosome-based diagnostics labor and cost-effective [14, 15]. Since proteins and peptides are the most widely studied macromolecules as biomarkers, identifying and annotating exosomal proteins can help develop the least-invasive novel diagnostic methods as well as therapies for various diseases [16–18]. The proteins extracted from the circulating exosomes can give us comprehensive information about a specific disease – for example – exosomal proteins can give us important evidence about distal tumors, which is otherwise difficult to obtain due to complex diagnostic methods like tissue biopsy [16]. Extracting proteins from exosomes is more efficient than extracting them from blood, as blood has many substances [19].

Identifying proteins secreted by cells via exosomes has its own challenges, as cells produce a wide range of highly similar proteins. In addition, exosomes arise from a range of different cell types, and it would be difficult to determine their origin tissue unless they carry extremely specific cargo [20]. Thus, it is crucial to develop a computational method that can predict proteins secreted by exosomes. In this direction, there are several existing methods to predict classical and non-classical secreted proteins that include SRTpred, OutCyte, SecretP, SPRED, and SecretomeP 2.0 [21–25]. None of them has been specifically trained on proteins secreted by exosomes or have discovered motifs found in exosomal proteins. There is only one method ExoPred, that is trained on exosomal proteins for vertebrates [26]. To complement presently available methods, we made a systematic attempt to build a classifier that can annotate human exosomal proteins accurately. We have used a wide range of model-building techniques, different types of protein features, and a motif-based approach (see Figure 2). In addition, we have provided users with a novel method to predict exosomal motifs in the sequences. This can help researchers in designing and discovering new exosomal proteins.

Significance Statement

Problem: Identification of secretory proteins in body fluids is one of the key challenges in the development of non-invasive diagnostics. It has been shown in the past that a significant number of proteins are secreted by cells via exosomes called exosomal proteins.

What is already known: The pre-existing webservers are able to predict whether a protein is secreted from unconventional pathways. There is only one existing software that particularly predicts exosomal proteins; however, it is not able to predict the same accurately.

What this paper adds: We have attempted to create a webserver that is able to predict exosomal proteins accurately. In addition, it also gives the users functional motifs specific to exosomal proteins which we believe will be useful in developing novel protein sequences for exosomal drug delivery and getting an understanding of the mechanism of how proteins are transported via exosomes.

2 | MATERIALS AND METHODS

2.1 | Compilation and processing of the dataset

The data used in this research work was retrieved from UniProt release 2022_02 (Released on May 25, 2022) and from the ExoPred dataset [26, 27]. We retrieved 2178 exosomal proteins from UniProt using the following queries; (i) (go:0070062) AND (reviewed:true) AND (organism_id:9606), (ii) "extracellular exosome" AND (reviewed:true) AND (organism_id:9606), and (iii) "exosome" AND (reviewed:true) AND (organism_id:9606). In addition, we retrieved 2551 exosomal proteins from the ExoPred dataset, which are reviewed proteins belonging to humans. After compiling the data extracted from UniProt and ExoPred, we had a total of 3915 exosomal proteins. Similarly, we extracted 18,207 non-exosomal proteins from UniProt using the following query, NOT (go:0070062) NOT Exosomes NOT "Extracellular exosome" NOT Exosome AND (reviewed:true) AND (organism_id:9606). We also combined these non-exosomal proteins with the non-exosomal proteins from the ExoPred dataset. Finally, we got 20,330 unique non-exosomal proteins after removing duplicates. We also removed proteins consisting of non-standard amino acids "BJOUXZ" and sequences with lengths <55 and >1500. Finally, we obtained 2831 non-redundant exosomal proteins after discarding redundant sequences using CD-HIT software where no two proteins have more than 40% similarity [28]. Similarly, we obtained 10,680 non-exosomal proteins after removing redundant sequences. The final dataset contains 2831 exosomal and 2831 non-exosomal (randomly selected from 10,680 non-exosomal sequences) proteins.

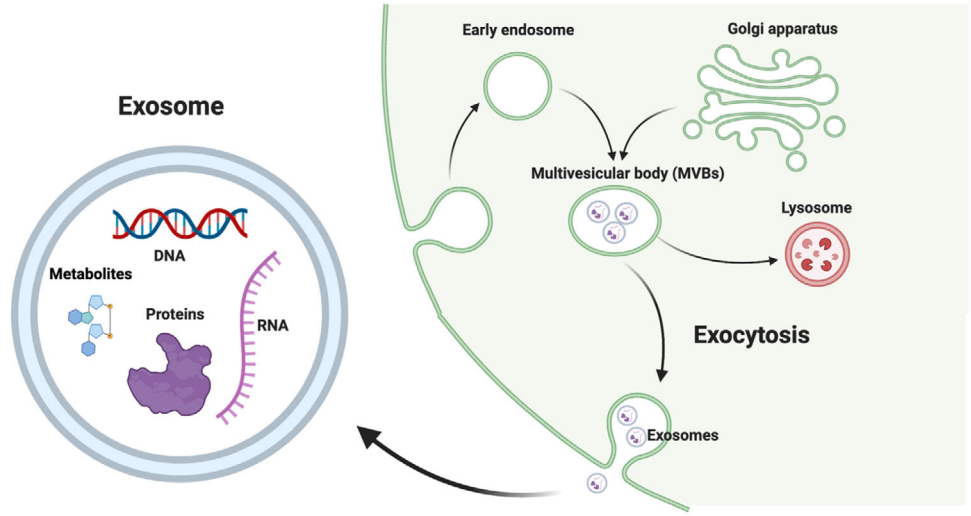

FIGURE 1 Mechanism of formation of exosomes.

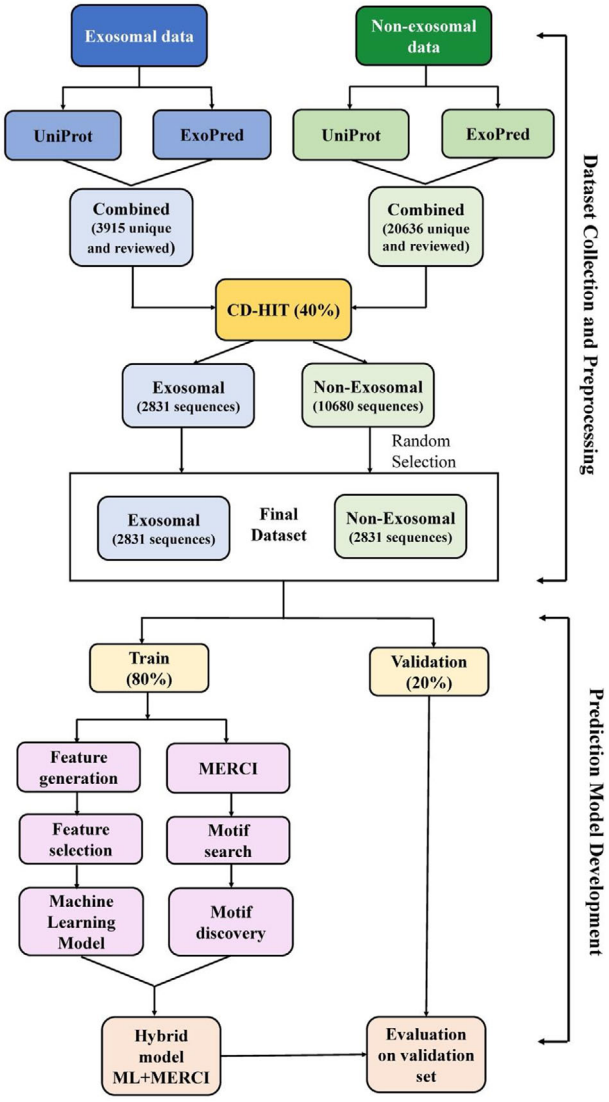

FIGURE 2 Flowchart of the methodology followed in the study.

2.2 | Feature generation

To develop a prediction model to classify proteins, we need a set of features for every protein.

A number of feature encoding techniques have been used in previous studies [29–32]. We used a standalone tool called Pfeature to compute numerous features for the proteins, including evolutionary information-based features and composition-based features [33].

2.2.1 | Composition-based features

The composition-based feature module available on Pfeature provides a vector of 9163 features for every protein in the positive (exosomal) and negative (non-exosomal) dataset like amino acid composition (AAC), tri-peptide composition (TPC), di-peptide composition (DPC), and many more.

2.2.2 | Evolutionary features

The evolutionary features of a protein are known to provide additional important information about proteins than its other primary sequence features [34, 35]. The evolutionary information can be retrieved by calculating the position-specific scoring matrix (PSSM) profile using Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) for each protein [36]. In PSSM, we obtain a matrix containing the dimensions 20 × length of sequence for protein or peptide sequences. As we are using multiple sequences together for the prediction, we need a fixed-length vector to develop machine learning (ML) models. Hence, we have used PSSM-400 composition profiles as evolutionary features, which have been described in earlier studies [35]. PSSM-400 is a fixed 20 × 20 dimension vector for a protein sequence which comprises the measure of occurrences of

20 amino acids in the sequence. We have created a PSSM matrix for each sequence which was first normalized within the range of 0–1 and converted into a PSSM composition of size 20 × 20 vector [33].

2.3 | Feature selection

It has been shown earlier that all the features extracted from a protein are not relevant, and there is a need to select only the useful ones from a big set of features [37]. To achieve the same, we applied RFE feature selection technique using Logistic Regression (LR) as the estimator [38]. We selected the top 20 and top 50 most relevant compositional features and evolutionary features (PSSM composition), respectively. This feature selection method keeps removing the weakest features from the set until a specified number of features has been reached. The features were selected from the standardized data that was obtained using StandardScaler method [39]. The features that were top-ranked were then used to create several machine-learning prediction models for the dataset. The features used in the ML models are described in Table S1.

2.4 | Similarity search using BLAST

BLAST version-2.2.29+ is widely used to identify and annotate protein and nucleotide sequences [40]. In this research study, we tried to use BLAST for the identification of exosomal proteins. It is based on the protein sequence similarity with exosomal and non-exosomal protein sequences. The protein query sequences were made to hit against a database of exosomal and non-exosomal protein sequences.

We applied three approaches to identify exosomal sequences, which involved taking into account the top hit, top three hits, and top five hits at various *E*-value cut-offs. In the first strategy, that is, first hit, – the sequence is identified as exosomal or non-exosomal based on its first hit against the whole database. However, for the top three and five hits, a voting approach is considered, and a sequence is identified as exosomal if top three or five hits have the maximum of exosomal proteins. The non-exosomal proteins are also characterized in the same manner. For this, a minimum of three or five hits must be available for voting. The performance of these three strategies was recorded for different *E*-values. Several researchers have used this methodology to identify a protein sequence [35, 41].

2.5 | Motif search

It is essential to recognize the functional motifs present in the protein or peptide sequences for their functional annotation as well as to classify the negative and positive datasets. In this study, we used Motif Emerging with Classes Identification (MERCI) program to find motifs in both exosomal and non-exosomal protein sequences [42]. MERCI selects specific motifs in the positive dataset by comparing negative and positive input sequences. Hence, to retrieve the particular motifs in exosomal and non-exosomal protein sequences, we followed a two-step procedure that involved – (a) Providing exosomal proteins as positive input and non-exosomal proteins as negative input and finding motifs for exosomal protein sequences, (b) Reversing the order for positive and negative input to find motifs for non-exosomal protein sequences.

We used different options available in MERCI to extract motifs that are exclusive as well as inclusive to both sets. By default, MERCI takes the maximal frequency of the negative sequences (fn) as zero, which gives only exclusive motifs, that is, the motifs that are not common in positive and negative sets. We increased this value to fn = 8 to obtain inclusive motifs as well. Within the exclusive and inclusive motifs, we got different kinds of motifs by specifying some values that include – (a) No gap, (b) Gap = 1, (c) Gap = 2, and (d) Class = Koolman–Rohm. After that, the unique proteins containing motifs were selected to compute the overall coverage of motifs in the protein sequences.

2.6 | ML classifiers

We have employed several ML algorithms to differentiate between exosomal and non-exosomal proteins. These algorithms involve Gaussian Naïve Bayes (GNB), K-Nearest Neighbors (KNN), Decision Tree (DT), Extreme Gradient Boosting (XGB), Logistic RegreLR, Support Vector classifier (SVC), and random forest (RF). The parameters of these algorithms were optimized using hyperparameter tuning.

2.7 | Performance metrics calculation and cross-validation

The whole dataset was divided into the ratio of 80:20, where 80% comprised the training and 20% validation data. The five-fold cross-validation technique was applied to 80% of the training data to assess the ML models, and the remaining 20% was kept unknown to the models. In the five-fold cross-validation technique, 80% of training data is split into five parts where four folds are used for training, and the left one fold is used as a test set for internal validation purposes. This procedure is reiterated five times so that every fold gets a chance to be the test fold. The ML models used in this study have been evaluated using performance metrics which include parameters dependent and independent of the threshold. The different standard evaluation metrics that have been used in this study include sensitivity, specificity, Matthews correlation coefficient (MCC), accuracy, and area under the receiver operating characteristics (AUROC). Out of these, AUROC is threshold-independent, and the rest of the parameters are threshold-dependent. The threshold-dependent parameters, like specificity, sensitivity, and MCC, were optimized to obtain a threshold with the maximum values. These metrics have been previously used in studies to estimate the performance of ML models [43–45].

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \quad (1)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FP} + \text{TN}} \times 100 \quad (2)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \times 100 \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \quad (4)$$

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

2.8 | Hybrid model

To improve the prediction of ML models, we applied a hybrid approach that integrates and employs the various results obtained in this study. The hybrid approach uses a weighted scoring method in which the scores are calculated by combining two methods (i) Motif-based approach and (ii) ML-based methods. In this hybrid model, we assigned a score of +0.5 if a protein sequence had an exosomal motif and −0.5 if it had a non-exosomal motif, and 0 if no motif was found. These scores were combined with the ML prediction scores, which were obtained using the predict_proba() function. It gave us the probability of a protein belonging to a particular class instead of a binary result. The motif score and ML score together formed an overall score for every protein ranging from −0.5 to +1.5. The scoring method is described in Equation (5). The sequences were categorized as exosomal and non-exosomal by analyzing the overall scores. A number of studies have used this hybrid approach earlier [41, 46].

$$S' = \begin{cases} S + 0.5 & \textit{If exosomal motif present} \\ S - 0.5 & \textit{If non-exosomal motif present} \\ S & \textit{If no motif is found} \end{cases} \quad (5)$$

Here, S = Prediction score obtained from ML-based approach, S' = Hybrid score ranging from −0.5 to 1.5 obtained by ML-based and motif-based approaches.

3 | RESULTS

3.1 | Amino acid composition analysis

After analyzing and comparing the AAC of exosomal and non-exosomal proteins, we have discovered that there is only a slight amount of difference in the average AACs. However, we performed a two-sided Mann–Whitney *U* test on the data to compare the AACs of both exosomal and non-exosomal protein groups. The two-sided Mann–Whitney *U* test is used to compare the central tendencies of the two groups without making any assumption on the distribution of the data. We found that the difference between the averages in these groups was significant for about 15 amino acids, and the *p*-values for each amino acid are given in Table S2 and shown in Figure 3. The maximum difference between averages was observed in AACs of serine (0.93) followed by leucine (0.76) and proline (0.64) with a p-value < 0.05.

3.2 | BLAST performance

BLAST is widely used to annotate and recognize the role of a query protein sequence on the basis of similarity search. We attempted to utilize BLAST in this study to classify proteins as exosomal and non-exosomal. We used five-fold cross-validation to evaluate the performance of BLAST. Firstly, the sequences in four folds are used to create a database, and the sequences present in the fifth fold are used to hit against the sequences present in the respective database. This procedure was repeated five times. To evaluate BLAST performance on the validation dataset, we constructed a database using all the training sequences, and those in the validation set were searched against the database. We used BLAST in three ways – (a) top hit, (b) top 3 hits, and (c) top 5 hits. The top hit is a standard method that assigns a class to the protein based on the first hit, whereas the top 3 and top 5 criteria assign a class to the protein on the basis of the class that appears the maximum number of times in the first 3 and 5 hits. We have considered the top 3 and 5 hits as sometimes the top hit is not always the most relevant one. However, even after trying all these criteria, we were getting a large number of false positives and false negatives. We obtained 18.06%, 11.08%, and 8.39% sensitivity (number of correct hits) for the training dataset, whereas the sensitivity of 17.92%, 11.39%, and 8.65% was obtained for the validation dataset for top 1, top 3, and top 5 hits, respectively. With the increment of the *E*-value, the error rate was also increasing. For the training set, we got 9.45%, 4.75%, and 3.2%, and for the validation set, we got 12.97%, 7.41%, and 6% for top 1, top 3, and top 5 hits, respectively. The results for BLAST are shown in Table 1.

3.3 | ML models

For each protein sequence in the dataset, a total of 9163 features were computed that constituted more than ten types of compositional features. Along with the composition-based features, evolutionary features were also computed. We first developed ML models on features like AAC and PSSM Composition which led to an AUROC of 0.70 on RF model and 0.72 on LR model, respectively. The results for AAC and PSSM composition are given in Table 2.

In order to improve the model's performance, we performed feature selection on the set of 9163 features using the feature selection technique – RFE. The best-performing ML model was RF model on 70 features containing the top 20 compositional features and top 50 evolutionary (PSSM) features. It obtained an AUROC of 0.73 for the independent validation set. The detailed results for all ML models performed for selected features is given in Table 3. The ML models were developed using the scikit-learn package in Python, and hyperparameter

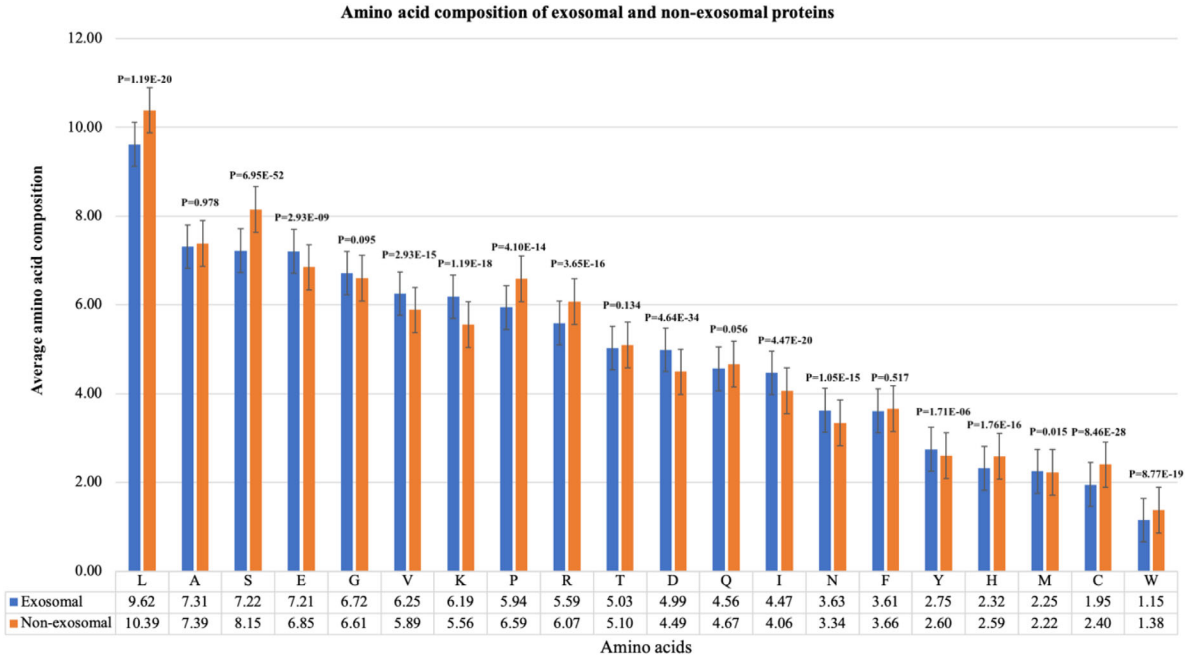

FIGURE 3 Amino acid composition analysis for exosomal and non-exosomal proteins.

tuning was done to augment the algorithms using grid search [39]. Grid search finds the combination of best hyperparameters for a given model by exhaustively generating candidates from the grid of parameter values. The parameters used for each algorithm are described in Table S3.

3.4 | Compositional features

The AAC for the exosomal and non-exosomal proteins was calculated to build the ML models. The RF model was observed to perform well than other models and was able to achieve an AUROC of 0.71 and 0.70 on training and validation sets, respectively. The performance of AAC for the dataset is given in Table 2.

3.5 | Evolutionary features

The ML models were also developed on the basis of evolutionary information. To obtain the evolutionary information, we computed the PSSM profiles of each protein which were then fed to our ML models. It was observed that the LR model was performing best on these features and achieved the AUROC of about 0.73 and 0.72 for training and validation sets. The performance of PSSM-based models is given in Table 2.

3.6 | Feature selection

A total of 9163 composition-based features were generated, which were short-listed to the top 20 features using the RFE feature selection method based on the LR estimator. These 20 features attained an AUROC of 0.71 on training and 0.71 on validation sets based on the SVC. We also selected the top 50 features for evolutionary information-based features using the same technique, which yielded the AUROCs of 0.74 on training and 0.71 on validation sets using a RF classifier. We also compiled the top 20 compositional and top 50 evolutionary features, which resulted in a matrix of a total of 70 features. The combination of these features was able to achieve the AUROCs of 0.75 on training and 0.73 on validation sets. The results for all the selected features are shown in Table 3.

3.7 | Top selected features

The top compositional features include Amino Acid Index (AAI), Atom Composition (ATC), Pseudo Amino Acid Composition (PAAC), Shannon Entropy (SEP), Quasi-Sequence Order (QSO), and Shannon Entropy of Residue Level (SER). Amongst these, it was observed that three of the relevant features include SER, QSO, and PAAC of tryptophan (W), which indicates that tryptophan can be an important amino acid for differentiating between exosomal and non-exosomal proteins. Along with this, it was observed that the ATC of Nitrogen and Sulphur were also two of the critical features in predicting exosomal proteins.

3.8 | Motif search

We attempted to identify the exclusive and inclusive set of motifs present in exosomal and non-exosomal proteins using the publicly available MERCI program. To achieve this, we extracted motifs using different parameters like – (a) no gap, (b) gap = 1, (c) gap = 2, and

TABLE 1 Results for top 1, top 3, and top 5 hits in Basic Local Alignment Search Tool (BLAST) for validation set searched against the training set database (here, sens = sensitivity and spec = specificity).

Top 1		Training				Validation			
		Exosomal		Non-exosomal		Exosomal		Non-exosomal	
	e-values	Sens	Error	Spec	Error	Sens	Error	Spec	Error
	10^{-6}	18.06	9.45	10.29	10.86	17.92	9.97	13.24	12.97
	10^{-5}	18.75	9.74	10.8	11.28	18.45	10.41	14.03	13.06
	10^{-4}	19.17	10.11	11.3	11.57	19.24	10.68	14.56	13.77
	10^{-3}	19.96	10.69	11.99	12.14	19.59	11.3	15.45	13.86
	10^{-2}	20.71	11.35	12.83	12.89	20.12	11.83	16.42	14.74
	10^{-1}	22.26	12.52	14.09	14.68	21.09	12.53	18.01	16.24
Top 3		**Training**				**Validation**			
		Exosomal		**Non-exosomal**		**Exosomal**		**Non-exosomal**	
	***e*-values**	**Sens**	**Error**	**Spec**	**error**	**Sens**	**Error**	**Spec**	**Error**
	10^{-6}	11.08	4.75	6.12	6.65	11.39	4.5	6.35	7.41
	10^{-5}	11.99	5.03	6.49	6.98	12	4.94	6.8	7.86
	10^{-4}	12.74	5.37	6.91	7.51	12.53	5.12	7.15	8.38
	10^{-3}	13.58	5.85	7.37	7.84	13.24	5.65	7.86	8.74
	10-2	14.55	6.51	8.06	8.54	13.95	6.18	8.91	9.36
	10^{-1}	15.9	7.24	9.07	9.32	15.09	6.97	10.24	10.59
Top 5		**Training**				**Validation**			
		Exosomal		**Non-exosomal**		**Exosomal**		**Non-exosomal**	
	***e*-values**	**Sens**	**Error**	**Spec**	**Error**	**Sens**	**Error**	**Spec**	**Error**
	10^{-6}	8.39	3.2	4.22	5.12	8.65	3.35	4.94	6
	10^{-5}	8.9	3.44	4.7	5.5	8.91	3.8	5.12	6.62
	10^{-4}	9.43	3.73	4.86	5.9	9.89	4.06	5.74	6.88
	10^{-3}	10.16	4.15	5.45	6.34	10.68	4.41	6.53	6.97
	10^{-2}	10.97	4.46	6.01	6.76	11.39	4.85	6.88	7.68
	10^{-1}	12.1	5.06	7	7.33	12.53	5.74	7.68	8.38

(d) class = Koolman–Rohm. By default, MERCI takes fn (maximal frequency in negative sequences) as zero, which gives exclusive motifs; we increased it to fn = 8 to get inclusive motifs for both negative and positive datasets. Altogether, MERCI provided 89 motifs in exosomal and 130 motifs in a non-exosomal set that covered 1441 exosomal and 1373 non-exosomal sequences. The top 5 motifs in each category for the exclusive and inclusive sets and the number of sequences they occurred in are given in Table 4. It is observed that most of the motifs crucial for predicting exosomal proteins consisted of aliphatic amino acids.

3.9 | Hybrid approach

Since we were getting a good sequence coverage using the motif search, we decided to combine motif prediction with ML-based prediction. In this hybrid model, we allotted a protein sequence a score of +0.5 if an exosomal motif was present, −0.5 if a non-exosomal motif was present, and 0 if none were present. These scores were compiled with the scores obtained from ML-based predictions. After merging the motif prediction scores with RF model prediction based on AAC, we acquired an AUROC of 0.86 for training and 0.84 for the validation dataset. We also merged the motif prediction scores with RF model prediction based on the top 70 features (20 compositional and 50 evolutionary), which attained an AUROC of 0.87 for training and 0.85 for the validation set. The performance of other models has been explained in Table 5. The AUROC plots for training and validation sets for ML and hybrid models are illustrated in Figure 4.

3.10 | Web server development

We developed a web server – ExoProPred (https://webs.iiitd.edu.in/raghava/exopropred/) to predict exosomal proteins. We have integrated our two best-performing hybrid models — (a) AAC combined with MERCI and (b) Top 70 features combined with MERCI. The web

TABLE 2 Results for ML models developed for AAC and PSSM composition features.

Amino acid composition (AAC)										
Training						**Validation**				
Model	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	59.75	59.92	59.84	0.62	0.20	50.18	61.75	56.13	0.58	0.12
RF	**64.58**	**66.15**	**65.36**	**0.71**	**0.31**	**64.91**	**65.01**	**64.96**	**0.70**	**0.30**
LR	63.04	65.21	64.12	0.69	0.28	59.09	63.47	61.34	0.67	0.23
XGB	63.92	64.68	64.30	0.70	0.29	61.64	65.35	63.55	0.70	0.27
KNN	62.87	65.30	64.08	0.69	0.28	64.36	60.89	62.58	0.68	0.25
GNB	61.68	62.32	62.00	0.67	0.24	60.00	62.26	61.17	0.64	0.22
SVC	64.93	64.99	64.96	0.70	0.30	62.36	61.92	62.14	0.68	0.24
PSSM composition										
Training						**Validation**				
Model	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	57.87	58.26	58.06	0.62	0.16	56.36	54.72	55.52	0.59	0.11
RF	68.30	66.76	67.54	0.73	0.35	66.73	65.87	66.28	0.71	0.33
LR	**67.91**	**67.29**	**67.60**	**0.73**	**0.35**	**65.09**	**67.24**	**66.20**	**0.72**	**0.32**
XGB	65.28	64.80	65.04	0.71	0.30	64.91	61.41	63.11	0.70	0.26
KN	66.37	64.93	65.66	0.71	0.31	65.46	59.18	62.22	0.69	0.25
GNB	67.51	60.48	64.02	0.68	0.28	63.82	59.52	61.61	0.67	0.23
SVC	67.82	66.89	67.36	0.73	0.35	65.82	65.87	65.84	0.71	0.32

AUC, areas under the curve; DT, Decision Tree; GNB, Gaussian Naïve Bayes; KNN, K-Nearest Neighbors; LR, Logistic Regression; MCC, Matthews correlation coefficient; ML, machine learning; PSSM, position-specific scoring matrix; RF, random forest; SVC, Support Vector classifier; XGB, Extreme Gradient Boosting.

server incorporates the key modules, including (a) prediction, (b) motif scan, and (c) download. The "prediction module" allows users to submit their query protein sequences in FASTA format. This module can predict exosomal and non-exosomal proteins effectively. The second module, "motif scan," can identify the motifs present in exosomal and non-exosomal protein sequences using the MERCI software. This module can also scan or map the motifs present in the protein sequence query entered by the user and differentiates between them as exosomal or non-exosomal sequences. The web server has been developed on a responsive HTML template and is compatible with various operating systems. We also built a Python-based standalone package of ExoProPred to help users easily predict and classify the sequences at a large scale which can be downloaded from the "download module" on the web server.

3.11 | Comparison with other prediction tools

Presently, there is only one tool that predicts exosomal proteins – ExoPred. Other tools, such as SecretomeP 2.0, and Outcyte, predict whether the protein is following an unconventional pathway [21, 22]. We entered our validation set of 569 sequences into each of the servers after subtracting the sequences taken from the ExoPred dataset and performing a comparative analysis. ExoPred was able to predict the sequences with 66.08% accuracy [26]. However, it had very low sensitivity but high specificity, which means it is able to predict the non-exosomal sequences but not able to classify exosomal sequences correctly. For SecretomeP 2.0, we selected the "mammalian" option on the web server, and as indicated on their webpage, proteins with "NN score" higher than 0.6 are said to be secreted via unconventional pathways. After setting this threshold, we found that it was able to predict exosomal sequences with 54.83% accuracy [24]. In the Outcyte web server, we selected Outcyte UPS (Unconventional protein secretion option) and obtained an accuracy of 61.16% for our validation set, with low sensitivity and comparatively higher specificity [21]. On entering these sequences on our web server – ExoProPred, we obtained an accuracy of 79.4%, which is higher than all the above-mentioned tools. ExoProPred is also able to achieve a balanced sensitivity and specificity along with the highest accuracy. The full comparison of prediction by web servers is given in Table 6.

4 | DISCUSSION

There is a need to develop non-invasive diagnostic methods and therapies to prevent patients from going through painful medical procedures to get treatment. Exosomal biomarkers can be found in body fluids (saliva, blood, urine, etc.) in abundance and can be used to detect a disease or develop a treatment for different types of conditions [11]. These biomarkers are derived from the parent cells and are even more

TABLE 3 Results for ML models developed for the top 20 composition features, top 50 evolutionary (PSSM) features, and combination of top selected composition and evolutionary (PSSM) features.

20 selected composition-based features										
Training						**Validation**				
Model	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
DT	59.75	59.16	59.46	0.62	0.19	60.55	55.06	57.72	0.61	0.16
RF	63.00	64.37	63.68	0.69	0.27	60.00	62.95	61.52	0.68	0.23
LR	64.67	65.61	65.14	0.71	0.30	64.18	63.47	63.81	0.70	0.28
XGB	64.23	64.37	64.30	0.69	0.29	63.27	64.15	63.73	0.70	0.27
KNN	62.03	63.21	62.62	0.67	0.25	61.27	61.75	61.52	0.67	0.23
GNB	58.44	58.63	58.53	0.63	0.17	54.73	57.80	56.31	0.62	0.13
SVC	**65.28**	**65.13**	**65.20**	**0.71**	**0.30**	**66.36**	**65.52**	**65.93**	**0.71**	**0.32**
50 selected evolutionary-based features (PSSM Composition)										
Training						**Validation**				
Model	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	58.40	59.55	58.97	0.62	0.18	58.91	57.46	58.16	0.62	0.16
RF	**67.21**	**69.25**	**68.22**	**0.74**	**0.37**	**67.64**	**65.70**	**66.64**	**0.71**	**0.33**
LR	67.69	66.67	67.18	0.73	0.34	65.09	63.47	64.25	0.69	0.29
XGB	65.67	65.82	65.75	0.72	0.32	66.73	61.92	64.25	0.69	0.29
KN	64.58	65.69	65.13	0.71	0.30	65.46	60.21	62.75	0.68	0.26
GNB	65.19	65.38	65.28	0.70	0.31	64.00	62.26	63.11	0.68	0.26
SVC	66.94	67.33	67.14	0.73	0.34	64.36	63.29	63.81	0.69	0.28
70 selected features (20 compositional and 50 evolutionary (PSSM))										
Training						**Validation**				
Model	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	59.49	60.13	59.81	0.63	0.20	56.00	64.15	60.19	0.61	0.20
RF	67.08	69.47	68.26	0.75	0.37	66.91	67.07	66.99	0.72	0.34
LR	**68.83**	**67.82**	**68.33**	**0.75**	**0.37**	**68.55**	**64.67**	**66.55**	**0.73**	**0.33**
XGB	66.64	66.22	66.43	0.73	0.33	64.91	64.84	64.87	0.72	0.30
KN	63.57	64.04	63.80	0.69	0.28	66.00	63.81	64.87	0.70	0.30
GNB	64.14	64.53	64.33	0.70	0.29	64.36	63.12	63.73	0.68	0.28
SVC	68.70	67.82	68.26	0.75	0.37	66.91	65.87	66.37	0.72	0.33

AUC, areas under the curve; DT, Decision Tree; GNB, Gaussian Naïve Bayes; KNN, K-Nearest Neighbors; LR, Logistic Regression; MCC, Matthews correlation coefficient; ML, machine learning; PSSM, position-specific scoring matrix; RF, random forest; SVC, Support Vector classifier; XGB, Extreme Gradient Boosting.

specific and sensitive than those extracted directly from the body fluids because the exosome is highly stable and non-immunogenic [47]. Among the disease biomarkers, proteins have been broadly studied and can be used for the diagnosis, prognosis, and treatment of specific diseases [16–18]. However, it is difficult to identify these proteins as they are extremely similar to those produced by the cells, and there is a mixture of exosomes in the biofluids derived from different types of cells [20]. To overcome this limitation, we made an effort to develop a prediction server ExoProPred, that classifies the proteins into exosomal and non-exosomal.

In this study, we created a dataset for 2831 exosomal and 2831 non-exosomal proteins extracted from UniProt and ExoPred servers [26, 27]. A number of features were generated for the protein sequences (~9163 features). Firstly, we used features like AAC (20 features) and PSSM composition (400 features) to develop ML models, which resulted in an AUROC of 0.70 on RF model and 0.72 on LR model, respectively. The possible reasons why algorithms like LR outperformed powerful ML algorithms such as SVC and RF in PSSM could be — (a) LR assumes a linear relationship between the predictor variables and outcome variable. PSSM scores are a natural fit for this assumption as they represent a linearly scaled measure of the evolutionary conservation at each residue position, (b) In PSSM-based classification tasks, some residues may have weak or noisy signals, and even though algorithms like SVM are more robust to noise than LR, they might fail if the noise is too high and there are many outliers in the data.

TABLE 4 Top 5 motifs exclusive and inclusive motifs and the number of sequences in which they occurred (a) no gap, (b) gap = 1, (c) gap = 2, and (d) class = Koolman–Rohm (fn = maximal frequency in negative sequences, pos = occurrence in positive sequences, neg = occurrence in negative sequences).

Exclusive motifs (fn = 0)			Inclusive motifs (fn = 8)		
Motifs	**Pos**	**Neg**	**Motifs**	**Pos**	**Neg**
No gap					
IATG	14	0	HSASA	32	7
NRAL	13	0	PVLRRN	32	7
RIHTG	12	0	RLKCH	31	7
EKYL	12	0	SPPKC	31	8
IKAK	12	0	RLKTH	30	8
Gap = 1					
E R D gap E R	16	0	A I E gap T	41	8
G G L gap V L	16	0	P F gap R L	41	8
Q gap L S R L	16	0	I gap R V R	39	8
A L A E gap G	15	0	D R gap A I	37	7
A I gap E E L	14	0	D gap R A I	37	8
Gap = 2					
I gap I gap S G G	22	0	F gap D R gap F	40	8
E E V gap G gap K	19	0	F D gap R gap F	39	8
D E gap G gap Q V	18	0	R D gap D gap Y	37	7
E L E E gap L gap Q	18	0	E K A gap L gap A	36	7
G D A gap D gap L	18	0	–	–	–
Class = Koolman–Rohm					
Neutral G K T S	20	0	A I acidic T	43	6
E A E aliphatic aliphatic neutral aliphatic	20	0	D aliphatic D acidic aliphatic aliphatic	42	8
Aliphatic N aliphatic basic K aliphatic aliphatic	19	0	L E basic aliphatic aliphatic E	41	8
G acidic acidic K acidic	18	0	Acidic aliphatic K neutral Y	41	8
F aliphatic K acidic F	18	0	Acidic acidic aliphatic K aliphatic aliphatic aliphatic	40	8

Secondly, we tried to increase the performance of ML models by selecting only the relevant features that were mined from a big set of features using Recursive Feature Elimination (RFE). The best-performing model was RF model that was trained on 70 features (20 compositional and 40 PSSM features). It obtained an AUROC of 0.73 on an independent validation set. The top 20 compositional features included AAI, ATC, PAAC, SEP, QSO, and SER. Additionally, it was also observed that amino acids like serine, leucine, and proline showed the maximum difference in average AACs between exosomal and non-exosomal proteins.

Besides the development of ML models on selected essential features, we also applied the BLAST tool to identify the exosomal proteins, as this tool has been widely used to annotate the query proteins [40]. However, we were not able to obtain very high performance with BLAST. An explanation for this could be that exosomal proteins are very similar to the proteins present in their parent cell; hence, it must be difficult to point out which protein belongs to the exosome. We decided to exclude BLAST-based performance from our hybrid model, and to boost the performance of the hybrid model; we added a motif-based approach using MERCI in which we obtained 89 exosomal and 130 non-exosomal motifs covering 1441 exosomal and 1373 non-exosomal sequences [42]. In this study, we have identified novel exosomal and non-exosomal motifs using various methods like (1) no gap, (2) gap = 1, (3) gap = 2, and (4) class = Koolman–Rohm which are found exclusively as well as inclusively in both classes (exosomal and non-exosomal). On analyzing the top motifs obtained for the classification of exosomal and non-exosomal sequences, we observed that most motifs contained aliphatic amino acids.

The motif-based approach was able to cover a high amount of exosomal sequences. Hence, we combined this approach with the top-selected features to develop a hybrid model to predict exosomal protein sequences. We are able to achieve an accuracy of 78% and an AUROC of 0.85 with balanced specificity and sensitivity on an independent validation set. In addition to this, we obtained the highest accuracy for a validation set of 569 sequences when compared to other prediction web servers like Outcyte, ExoPred, and SecretomeP 2.0. The accuracy obtained by Outcyte, ExoPred, SecretomeP 2.0, and ExoProPred are 61.16%, 66.08%, 54.83, and 79.40%, respectively.

TABLE 5 Results of hybrid approach (a) MERCI + ML (AAC), (b) MERCI + ML (top 20 compositional features), (c) MERCI + ML (top 50 PSSM features), (d) MERCI + ML (top 70 features – compositional and evolutionary).

AAC										
Training						**Validation**				
Model	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
DT	72.42	73.71	73.06	0.8	0.46	66.73	74.44	70.7	0.78	0.41
RF	76.85	77.31	77.08	0.86	0.54	77.27	75.64	76.43	0.84	0.53
LR	76.46	75.89	76.18	0.85	0.52	74.91	74.44	74.67	0.84	0.49
XGB	76.11	76.69	76.4	0.85	0.53	76.55	75.99	76.26	0.84	0.53
KNN	76.33	75.93	76.13	0.85	0.52	78.18	72.9	75.46	0.84	0.51
GNB	73.91	73.67	73.79	0.82	0.48	73.27	70.84	72.02	0.79	0.44
SVC	76.24	77.58	76.9	0.85	0.54	74.91	74.44	74.67	0.84	0.49
20 selected features (compositional)										
Training						**Validation**				
Model	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	71.85	73.62	72.73	0.81	0.45	72.55	71.01	71.76	0.80	0.44
RF	75.67	74.96	75.31	0.85	0.51	75.64	74.27	74.93	0.83	0.50
LR	76.85	76.82	76.84	0.86	0.54	76.00	74.96	75.46	0.85	0.51
XGB	75.76	75.67	75.71	0.84	0.51	75.27	74.96	75.11	0.84	0.50
KNN	74.05	75.53	74.78	0.84	0.50	74.00	74.44	74.23	0.83	0.48
GNB	70.89	70.73	70.81	0.79	0.42	70.18	68.95	69.55	0.78	0.39
SVC	77.33	76.29	76.82	0.85	0.54	76.91	76.16	76.52	0.84	0.53
50 selected evolutionary features (PSSM)										
Training						**Validation**				
Model	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	73.13	71.83	72.48	0.80	0.45	72.00	71.70	71.84	0.80	0.44
RF	77.29	79.93	78.60	0.87	0.57	77.64	77.19	77.41	0.85	0.55
LR	77.55	78.91	78.22	0.86	0.56	76.00	75.47	75.73	0.84	0.51
XGB	75.93	77.57	76.74	0.86	0.54	78.00	75.81	76.88	0.84	0.54
KNN	75.41	77.88	76.63	0.86	0.53	76.55	73.07	74.76	0.84	0.50
GNB	77.90	73.56	75.75	0.83	0.52	77.82	71.36	74.49	0.81	0.49
SVC	78.21	78.50	78.36	0.86	0.57	76.18	74.27	75.20	0.84	0.50
70 selected features (20 compositional and 50 evolutionary)										
Training						**Validation**				
Model	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	72.56	73.12	72.84	0.80	0.46	71.09	73.41	72.29	0.79	0.45
RF	78.17	78.86	78.51	0.87	0.57	79.82	76.16	77.93	0.85	0.56
LR	78.83	77.84	78.33	0.87	0.57	78.18	75.64	76.88	0.85	0.54
XGB	77.51	77.13	77.32	0.86	0.55	75.82	75.30	75.55	0.85	0.51
KNN	74.27	76.68	75.46	0.85	0.51	77.27	75.47	76.35	0.84	0.53
GNB	74.40	74.23	74.32	0.82	0.49	74.36	71.18	72.73	0.81	0.46
SVC	77.99	79.22	78.60	0.87	0.57	76.91	77.99	77.05	0.85	0.54

AAC, amino acid composition; AUC, areas under the curve; DT, Decision Tree; GNB, Gaussian Naïve Bayes; KNN, K-Nearest Neighbors; LR, Logistic Regression; MCC, Matthews correlation coefficient; MERCI, Motif Emerging with Classes Identification; ML, machine learning; PSSM, position-specific scoring matrix; RF, random forest; SVC, Support Vector classifier; XGB, Extreme Gradient Boosting.

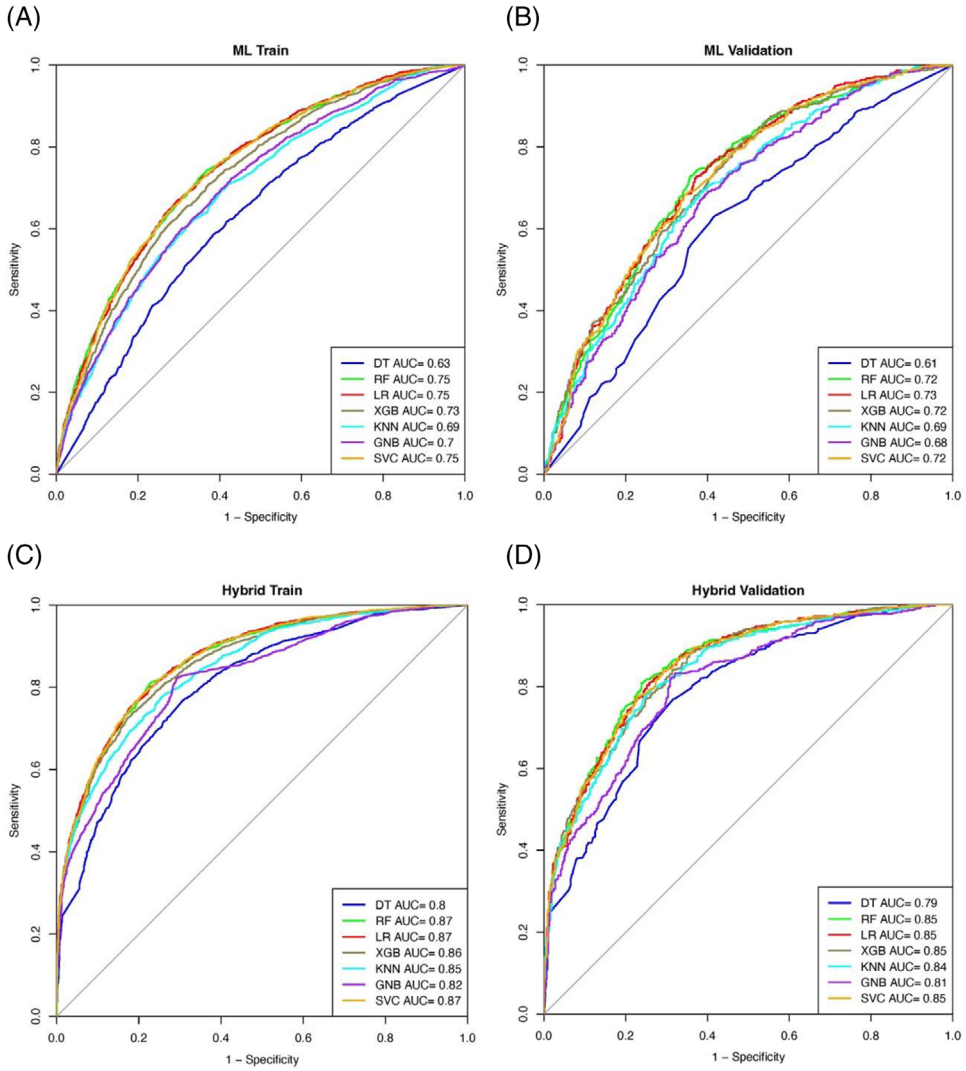

FIGURE 4 Area under the receiver operating characteristics (AUROC) plots for (A) training set in ML model, (B) validation set in ML model, (C) training set in the hybrid model, and (D) validation set in the hybrid model.

TABLE 6 Comparison of prediction by web servers ExoPred, SecretomeP 2.0, and Outcyte with ExoProPred on a validation dataset.

Prediction model	TP	FP	TN	FN	Sens	Spec	Acc
Exored	26	51	350	142	15.48%	87.28%	66.08%
SecretomeP 2.0	80	169	232	88	47.62%	57.85%	54.83%
Outcyte	47	100	301	121	27.97%	75.06%	61.16%
ExoProPred	**133**	**83**	**318**	**34**	**79.64%**	**79.30%**	**79.40%**

FN, false negative; FP, false positive; TN, true negative; TP, true positive.

We have created a platform that allows users to classify exosomal and non-exosomal protein sequences. In addition to the prediction of exosomal proteins, we have incorporated a motif-search program in our web server to help users discover exosomal and non-exosomal motifs in their query sequences to aid in the prediction as well as identification of new exosomal proteins. In this web server, we have implemented our best-performing model – the hybrid model.

5 | CONCLUSION

Exosomal proteins have diverse applications in healthcare, particularly in developing non-invasive disease biomarkers. These exosomal proteins are valuable in liquid biopsy, allowing non-invasive sampling for disease detection and monitoring. This study presents a highly accurate method for predicting exosomal proteins, which performs better

than the existing method. In addition to ML models called black box, we used similarity and motif-based approaches. In case a query sequence is highly similar to a known exosomal protein, we assign the query sequence as an exosomal protein; confidence in prediction depends upon the level of similarity. In addition, we discovered motifs associated with exosomal proteins to identify exosomal proteins. One of the objectives of this study is to facilitate researchers working in healthcare, thus, we developed an online web server and offline standalone software. We believe our study will benefit scientists worldwide studying protein or peptide diagnostics and therapies. The web server is freely available to encourage the use of this prediction method in research. We hope the development of ExoProPred enables the exploration of the potential of exosomal proteins and helps in the development of non-invasive diagnostic and therapeutic techniques for a range of ailments.

AUTHOR CONTRIBUTIONS

Akanksha Arora, Neelam Sharma, and Naorem Leimarembi Devi collected and processed the data. Akanksha Arora, Sumeet Patiyal, Neelam Sharma, and Naorem Leimarembi Devi implemented the algorithms. Akanksha Arora and Sumeet Patiyal developed the prediction models. Akanksha Arora and Sumeet Patiyal developed the front end and back end of the web server. Akanksha Arora, Dashleen Kaur, and Gajendra P. S. Raghava prepared the manuscript. Gajendra P. S. Raghava conceived and coordinated the project. All authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors are thankful to Council of Scientific and Industrial Research (CSIR), Department of Biotechnology (DBT), Department of Science and Technology (DST-INSPIRE), and DBT-RA program for providing fellowships and the financial support. Authors are also thankful to the Department of Computational Biology, IIITD, New Delhi for infrastructure and facilities. The authors thank the Department of Biotechnology (DBT) for providing an infrastructure grant to the institute (grant number – BT/PR40158/BTIS/137/24/2021). The authors would like to acknowledge that the figures were created using BioRender.com.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial and non-financial interests.

BIORXIV LINK

https://www.biorxiv.org/content/10.1101/2023.01.30.526378v1

DATA AVAILABILITY STATEMENT

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

ORCID

Gajendra P. S. Raghava https://orcid.org/0000-0002-8902-2876

REFERENCES

1. Kravchenko-Balasha, N., Shin, Y. S., Sutherland, A., Levine, R. D., & Heath, J. R. (2016). Intercellular signaling through secreted proteins induces free-energy gradient-directed cell movement. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 5520–5525.
2. Gomez-Navarro, N., & Miller, E. (2016). Protein sorting at the ER–Golgi interface. *The Journal of Cell Biology*, *215*, 769–778.
3. Lopez-Verrilli, M. A., & Court, F. A. (2013). Exosomes: Mediators of communication in eukaryotes. *Biological Research*, *46*, 5–11.
4. Meldolesi, J. (2022). Unconventional protein secretion dependent on two extracellular vesicles: Exosomes and ectosomes. *Frontiers in Cell and Developmental Biology*, *10*, 877344.
5. Kuo, I.-Y., Hsieh, C.-H., Kuo, W.-T., Chang, C.-P., & Wang, Y.-C. (2022). Recent advances in conventional and unconventional vesicular secretion pathways in the tumor microenvironment. *Journal of Biomedical Science*, *29*, 56.
6. Doyle, L. M., & Wang, M. Z. (2019). Overview of extracellular vesicles, their origin, composition, purpose, and methods for exosome isolation and analysis. *Cells*, *8*, 727.
7. Han, Y., Jia, L., Zheng, Y., & Li, W. (2018). Salivary exosomes: Emerging roles in systemic disease. *International Journal of Biological Sciences*, *14*, 633–643.
8. Abels, E. R., & Breakefield, X. O. (2016). Introduction to extracellular vesicles: Biogenesis, RNA cargo selection, content, release, and uptake. *Cellular and Molecular Neurobiology*, *36*, 301–312.
9. Bellingham, S. A., Coleman, B. M., & Hill, A. F. (2012). Small RNA deep sequencing reveals a distinct miRNA signature released in exosomes from prion-infected neuronal cells. *Nucleic Acids Research*, *40*, 10937–10949.
10. Kalluri, R., & LeBleu, V. S. (2020). The biology, function, and biomedical applications of exosomes. *Science (New York, N.Y.)*, 367.
11. Huda, M. N., Nafiujjaman, M., Deaguero, I. G., Okonkwo, J., Hill, M. L., Kim, T., & Nurunnabi, M. (2021). Potential use of exosomes as diagnostic biomarkers and in targeted drug delivery: Progress in clinical and preclinical applications. *ACS Biomaterials Science & Engineering*, *7*, 2106–2149.
12. Zhou, B., Xu, K., Zheng, X., Chen, T., Wang, J., Song, Y., Shao, Y., & Zheng, S. (2020). Application of exosomes as liquid biopsy in clinical diagnosis. *Signal Transduction and Targeted Therapy*, *5*, 144.
13. Sun, B., Li, Y., Zhou, Y., Ng, T. K., Zhao, C., Gan, Q., Gu, X., & Xiang, J. (2019). Circulating exosomal CPNE3 as a diagnostic and prognostic biomarker for colorectal cancer. *Journal of Cellular Physiology*, *234*, 1416–1425.
14. Théry, C., Amigorena, S., Raposo, G., & Clayton, A. (2006). Isolation and characterization of exosomes from cell culture supernatants and biological fluids. *Current Protocols in Cell Biology*, Chapter *3*, Unit 3.22.
15. Yu, W., Hurley, J., Roberts, D., Chakrabortty, S. K., Enderle, D., Noerholm, M., Breakefield, X. O., & Skog, J. K. (2021). Exosome-based liquid biopsies in cancer: Opportunities and challenges. *Annals of Oncology : Official Journal of the European Society for Medical Oncology*, *32*, 466–477.
16. Hu, C., Jiang, W., Lv, M., Fan, S., Lu, Y., Wu, Q., & Pi, J. (2022). Potentiality of exosomal proteins as novel cancer biomarkers for liquid biopsy. *Frontiers in Immunology*, *13*, 792046.
17. Jeppesen, D. K., Nawrocki, A., Jensen, S. G., Thorsen, K., Whitehead, B., Howard, K. A., Dyrskjøt, L., Ørntoft, T. F., Larsen, M. R., & Ostenfeld, M. S. (2014). Quantitative proteomics of fractionated membrane and lumen exosome proteins from isogenic metastatic and nonmetastatic bladder cancer cells reveal differential expression of EMT factors. *Proteomics*, *14*, 699–712.
18. Poersch, A., Grassi, M. L., Carvalho, V. P. D., Lanfredi, G. P., Palma, C. D. S., Greene, L. J., De Sousa, C. B., Carrara, H. H. A., Candido Dos Reis, F. J., & Faça, V. M. (2016). A proteomic signature of ovarian cancer tumor fluid identified by highthroughput and verified by targeted proteomics. *Journal of Proteomics*, *145*, 226–236.

19. Yi, X., Chen, J., Huang, D., Feng, S., Yang, T., Li, Z., Wang, X., Zhao, M., Wu, J., & Zhong, T. (2022). Current perspectives on clinical use of exosomes as novel biomarkers for cancer diagnosis. *Frontiers in Oncology*, *12*, 966981.
20. Li, X., Corbett, A. L., Taatizadeh, E., Tasnim, N., Little, J. P., Garnis, C., Daugaard, M., Guns, E., Hoorfar, M., & Li, I. T. S. (2019). Challenges and opportunities in exosome research—Perspectives from biology, engineering, and cancer therapy. *APL Bioengineering*, *3*, 011503.
21. Zhao, L., Poschmann, G., Waldera-Lupa, D., Rafiee, N., Kollmann, M., & Stühler, K. (2019). OutCyte: A novel tool for predicting unconventional protein secretion. *Scientific Reports*, *9*, 19448.
22. Yu, L., Guo, Y., Zhang, Z., Li, Y., Li, M., Li, G., Xiong, W., & Zeng, Y. (2010). SecretP: A new method for predicting mammalian secreted proteins. *Peptides*, *31*, 574–578.
23. Kandaswamy, K. K., Pugalenthi, G., Hartmann, E., Kalies, K.-U., Möller, S., Suganthan, P. N., & Martinetz, T. (2010). SPRED: A machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes. *Biochemical and Biophysical Research Communications*, *391*, 1306–1311.
24. Bendtsen, J. D., Jensen, L. J., Blom, N., Von Heijne, G., & Brunak, S. (2004). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Engineering, Design & Selection : PEDS*, *17*, 349–356.
25. Garg, A., & Raghava, G. P. S. (2008). A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In Silico Biology*, *8*, 129–140.
26. Ras-Carmona, A., Gomez-Perosanz, M., & Reche, P. A. (2021). Prediction of unconventional protein secretion by exosomes. *BMC Bioinformatics*, *22*, 333.
27. Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L. G., … Teodoro, D. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*, D480–D489.
28. Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, *28*, 3150–3152.
29. Hasan, M. M., Khatun, M. S., & Kurata, H. (2020). iLBE for computational identification of linear B-cell epitopes by integrating sequence and evolutionary features. *Genomics, Proteomics & Bioinformatics*, *18*, 593–600.
30. Kaur, D., Arora, A., Vigneshwar, P., & Raghava, G. P. S. (2023). Prediction of peptide hormones using an ensemble of machine learning and similarity-based methods. *bioRxiv*, 05.
31. Aggarwal, S., Dhall, A., Patiyal, S., Choudhury, S., Arora, A., & Raghava, G. P. S. (2023). An ensemble method for prediction of phage-based therapy against bacterial infections. *Frontiers in Microbiology*, *14*, 1148579.
32. Mathur, M., Patiyal, S., Dhall, A., Jain, S., Tomer, R., Arora, A., & Raghava, G. P. S. (2021). Nfeature: A platform for computing features of nucleotide sequences. *BioRxiv*, 12.
33. Pande, A., Patiyal, S., Lathwal, A., Arora, C., Kaur, D., Dhall, A., Mishra, G., Kaur, H., Sharma, N., Jain, S., Usmani, S. S., Agrawal, P., Kumar, R., Kumar, V., & Raghava, G. P. S. (2022). Pfeature: A tool for computing wide range of protein features and building prediction models. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *30*(2), 204–222.
34. Kumar, M., Gromiha, M. M., & Raghava, G. P. (2007). Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics*, *8*, 463.
35. Sharma, N., Patiyal, S., Dhall, A., Pande, A., Arora, C., & Raghava, G. P. S. (2021). AlgPred 2.0: An improved method for predicting allergenic proteins and mapping of IgE epitopes. *Briefings in Bioinformatics*, *22*, bbaa294.
36. Altschul, S. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*, 3389–3402.
37. Ren, K., Zeng, Y., Cao, Z., & Zhang, Y. (2022). ID-RDRL: A deep reinforcement learning-based feature selection intrusion detection model. *Scientific Reports*, *12*, 15370.
38. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*, 14.
39. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
40. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–410.
41. Sharma, N., Naorem, L. D., Jain, S., & Raghava, G. P. S. (2022). ToxinPred2: An improved method for predicting toxicity of proteins. *Briefings in Bioinformatics*, *23*, bbac174.
42. Vens, C., Rosso, M.-N., & Danchin, E. G. J. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics (Oxford, England)*, *27*, 1231–1238.
43. Jain, S., Dhall, A., Patiyal, S., & Raghava, G. P. S. (2022). IL13Pred: A method for predicting immunoregulatory cytokine IL-13 inducing peptides. *Computers in Biology and Medicine*, *143*, 105297.
44. Dhall, A., Patiyal, S., Sharma, N., Usmani, S. S., & Raghava, G. P. S. (2021). Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Briefings in Bioinformatics*, 936–945.
45. Jiao, S., Zou, Q., Guo, H., & Shi, L. (2021). iTTCA-RF: A random forest predictor for tumor T cell antigens. *Journal of Translational Medicine*, *19*, 449.
46. Dhanda, S. K., Vir, P., & Raghava, G. P. (2013). Designing of interferon-gamma inducing MHC class-II binders. *Biology Direct*, *8*, 30.
47. Boukouris, S., & Mathivanan, S. (2015). Exosomes in bodily fluids are a highly stable resource of disease biomarkers. *Proteomics – Clinical Applications*, *9*, 358–367.

SUPPORTING INFORMATION

Additional supporting information may be found online https://doi.org/10.1002/pmic.202300231 in the Supporting Information section at the end of the article.

How to cite this article: Arora, A., Patiyal, S., Sharma, N., Devi, N. L., Kaur, D., & Raghava, G. P. S. (2024). A random forest model for predicting exosomal proteins using evolutionary information and motifs. *Proteomics*, *24*, e2300231. https://doi.org/10.1002/pmic.202300231