Development of *in silico* tools for advancing non-invasive diagnostics and therapeutics

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

By

Akanksha

PHD20208

Under the guidance of

Prof. Gajendra Pal Singh Raghava

Professor, IIIT Delhi

Department of Computational Biology

Indraprastha Institute of Information Technology, New Delhi

September 2025

Certificate

This is to certify that the thesis titled "**Development of *in silico* tools for advancing non-invasive diagnostics and therapeutics**" being submitted by **Akanksha** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulation relating to the degree.

The results contained in this thesis have not been submitted in part or in full to any other university or institute for the award of any degree/diploma.

September, 2025

Prof. Gajendra Pal Singh Raghava
Professor
Indraprastha Institute of Information Technology Delhi
New Delhi-110020

Declaration

This is to be certified that the dissertation titled "**Development of *in silico* tools for advancing non-invasive diagnostics and therapeutics**" being submitted by **Akanksha** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by me. This research work has been carried out under the supervision of Prof. **Gajendra Pal Singh Raghava**.

The results contained in this thesis have not been submitted in part or in full to any other university or institute for the award of any degree/diploma.

Akanksha
PhD Student
Roll number - PHD20208
Indraprastha Institute of Information Technology, Delhi

Acknowledgements

"Dream is not that which you see while sleeping; it is something that does not let you sleep." — Dr. A.P.J. Abdul Kalam. This thesis is the result of my dream of becoming a scientist, which kept me awake. It was a journey filled with challenges, growth, and immense learning. As I reflect upon the years of hard work and perseverance, I am deeply grateful to all those whose support, guidance, and belief made this accomplishment possible. It is with heartfelt appreciation that I acknowledge the individuals and institutions who stood beside me through every step of this journey.

I would like to express my deepest gratitude to my advisor, Prof. Gajendra Pal Singh Raghava, for his help, insightful guidance, and constant encouragement throughout the course of my PhD. His deep expertise, patience, and belief in my potential have played a pivotal role in shaping both my research and my growth as a researcher. I am especially thankful for the freedom he gave me to explore my ideas, while always being there with thoughtful advice whenever I needed direction. It was an incredibly enriching experience working under his mentorship.

I would also like to thank my Doctoral Committee members Dr. Gaurav Ahuja and Dr. Jaspreet Kaur Dhanjal for their valuable comments and inputs throughout my Ph.D., which helped improve my research work. I am extremely thankful to all the research and teaching staff at IIIT Delhi for creating a stimulating academic environment and for their valuable contributions to my learning throughout the years. I want to say thanks to Mr. Raju Biswas, Mrs. Anshu Dureja, Mrs. Shipra Jain, and Mr. Imran Khan with others in the administrative department for helping me with all the academia related work. I also want to thank all the admin facility staff, FMS, and mess staff for providing me with great facilities at IIIT Delhi. I would also like to thank Dance Fitness teacher - Mr. Zubair Rana, Yoga teacher - Mr. Ajay Saxena, and the Gym Trainer at IIIT Delhi for their extremely rejuvenating sessions, which played a vital role in helping me maintain balance, focus, and well-being during my PhD journey. I am also thankful to the IT helpdesk and other system administrator staff for the management of computer facilities at IIIT Delhi, which aided in the smooth conduct of research work.

I also gratefully acknowledge the financial support provided by Council of Scientific and Industrial Research (CSIR) which played a crucial role in supporting my research and academic endeavors throughout the course of my PhD. I also want to thank IIIT Delhi and the Academic

Affairs Committee at IIIT Delhi for providing me with an Overseas Research Fellowship (ORF) to pursue six-month internship at the University of California, Los Angeles (UCLA). I am also grateful to the School of Dentistry, UCLA, for providing me with additional funding.

I am sincerely grateful to Dr. David T.W. Wong for providing me with the opportunity to intern in his lab and for his inspiring mentorship. His visionary approach to research and encouragement throughout my time at UCLA greatly enriched my academic journey and deepened my interest in translational research. I would also like to extend my heartfelt thanks to my lab mates at the Wong Lab — Jordan Cheng, Neeti Swarup, Aida Mohammadi, and Irene Choi, whose support and friendship made my experience truly memorable. I am grateful to have found lifelong friends in each of them.

I also want to thank my colleagues and lab mates at IIIT Delhi. I am really grateful to my seniors, Dr. Neelam Sharma, Dr. Dilraj Kaur, Dr. Sumeet Patiyal, Dr. Leimarembi Devi, Dr. Anjali Dhall, Dr. Chakit Arora, Dr. Neetesh Pandey, and Mrs. Shipra Jain for their help and guidance whenever required. I would also like to thank my colleagues Dashleen, Nisha, Purva, Anshuma, Shivani, Pankaj, Naman, Saloni, Pratik, Pushpendra, Anand, Kunal, Shubham, Ritu, Nishant.

I am immensely grateful to the friends I made at IIIT Delhi – Anam Fatima, Sumedha Chugh, Samriddhi Gupta, Aseem Srivastava, Dilraj Kaur, Neelam Sharma, Dashleen Kaur, Leimarembi Devi, Prateek Paul, Kapali Suri, and GL Harika for their support. I would also like to thank my best friends – Saloni Bishnoi and Shreya Jain, who stood by me through every high and low of this journey. Their constant encouragement, heartfelt conversations, and moments of laughter became my anchor during the most challenging times.

I owe my deepest gratitude to my family for their love and encouragement. To my parents, who have always believed in me and stood by every decision with wholehearted support. I am also grateful to my younger sister Angel Arora who brings lightness and laughter into my life, even during the most stressful times. I would also like to extend my heartfelt thanks to my parents-in-law for their constant love and blessings. Their kind words always provided me comfort and reassurance.

Lastly, I would like to thank my husband, Siddharth Agrawal, who has been the strongest pillar of support throughout my Ph.D. journey. This milestone would not have been possible without

his unwavering encouragement. He kept me grounded and focused even from miles away, often putting our time together on hold so I could pursue my goals. Whether it was accompanying me to conferences and interviews, or simply reminding me of my worth when I doubted myself, his quiet strength and steadfast belief in me made all the difference. I cannot thank my husband enough for being my calm in the chaos, my constant in the uncertainty, and my greatest source of strength.

Abstract

Non-invasive diagnostics and therapeutics have the potential to revolutionize clinical practices by providing patient-friendly, accessible, and pain-free alternatives to traditional invasive methods. Among various biofluids, saliva stands out due to its easy collection and abundance of biomarkers that are reflective of systemic health. To accelerate the research on saliva-based non-invasive diagnostics and therapeutics, we built SalivaDB, a comprehensive database compiling 15,821 biomarker entries from various sources like research articles and databases. It contains information on about 201 diseases for biomarker categories like proteins, metabolites, microbes, miRNAs, and genes. SalivaDB is a user-friendly web-based platform, helping researchers by reducing research effort and time required for the discovery and validation of clinically relevant salivary biomarkers.

One significant pathway through which these biomarkers enter saliva is via exosomes, which are secreted by approximately all cell types. Exosomes carry biomolecules like RNAs, proteins, and lipids, making them highly promising for early disease detection and targeted therapeutic delivery. To understand the role of exosomal biomarkers in diagnostics and therapeutics, we developed computational tools to predict major exosomal molecules. We first developed ExoProPred, which is a tool for predicting exosomal proteins. We initially applied a similarity-search-based method using BLAST, which was ineffective due to low sequence similarity among exosomal proteins. Subsequently, we applied a motif-based approach which revealed recurrent motifs that were unique to exosomal proteins. Although this method showed high accuracy, it had limited coverage. We then employed machine learning models using compositional and evolutionary features, achieving an AUROC of 0.73. To further enhance prediction performance, we developed a hybrid approach that integrated Machine Learning (ML) with motif analysis, resulting in a significantly improved AUROC of 0.85.

After predicting exosomal proteins, we focused on miRNA which is another most commonly found biomarker in exosomes. We developed EmiRPred which is a computational tool to predict exosomal miRNAs. We built AI-based models (ML, DL, LLM) using a number of features including composition features, binary features, structural features, and structural images. We integrated these AI-based models with alignment-based approaches like (motif-search and similarity-search) to effectively predict exosomal miRNAs. This integrated ensemble model achieved an AUC of 0.73 on independent validation set. The tools ExoProPred

and EmiRPred are accessible via web server, downloadable standalone, and python packages, supporting broad usability within the research community.

In addition to predicting exosome-associated biomolecules, we also aimed to predict the molecules that are highly expressed in exosomes. It is observed that expression levels of miRNA varied significantly amongst different subcellular locations and also changed notably in the presence of disease. The identification of abundant miRNAs in exosomes is crucial for exploring their physiological roles and potential implications in disease diagnostics and therapeutics. Therefore, to provide deeper insights into baseline miRNA profiles in exosomes, we developed AdmirePred, a prediction tool to identify highly abundant miRNAs within blood exosomes. In this study, we used alignment-based methods like motif-search and similarity-search and alignment-free methods like machine learning algorithms. We leveraged a combination of both these approaches to develop a hybrid method that achieved an AUC of 0.854 on an independent validation set. AdmirePred is available as a standalone software, web server and a Python package.

To demonstrate the practical utility of extracellular saliva-based biomarkers, we investigated extracellular RNA biomarkers in saliva for diagnosing Gastric Cancer (GC). Extracellular RNA (exRNA) originates from cells and is released into body fluids through active secretion via extracellular vesicles (exosomes) or passive release during cell death. We identified different sets of biomarkers including primary and secondary biomarkers in this study. The best performing set of features was an eight-gene biomarker panel with a high validation AUC of 0.905 and an MCC of 0.770. The biomarker panel identified in our study performed better than previously discussed biomarkers in the literature. These findings underscore saliva's significant potential as a reliable and efficient diagnostic fluid for early and accurate GC detection.

Overall, we present a wide-ranging collection of curated resources, computational tools, and methods that collectively aim to advance non-invasive diagnostics and therapeutics. These tools and findings are designed to support the transition from traditional invasive procedures to more accessible, efficient, and patient-friendly diagnostic approaches.

List of Publications, Poster Presentations, and Patents

Publications

Thesis related

Published

1. **Arora, A.**, Kaur, D., Patiyal, S., Kaur, D., Tomer, R., & Raghava, G. P. (2023). SalivaDB—a comprehensive database for salivary biomarkers in humans. *Database*, *2023*, baad002.
2. **Arora, A.**, Patiyal, S., Sharma, N., Devi, N. L., Kaur, D., & Raghava, G. P. (2024). A random forest model for predicting exosomal proteins using evolutionary information and motifs. *Proteomics*, *24*(6), 2300231.
3. **Arora, A.,** & Raghava, G. P. S. (2025). Prediction of exosomal miRNA-based biomarkers for liquid biopsy. *Scientific reports, 15*(1), 31191.

Under Review

4. **Arora, A.**, & Raghava, G. P. (2025). AdmirePred: A method for predicting abundant miRNAs in Exosomes. *(Under Review)*
5. **Arora, A.**, & Raghava, G. P. (2025). Saliva-based Biomarkers for Predicting Gastric Cancer. *(Under Review)*

Others

Published

6. Kaur, D.[#], **Arora, A.[#]**, Vigneshwar, P., & Raghava, G. P. (2024). Prediction of peptide hormones using an ensemble of machine learning and similarity-based methods. *Proteomics*, 2400004.
7. Kaur, D., **Arora, A.**, Patiyal, S., & Raghava, G. P. S. (2023). Hmrbase2: a comprehensive database of hormones and their receptors. *Hormones*, *22*(3), 359-366.
8. Rathore, A. S. [#], Choudhury, S. P. [#], **Arora, A.**, Tijare, P., & Raghava, G. P. S. (2024). ToxinPred 3.0: An improved method for predicting the toxicity of peptides. Submitted to *Computers in Biology and Medicine, 179,* 108926.
9. Jarwal, A., Dhall, A., **Arora, A.**, Patiyal, S., Srivastava, A., & Raghava, G. P. (2024). A Deep Learning method for classification of HNSCC and HPV patients using single-cell transcriptomics. *Frontiers in Molecular Biosciences*, *11*, 1395721.

10. Aggarwal, S., Dhall, A., Patiyal, S., Choudhury, S., **Arora, A.**, & Raghava, G. P. (2023). An ensemble method for prediction of phage-based therapy against bacterial infections. *Frontiers in Microbiology*, *14*, 1148579.
11. Swarup, N., Cheng, J., Choi, I., Heo, Y.J., Kordi, M., Aziz, M., **Arora, A.**,.. & Wong, D. T. (2023). Multi-faceted attributes of salivary cell-free DNA as liquid biopsy biomarkers for gastric cancer detection. *Biomarker research*, *11*(1), p.90.

Under Review

12. Malik, S., Tomer, R., **Arora, A.**, & Raghava, G. P. (2025). Identification of Multiple Prognostic Biomarker Sets for Risk Stratification in SKCM. *bioRxiv*, 2025-02.
13. Srivastava, A., Dhall, A., Patiyal, S., **Arora, A.**, Jarwal, A., & Raghava, G. P. (2023). Prediction of alzheimer's disease from single cell transcriptomics using deep learning. *bioRxiv*, 2023-07.
14. Mathur, M., Patiyal, S., Dhall, A., Jain, S., Tomer, R., **Arora, A.**, & Raghava, G. P. (2021). Nfeature: A platform for computing features of nucleotide sequences. *BioRxiv*, 2021-12.
15. Jaganath, D., Sieberts, S. K., Raberahona, M., Huddart, S., Omberg, L., Rakotoarivelo, R., ... & CODA TB **DREAM Challenge Consortium**. (2024). Accelerating cough-based algorithms for pulmonary tuberculosis screening: Results from the CODA TB DREAM Challenge. *medRxiv*, 2024-05.

Poster Presentations

1. **Arora A.,** Raghava, G.P. S. (2024). Exploring Exosomal miRNA as Potential Biomarkers for Liquid Biopsy. *Machine Learning in Computational Biology*. September 5-6. University of Washington, Seattle, USA.
2. **Arora, A.**, Kaur, D., Patiyal, S., Kaur, D., Tomer, R., & Raghava, G. P. S. (2023). A database for salivary biomarkers in humans. *NIH Extracellular RNA Communication Consortium (ERCC19).* May 1-2. NIH, Bethesda, Maryland, USA.
3. Swarup, N., Choi, I., Kim, J., Cheng, J., **Arora, A.,** Aziz, A., Wong, D.T. (2024). Novel cell-free sequencing genomic data: Technique and Application of Broad Range cell-free DNA sequencing. *Cancer Biomarkers AI and Bioinformatics Workshop*. August 13-15. National Institutes of Health (NIH). California Institute of Technology (CalTech), USA.
4. McGill, C., Vanjari, G., Yu, A., Cheng, J., **Arora A.**, Kaczor-Urbanowicz, K., & Wong, D. T. W. (2023). Characterizing mitochondrial cell-free DNA in plasma as a novel liquid biopsy biomarker for oral cancer detection. *(B.I.G.)*. Aug 11. UCLA.

5. Hyun, J.C.,…, **Arora, A.**, Aberle, D., Wong, D.T. and Tang, C.Y., 2024. Abstract LB108: Specific plasma DNA end-motifs distinguish patients with lung adenocarcinoma versus benign nodules. *Cancer Research*, *84*(7_Supplement), pp.LB108-LB108.
6. Yip, S.H.,…, **Arora, A.**, Aberle, D., Wong, D.T. and Tang, C.Y., 2024. Abstract LB010: Identification of malignant lung indeterminate pulmonary nodules with Plasma WGS using exon read depth of short DNA fragments. *Cancer Research*, *84*(7_Supplement), pp.LB010-LB010.

Patents

Diagnostic System for Genetic Information Processing with Ailment Prediction Mechanism (Inventors: Jason C. Hyun, Shun Hang Yip, Raghuraman Ramamurthy, Edmund W. Wong, Cheuk Ying Tang, Neeti Swarup, Jordan C. Cheng, Irene Choi, **FNU Akanksha (also known as Akanksha Arora)**, and David Tai Wai Wong)

1. End Motif IPN (Intermediate Pulmonary Nodules), Filed: January 8, 2024
2. Exon IPN, Filed: January 8, 2024
3. End Motif OPML (Oral Pre-malignant lesions), Filed: February 2, 2024.
4. Exon OPML, Filed: February 2, 2024

Table of Contents

Abbreviations

AAC	Amino Acid Composition
Acc	Accuracy
AI	Artificial Intelligence
APAAC	Amphiphilic Pseudo Amino Acid Composition
ATC	Atom Composition
AUC or AUROC	Area under the Receiver Operating Characteristic Curve
BERT	Bidirectional Encoder Representations from Transformers
BLAST	Basic Local Alignment Search Tool
Blastn	nucleotide-nucleotide Basic Local Alignment Search Tool
Blastp	protein-protein Basic Local Alignment Search Tool
BTC	Bond Composition
CD-HIT	Cluster Database at High Identity with Tolerance
CDK	k-mer Composition
cfDNA	Cell Free Deoxyribonucleic Acid
circRNA	Circular Ribonucleic Acid
CNN	Convolutional Neural Networks
CTC	Conjoint Triad Calculation
CTD	Composition enhanced Transition Distribution
CV	Cross Validation
DAC	Dinucleotide-based Auto Correlation
DACC	Dinucleotide-based Auto Cross Correlation
DCC	Dinucleotide-based Cross Correlation
DDOR	Distance Distribution of Residue
DL	Deep Learning
DL	Deep Learning
DNA	Deoxyribonucleic Acid
DPC	Dipeptide Composition
DT	Decision Tree
e-value	Expected Value
ELISA	Enzyme-linked Immunosorbent Assay
ESCRT	Endosomal Sorting Complexes Required for Transport

ET	Extra Trees
EV	Extracellular Vesicle
exRNA	Extracellular Ribonucleic Acid
FDA	Food and Drug Administration
GB	Gradient Boosting
GC	Gastric Cancer
GEO	Gene Expression Omnibus
GNB	Gaussian Naïve Bayes
GO	Gene Ontology
HMDB	Human Metabolome Database
HTML	Hypertext Markup Language
IFS	Iterative Feature Selection
ILV	Intraluminal Vesicles
KNN	K-Nearest Neighbors
LLM	Large Language Model
lncRNA	Long non-coding Ribonucleic Acid
LR	Logistic Regression
MCC	Matthews Correlation Coefficient
MERCI	Motif EmeRging and with Classes Identification
miRNA	Micro Ribonucleic Acid
ML	Machine Learning
mRNA	Messenger Ribonucleic Acid
MVB	Multivesicular Bodies
NCBI	National Center for Biotechnology Information
ncRNA	Non-coding Ribonucleic Acid
p-value	Probability Value
PAAC	Pseudo Amino Acid Composition
PC_PDNC	Parallel Correlation Pseudo Dinucleotide Composition
PCP	Physicochemical Properties Composition
PDNC	Pseudo dinucleotide composition
piRNA	Piwi-Interacting Ribonucleic Acid
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool

QSO	Quasi-Sequence Order
RDK	Reverse Complement k-mer Composition
ResNet	Residual Network
RF	Random Forest
RFE	Recursive Feature Elimination
RNA	Ribonucleic Acid
RPM	Reads Per Million
RRI	Residue Repeat Information
SC	Stacking Classifier
SC_PDNC	Serial Correlation Pseudo Dinucleotide Composition
Sens	Sensitivity
SEP	Shannon Entropy of Protein
SER	Shannon Entropy of Residues
Sklearn	Sci-kit Learn
snoRNA	Small Nucleolar Ribonucleic Acid
SPC	Shannon Entropy of Physicochemical Property
Spec	Specificity
SVC	Support Vector Classifier
TF–IDF	Term Frequency - Inverse Document Frequency
TPC	Tripeptide Composition
tRNA	Transfer Ribonucleic Acid
VC	Voting Classifier
WHO	World Health Organization
XGB	Extreme Gradient Boosting

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The global landscape of healthcare is undergoing a transformative shift because of the increasing prevalence of chronic diseases such as cancer, diabetes, cardiovascular diseases, and neurodegenerative disorders. The World Health Organization (WHO) reports that non-communicable diseases account for around 74% of global deaths, accounting for over 41 million deaths each year. This number is expected to rise due to aging populations, environmental challenges, and lifestyle-related risk factors (World Health Organization, 2023). These conditions often develop silently, progressing to advanced stages before clinical symptoms appear, which limits the effectiveness of therapeutic interventions.

Traditional diagnostic methods such as tissue biopsies, imaging, or endoscopic procedures are widely used but have several limitations. These procedures are often invasive, costly, time-consuming, and pose risks of infection, bleeding, and sampling errors (Ludwig & Weinstein, 2005; L. Zhang et al., 2012). Moreover, regular monitoring with invasive methods is not practical for all patients, especially children, the elderly, or people in low-resource areas (Pantel & Alix-Panabières, 2013). Thus, the biomedical community is actively exploring alternative diagnostic approaches that are not only accurate and sensitive but also non-invasive, patient-friendly, and cost-effective (Cohen et al., 2018a; Siravegna et al., 2017a).

In the past few years, liquid biopsy has come out to be a revolutionary concept in non-invasive diagnostics. It involves the analysis of biological fluids such as blood, urine, saliva, and cerebrospinal fluid to detect biomarkers indicative of disease (Alix-Panabières & Pantel, 2016). Saliva has emerged as a promising diagnostic fluid among these fluids due to its non-invasive, easily accessible, and cost-effective nature (Figure 1.1). Saliva collection is painless and can be performed without needing trained personnel, making it ideal for mass screening and point-of-care testing (Malon et al., 2014). Saliva-based diagnostics are also safer in terms of biohazard risk, ease of handling, storage, and transportation.

Saliva reflects local and systemic physiological conditions, as it contains a rich array of biological molecules such as proteins, peptides, hormones, DNA, RNA, and metabolites (Javaid et al., 2016). These molecules can be valuable biomarkers for various diseases, including cancer, cardiovascular conditions, autoimmune disorders, and infectious diseases (Théry et al., 2009a). Recently, salivary biomarker detection has been explored for multiple

diseases. Saliva-based biomarkers offer a powerful alternative to conventional diagnostic methods (Nonaka & Wong, 2023).

Biomarkers can reach saliva through several pathways, with one major route being exosomes that transport biomarkers from distant diseased sites to saliva. These nanoscale extracellular vesicles (30–150 nm) are released by almost all cell types and are present in various body fluids, including saliva. Exosomes carry a rich cargo of bioactive molecules such as proteins, lipids, DNA, mRNA, and non-coding RNAs (Théry et al., 2002). They mirror the physiological and pathological conditions of their originating cells and play a role in intercellular communication.Therefore, they play a critical role in the progression of a number of diseases. Their molecular cargo is protected by a lipid bilayer, which preserves the integrity of biomarkers during circulation and sample processing. This bilayer protection enhances stability and detection sensitivity of the biomarkers (Valadi et al., 2007).

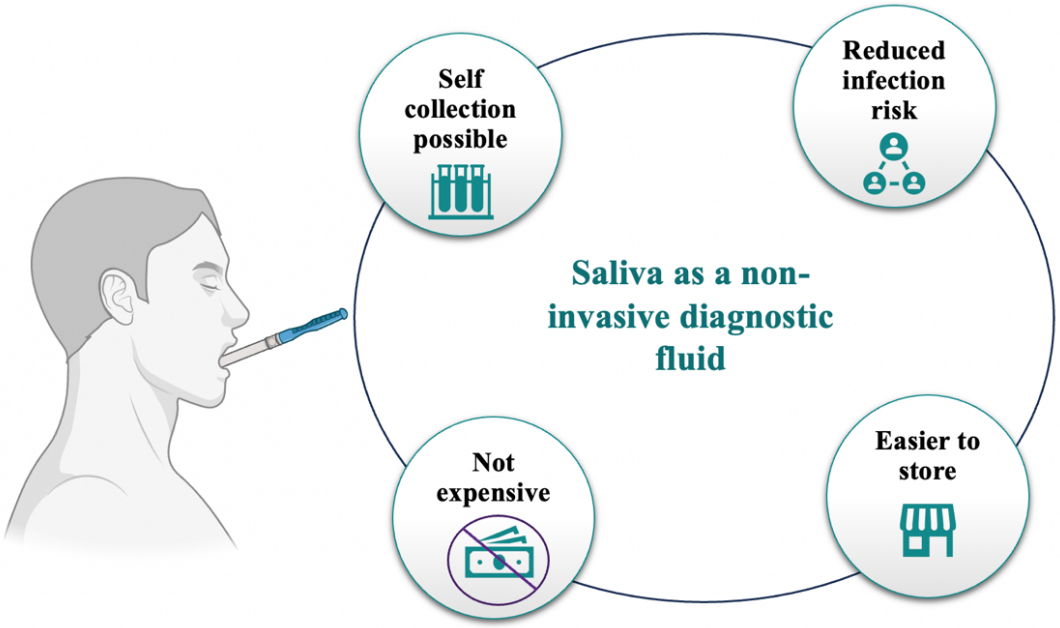


Figure 1.1: Saliva as a non-invasive medium for disease detection

While these sources (saliva and exosomes) offer a wealth of diagnostic information, it is a challenge to identify and translate their molecules into clinically usable biomarkers. The complexity of biological systems and diverse biological data require new approaches beyond traditional lab experiments. This is where *in silico* approaches like computational and statistical methods step in to improve biomedical research. These tools have proven invaluable in handling large-scale datasets, identifying patterns, and developing predictive models that aid

in biomarker discovery and disease classification (Ching et al., 2018). *In silico* tools can drastically reduce the time and cost associated with biomarker research by enabling high-throughput screening and reproducible analysis pipelines.

1.2 *In silico* Approaches in Biomedicine

In silico tools refer to computational methods that aid in understanding biological phenomena and solving biomedical problems. Over the last two decades, the explosion of omics data has necessitated using computational models to extract meaningful patterns and make predictive inferences (Manzoni et al., 2018). These approaches are critical in biomarker identification, disease classification, drug discovery, and therapeutic targeting (Alum, 2025; Blanco-González et al., 2023; Singh et al., 2023).

Development of *in silico* tools for biomedical applications offers several advantages:

- **Cost and time efficiency**: High-throughput data analysis reduces the need for exhaustive experimentation.
- **Reproducibility**: Algorithms provide consistent outputs for given inputs by minimizing variability.
- **Scalability**: Large datasets from diverse sources can be processed simultaneously.
- **Predictive power**: Machine learning models can uncover hidden associations, classify disease states, and accurately predict molecular interactions.

In silico tools have been successfully applied to various domains, including predicting cancer subtypes (M et al., 2024), screening drug-target interactions (Yamanishi et al., 2008), identifying non-coding RNAs (Panwar et al., 2014), and modeling disease networks (Barabási et al., 2011).

1.3 Pathways of Biomarker Transport to Saliva

Saliva harbors a diverse array of biomarkers that reflect both local and systemic physiological states. Several biological pathways transport molecules into the oral cavity and contribute to salivary biomarkers, such as passive diffusion, active transport, ultrafiltration, transcellular endo/exocytosis, exosome-mediated transport, and local secretions (See Table 1.1) (Javaid et al., 2016; C. Lau et al., 2013; Y.-H. Lee & Wong, 2009; Y. Li et al., 2004; Malathi et al., 2014; Taba et al., 2005; X. Zhang et al., 2015).

Table 1.1: Mechanism of transport of biomolecules to saliva

Pathway	Type of Biomarker	Source	Examples	References
Passive Diffusion	Small, non-polar molecules	Blood → Saliva	Cortisol, urea, ethanol, nicotine metabolites	(Y.-H. Lee & Wong, 2009)
Active Transport	Ions, glucose, small peptides	Blood → Salivary glands	Glucose, sodium, potassium, lactate	(Malathi et al., 2014)
Ultrafiltration	Low-molecular-weight solutes	Blood plasma → Saliva	Creatinine, electrolytes, some drugs	(Javaid et al., 2016)
Transcellular (Endo/Exo)	Macromolecules, cfDNA, cytokines	Blood → Acinar cells → Saliva	cfDNA fragments (e.g., TP53, EGFR), IL-6, TNF-α	(X. Zhang et al., 2015)
Exosome-Mediated	miRNA, mRNA, lncRNA, cfDNA, proteins	Circulating exosomes	miR-21, miR-31, mutated KRAS DNA, mRNA signatures	(C. Lau et al., 2013)
Gingival Crevicular Fluid (GCF)	Immune markers, microbial DNA/RNA	Gums/plasma leakage	IL-1β, CRP, bacterial DNA, neutrophil elastase	(Taba et al., 2005)
Local Secretion (Oral cells)	Host RNA, miRNA, proteins, microbial DNA	Oral mucosa, immune cells	miR-125a, defensins, oral microbiota DNA, cytokeratins	(Y. Li et al., 2004)

1.4 Exosome-mediated Transport

Among all the mechanisms listed above, the exosome-mediated pathway is the most promising for biomarker collection due to its biological stability and specificity. These vesicles can cross the blood-salivary gland barrier and be released into the salivary duct system either directly or after uptake by glandular cells. The lipid bilayer of exosomes protects their molecular contents from enzymatic degradation, enabling the reliable detection of low-abundance but highly informative biomarkers in saliva (C. Lau et al., 2013; X. Zhang et al., 2015). In addition, exosomes are free from many of the other salivary contaminants (Nonaka & Wong, 2023). The mechanism of transport of biomarkers to saliva via exosomes is shown in Figure 1.2.

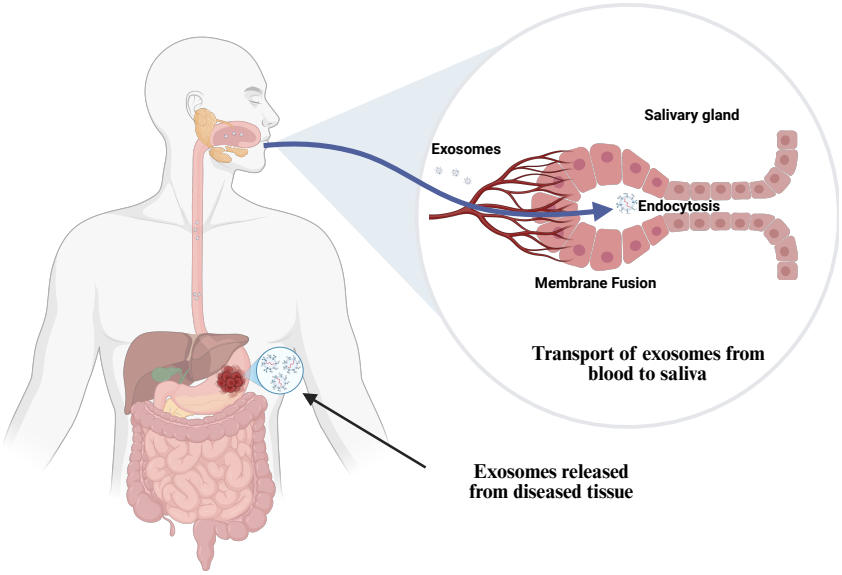


Figure 1.2: Mechanism of transport of biomarkers to saliva via exosomes

1.5 Biogenesis of Exosomes

Exosome biogenesis begins with the inward budding of the endosomal membrane leading to the formation of intraluminal vesicles (ILVs) within multivesicular bodies (MVBs). These MVBs either fuse with lysosomes for degradation or with the plasma membrane to release ILVs as exosomes into the extracellular space (Colombo et al., 2014; Kowal et al., 2014).

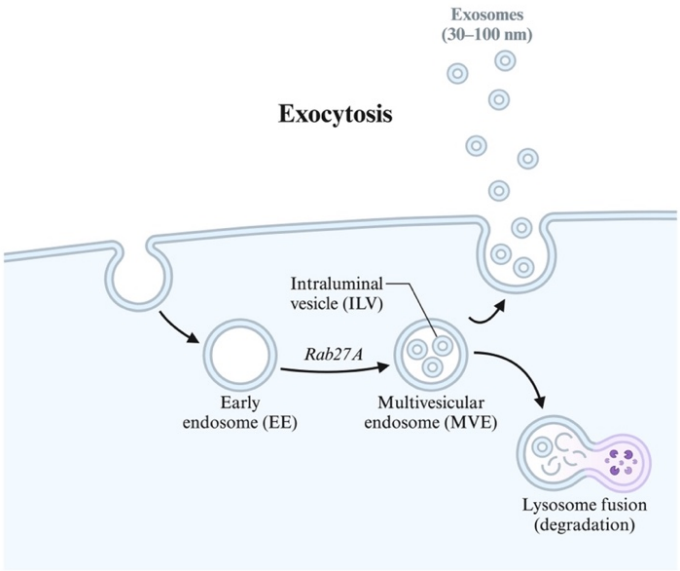


Figure 1.3: The mechanism of biogenesis of exosomes

The formation of ILVs can occur via ESCRT independent or dependent pathways. ESCRT-dependent pathways involve key proteins such as TSG101 and ALIX, whereas ESCRT-independent pathways rely on tetraspanins like CD63, CD81, and CD9, and lipid rafts enriched in ceramides (van Niel et al., 2018). Rab GTPases, such as Rab27a and Rab27b, are involved in the trafficking and secretion of MVBs to the plasma membrane (Ostrowski et al., 2010). Exosome biogenesis is influenced by various cellular states, including stress, hypoxia, and oncogenic transformation. This process plays a vital role in intercellular communication by transferring proteins, lipids, mRNAs, and miRNAs (Théry et al., 2009b). The method of biogenesis of exosomes is illustrated in Figure 1.3.

1.6 Composition of Exosomes

The exosomes have unique and rich molecular composition that reflect their original cell type which makes them potential reservoirs of disease biomarkers. The lipid composition of exosomes is distinct from the plasma membrane and includes high cholesterol levels, sphingomyelin, ceramide, and phosphatidylserine. These lipids confer rigidity and aid vesicle formation and fusion (Simons & Raposo, 2009). The protein content in exosomes is diverse yet enriched in specific classes. Common marker proteins include tetraspanins (CD9, CD63, CD81), which are involved in membrane organization and targeting; Alix and TSG101, which are part of the ESCRT complex involved in MVB sorting; and heat shock proteins (HSP70, HSP90), which contribute to protein folding and cellular stress responses (Kowal et al., 2016; Théry et al., 2002). Exosomes may also carry enzymes, signaling molecules, cytoskeletal proteins, and cell-type-specific markers, such as MHC molecules in immune cells or integrins in tumor-derived vesicles. The nucleic acid cargo includes a variety of RNA species like messenger RNAs (mRNAs), microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and circular RNAs (circRNAs), which can be functionally transferred to recipient cells to modulate gene expression (O'Brien et al., 2020a; Valadi et al., 2007). Exosomes also harbor double-stranded genomic DNA (dsDNA) that spans all chromosomes, potentially reflecting the mutational landscape of the parent cell and serving as a source of biomarkers for cancer diagnostics (Thakur et al., 2014). The molecular cargo of exosomes is not random but highly regulated and influenced by the cell's state and microenvironment (Jiang et al., 2022; Z. Wang et al., 2023). This selective enrichment and stability of contents make exosomes a promising tool for biomarker discovery and therapeutic delivery. The key components of exosomes are shown in Figure 1.4.

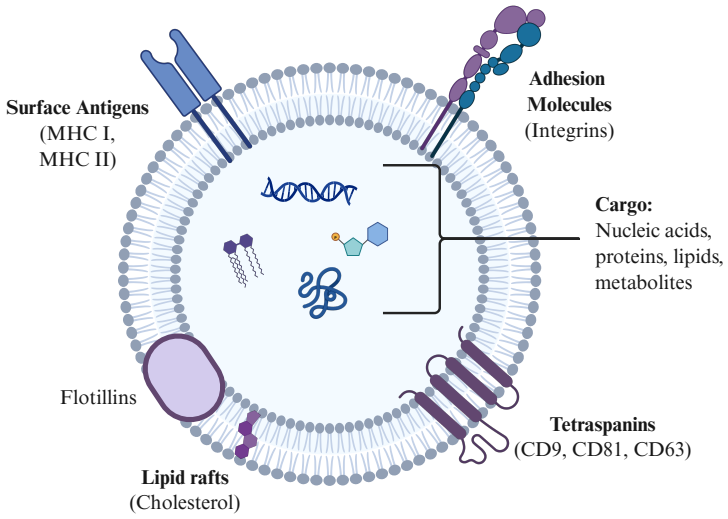


Figure 1.4: The key components of exosomes

1.7 Proposal's Origin

Non-invasive diagnostics and therapeutics are transforming modern healthcare by offering safer, more accessible, and patient-friendly alternatives to conventional invasive procedures. These approaches enable the detection and monitoring of diseases without the need for surgical intervention or complex sampling methods such as biopsies or endoscopies. While significant progress has been made in discovery of blood- and tissue-based biomarker, saliva remains underexplored. Unlike blood or tissue collection, collecting saliva is painless, doesn't need special equipment or trained staff, and is ideal for frequent, large-scale, or even at-home testing. Despite its simplicity, saliva is a molecularly rich fluid containing DNA, RNA, proteins, metabolites, and microbiome-derived components. It is capable of reflecting both local and systemic health conditions. Recent studies have demonstrated its potential in detecting various diseases, including cancers, cardiovascular disorders, metabolic conditions, and infections. To accelerate biomarker discovery in this domain, *in silico* approaches can be leveraged to identify salivary biomarkers associated with complex human diseases. One key mechanism through which biomarkers are transported into saliva is via exosomes, which are extracellular vesicles secreted by various cell types. These vesicles carry a rich molecular cargo comprising proteins, lipids, DNA, mRNA, microRNAs, and other non-coding RNAs. They carry information on the state of their original cell. While exosomes offer tremendous diagnostic and therapeutic potential, the molecular mechanisms governing the selective packaging of biomolecules into exosomes remain poorly understood. This presents a significant opportunity for further research.

1.8 Objectives of the thesis

The primary objective of this thesis is to develop a suite of *in silico* tools to advance non-invasive diagnostics and therapeutics. The study is structured into two major parts: saliva-based tools and exosome-based tools. The first part aimed to systematically compile and curate salivary biomarkers into a publicly accessible database called SalivaDB. This was followed by the development of artificial intelligence (AI) models for the prediction of gastric cancer (GC) using salivary biomarkers. The second part aimed to identify the molecules that are likely to be packaged into exosomes and function as disease biomarkers. We focused on computational prediction of exosomal proteins and microRNAs (miRNAs), where we developed tools like ExoProPred, EmiRPred, and AdmirePred. The outline of the thesis is shown in Figure 1.5. Altogether, these tools aim to facilitate biomarker discovery, reduce reliance on invasive diagnostic procedures, and support early disease detection through accessible biofluids. Hence, we have framed the following objectives:

- A comprehensive database for salivary biomarkers in humans
- A computational method for predicting exosomal proteins using motifs and evolutionary information
- A method for the prediction and design of exosomal miRNA-based biomarkers for liquid biopsy
- An in silico method for predicting abundant miRNAs in exosomes using miRNA expression
- An AI-Based method for predicting Gastric Cancer using extracellular RNA expression in Saliva

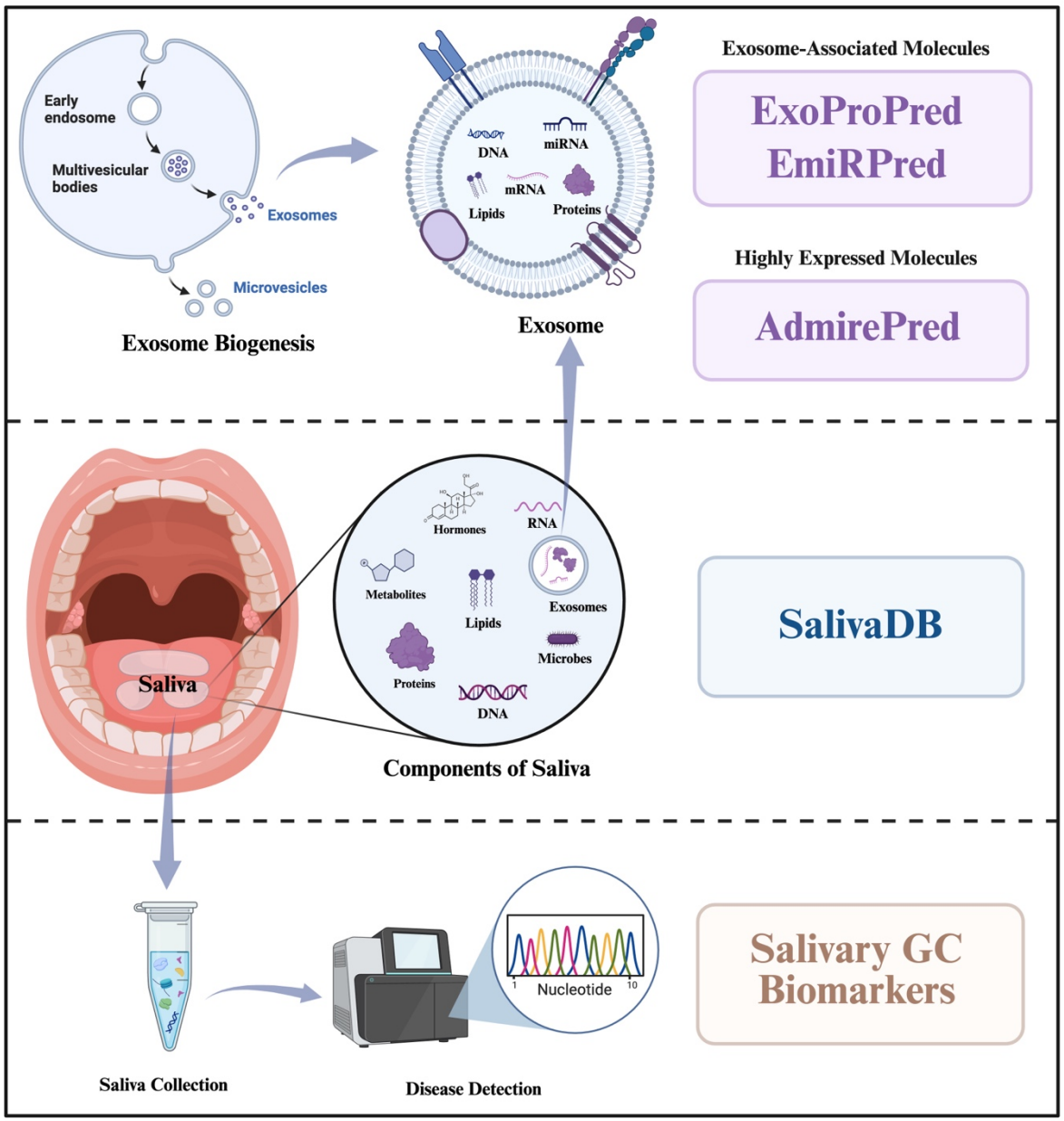


Figure 1.5: The workflow followed in the study

1.9 Organization of the chapters

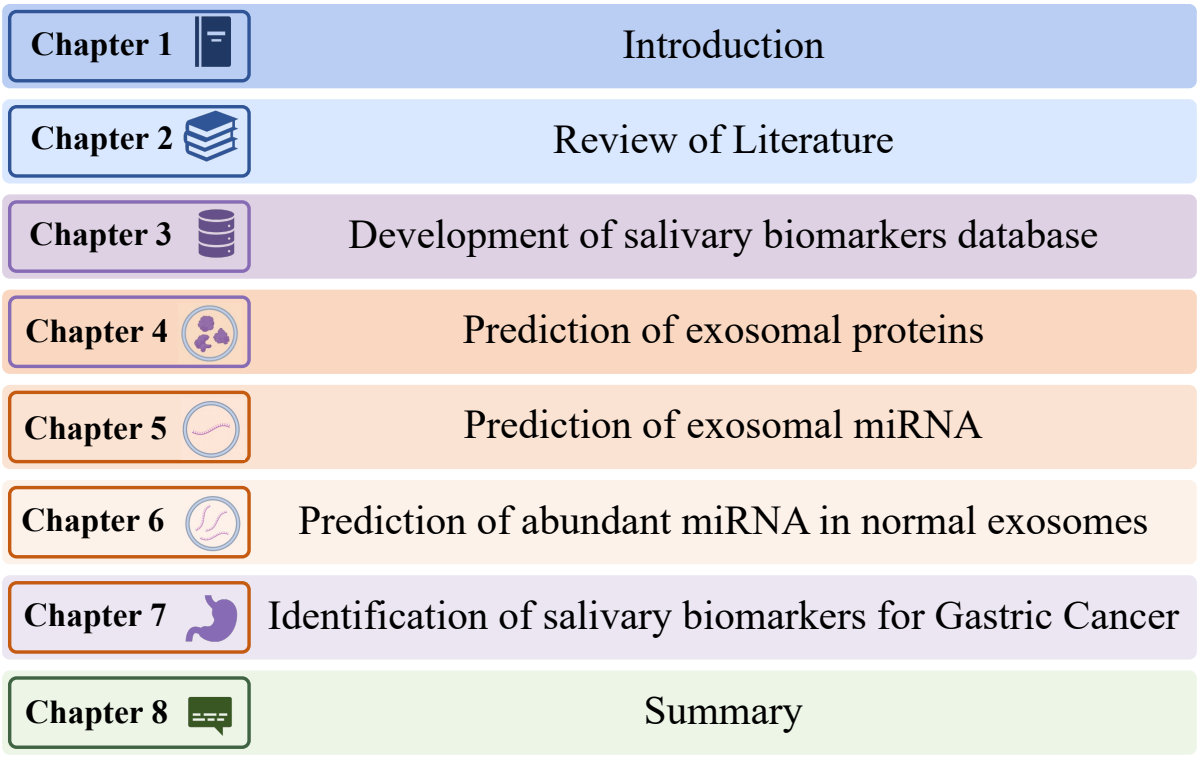


Figure 1.6: The organization of thesis and title of the chapters

Chapter 1: This chapter introduces the growing need for non-invasive, accessible diagnostics and therapeutics. It underscores the potential of saliva as a non-invasive medium and highlights the need for developing saliva-based tools. The chapter further discusses the biological pathways through which biomarkers reach saliva, with an emphasis on exosome-mediated transfer. It delves into the biogenesis and molecular composition of exosomes. It also addresses the current knowledge gap regarding the selective packaging of biomolecules within exosomes. The chapter concludes by outlining the structure of the thesis and the key objectives addressed in this study.

Chapter 2: This chapter presents an in-depth overview of the existing literature relevant to non-invasive biomarker discovery with a focus on saliva-based and exosome-based existing techniques and tools. It begins by outlining the clinical relevance of saliva-based biomarkers, listing known markers and available databases. The chapter then discusses the diagnostic and therapeutic potential of exosomes and traditional methods used to identify exosomal biomarkers. Further, it explores existing databases and *in silico* tools for predicting exosomal proteins and miRNAs.

Chapter 3: This chapter focuses on the first objective of the study where we developed a database of salivary biomarkers in humans. This study presents "SalivaDB" - a comprehensive, manually curated database developed to consolidate diverse salivary biomarker information from published literature and existing resources to a single platform. It integrates over 15,821 entries across five major biomolecule categories, including proteins, metabolites, microbes, miRNAs, and genes. The chapter details the methodology used for data collection, curation, and integration. It describes the technical architecture and features of the web-based platform of SalivaDB, including advanced search, browse, and sequence similarity search tools.

Chapter 4: This chapter describes the second objective, where we developed a computational method for predicting exosomal proteins using motifs and evolutionary information. This method can be used to classify proteins into exosomal and non-exosomal. A web server, a standalone and Python package "ExoProPred," has been developed to enable large-scale identification of exosomal proteins, facilitating research in exosome-based diagnostics and therapeutics.

Chapter 5: This chapter describes the third objective of the thesis, which corresponds to a computational method for the prediction and design of exosomal miRNA-based biomarkers for liquid biopsy. A web server, standalone, and a Python package called "EmiRPred" has been developed, which is a computational framework developed to predict exosomal microRNAs (miRNAs). The chapter describes alignment-based techniques, deep learning, machine learning, and large language models used in this study to distinguish exosomal from non-exosomal miRNAs using sequence, structure, and compositional features.

Chapter 6: This chapter introduces the fourth objective, where an *in silico* method is developed for predicting abundant miRNAs in exosomes using miRNA expression. A web server, standalone, and a Python package called "AdmirePred" have been built, which predicts microRNAs (miRNAs) that are highly abundant in blood-derived exosomes under normal conditions. This chapter describes the integration of alignment-based approaches with machine learning models trained on a range of sequence-derived features to develop this method.

Chapter 7: This chapter discusses the fifth objective, where an AI-based method is developed for predicting Gastric Cancer using extracellular RNA (exRNA) expression in Saliva. It describes the potential of using saliva as a non-invasive source for diagnosing Gastric Cancer. The study identifies a set of biomarkers that can accurately distinguish cancer patients from

healthy individuals using statistical methods and AI-based models developed on salivary exRNA expression data.

Chapter 8: This chapter presents a comprehensive summary of the work and provides a comprehensive view of the thesis. It explains the contribution of this work in the area of non-invasive diagnostics and therapeutics.

Chapter 2

Review of Literature

2.1 Overview of Non-Invasive Technologies

Modern diagnostics are increasingly focusing on non-invasive technologies to minimize patient discomfort and risk. Non-invasive methods avoid surgical biopsies or other intrusive procedures, relying instead on approaches like medical imaging and liquid biopsies. For example, imaging modalities such as MRI and ultrasound can detect internal abnormalities without incision. Similarly, liquid biopsies analyze body fluids (blood, saliva, urine, etc.) for disease indicators, offering a safer alternative to tissue biopsies (Heitzer et al., 2019). These approaches enable longitudinal monitoring, which means patients can be sampled repeatedly over time. This is valuable for tracking patient disease progression or treatment response (Wan et al., 2017). Overall, non-invasive diagnostics improve patient compliance and allow earlier detection of diseases by reducing the need for painful or risky procedures.

The circulating biomarkers in readily available biofluids represent a promising avenue in non-invasive diagnostics. The most common are blood-based tests (e.g., for circulating tumor DNA or proteins) (Cohen et al., 2018b). However, the routine phlebotomy performed to extract blood has some drawbacks, such as the need for a trained phlebotomist, risk of infection, and storage issues. The researchers are now exploring truly non-invasive biofluids like saliva and urine. Among these, saliva has gained substantial attention as a “mirror of the body” that can reflect systemic health (Kaczor-Urbanowicz et al., 2017)(Siravegna et al., 2017b). Saliva is emerging as an attractive diagnostic fluid in biomedical research and clinical practice. Its collection is fast, easy, inexpensive, and non-invasive, requiring only spitting or swabbing the mouth (Yoshizawa et al., 2013a). Despite the ease of collection, saliva contains many biomolecules including proteins, DNA, various forms of RNA, metabolites, and microbes (Wong, 2012).

2.2. Saliva-Based Biomarkers

In recent years, the field of salivaomics has developed to characterize the molecular constituents of saliva systematically. The salivaomics concept was introduced around 2008 and includes salivary proteome, transcriptome, microRNAome, metabolome, and microbiome (Shah, 2018). The researchers have identified numerous salivary biomarkers across these categories that correlate with disease states. For example, changes in specific salivary proteins (such as cytokines or enzymes) can indicate inflammation or cancer. At the same time, the presence of specific microbial species in saliva might signal oral or systemic infections (Belstrøm et al., 2017; Sivadasan et al., 2020). The advances in technology have enabled the

detection of circulating tumor DNA and tumor-specific mutations in saliva, extending salivary diagnostics beyond local diseases. A platform called EFIRM (electric field–induced release and measurement) was even shown to detect actionable EGFR mutations from lung tumors via saliva with high accuracy(Wei et al., 2014). The role of saliva-based biomarkers in healthcare is thus increasingly important, and a growing body of literature supports their validity for disease diagnostics (Surdu et al., 2025).

2.2.1 Known Saliva-Based Biomarkers

A multitude of specific biomarkers have been identified in saliva for various diseases. These biomarkers span different molecular types, including proteins (e.g., cytokines, enzymes, antibodies), nucleic acids (mRNAs, microRNAs, DNA), and metabolites. Saliva contains biomarkers for both localized diseases (e.g., periodontal disease, oral cancer) and systemic illnesses (e.g., diabetes, neurodegenerative diseases, cancers in distant organs). Inflammatory cytokines like IL-6, IL-8, and IL-1β are frequently elevated in saliva during oral infections or mucosal inflammation, making them useful indicators of conditions such as periodontitis or oral cancer (Yakob et al., 2014).

Table 2.1: FDA-approved saliva-based diagnostics

Disease/Condition	Biomarker(s) Detected	Company/Developer	Status	References
HIV/AIDS	HIV-1/2 antibodies in oral fluid	OraSure Technologies (OraQuick)	FDA-approved (OTC home test)	(Reynolds & Muwonga, 2004)
COVID-19 (SARS CoV 2)	Viral RNA (via RT-PCR on saliva)	SalivaDirect (Yale)	FDA Emergency Use Authorization	(Vogels et al., 2020)
Oral & Throat Cancer	Human and microbial RNA signature in saliva	Viome (CancerDetect Oral & Throat)	FDA Breakthrough Device (clinical validation – not yet approved)	(Banavar et al., 2023)
Cushing's Syndrome (Hypercortisolism)	Cortisol hormone level in saliva (late-night sample)	Salimetrics, Pantex (salivary cortisol ELISA kits)	FDA-cleared assay (diagnostic use since 2001)	(Nieman, 2018)

Metabolic markers such as certain amino acids have been found in altered concentrations in the saliva of patients with cancers and metabolic diseases (Sugimoto et al., 2010). Hormones like cortisol can be measured in saliva to assess stress axis activity or endocrine disorders (Hellhammer et al., 2009). Salivary microRNAs have gained attention as well, such as miR-21, which is a well-known oncomiR whose elevated levels in saliva have been linked to various cancers (Gai et al., 2018), and miR-146a/miR-155 are implicated in oral inflammatory conditions and malignancies (Menini et al., 2021). The presence of tumor-related DNA (such as HPV DNA or tumor mutation DNA) in saliva also enables non-invasive molecular diagnosis of head and neck cancers (Y. Wang et al., 2015) Table 2.1 shows the FDA-approved saliva-based diagnostic tests.

2.2.2 Available Databases for Salivary Biomarkers

Several databases like SalivaTecDB, Human Salivary Proteome Wiki, CancerPDF, and The Human Metabolome Database (HMDB) contain scattered information about salivary biomarkers found in humans. These are discussed below:

- **SalivaTecDB:** A database of salivary biomarkers that contains information on protein, miRNA, and microbial disease biomarkers found in saliva. It was last updated in 2017 (Arrais et al., 2013a; Rosa et al., 2012).
- **Human Salivary Proteome Wiki:** A resource maintained by the NIH that specifically catalogs salivary proteins (and their isoforms) identified in human saliva. It curates proteins detected in saliva, mapping them to gene/protein identifiers. However, it is limited to proteomic data and does not cover other biomarker types (W. W. Lau et al., 2021a).
- **HMDB (Human Metabolome Database):** HMDB contains data on metabolites found in humans, and also includes entries for metabolites detected in human saliva (Wishart et al., 2018a).
- **CancerPDF:** A manually curated database that contains information on cancer-related peptides found in various body fluids, including saliva. It includes information on mass, precursor protein, techniques for profiling and quantification of peptides (Bhalla, Verma, et al., 2017a).

These resources contain scattered information on salivary biomarkers (See Table 2.2). There is no comprehensive database that contains information on all types of saliva-based biomarkers found in humans and the diseases associated with them.

Table 2.2: Available databases for salivary biomarkers

Database	Contents (Saliva)	Website
Human Salivary Proteome Wiki	Proteins	https://www.salivaryproteome.org/
SalivaTecDB	miRNA, Microbes, Proteins	http://salivatec.viseu.ucp.pt/salivatec-db/
HMDB	Metabolites	https://www.hmdb.ca/
CancerPDF	Cancer Peptides	http://crdd.osdd.net/raghava/cancerpdf/

2.3 Exosomal Biomarkers

Exosomes play an important role in non-invasive diagnostics. Exosomes are nano-sized vesicles (approximately 30–150 nm in diameter) actively secreted by nearly all cell types as part of normal physiology and disease processes (Théry et al., 2018). Exosomes form when endosomal membranes invaginate to create multivesicular bodies, which then fuse with the plasma membrane to release the internal vesicles as exosomes (Colombo et al., 2014). Due to this biogenesis pathway, exosomes encapsulate a selective cargo of biomolecules from their parent cell, including proteins (cytosolic and membrane proteins), nucleic acids (mRNAs, microRNAs, DNA fragments), lipids, and metabolites (Mathieu et al., 2019). Exosome cargo reflects the pathological state of their cells of origin, making them valuable for disease detection. For example, tumor-derived exosomes contain oncogenic proteins (e.g., EGFR, PD-L1) and nucleic acids that serve as cancer biomarkers (Tkach & Théry, 2016). In neurodegenerative diseases, exosomes transport misfolded proteins like α-synuclein or tau, offering insights into disease progression (Thompson et al., 2016). The stability of exosomes in biofluids (e.g., blood, saliva, urine) further enhances their diagnostic utility (Yáñez-Mó et al., 2015).

2.3.1 Diagnostic Potential of Exosomes

Exosomes are considered powerful tools for diagnostics, often described as a “liquid biopsy” component that enables minimally invasive disease detection. They are found in all major biological fluids (blood plasma, saliva, cerebrospinal fluid, urine etc.), which makes them readily obtainable. Exosomes allow disease monitoring over time through simple fluid draws

as patients' fluids can be sampled longitudinally (B. Zhou et al., 2020). Several features mentioned below indicate the diagnostic potential of exosomes:

- **Rich Molecular Cargo:** Exosomes encapsulate a rich and complex cargo that reflects their cell of origin. This includes proteins from the cytosol and cell membrane, various RNA species (mRNAs, microRNAs, long non-coding RNAs), DNA fragments, and lipids. Such diversity means a single exosome sample can provide information on numerous biomarkers simultaneously. For instance, an exosome from a tumor cell might simultaneously contain a mutant DNA fragment, an oncogenic protein, and a signature microRNA, offering multiple pieces of evidence for the presence of cancer (Kalluri & LeBleu, 2020).
- **Disease-Specific Signatures:** Research has shown that exosomal content changes during disease (Kourembanas, 2015). Certain diseases have been the focus of developing exosome-based diagnostics, including cardiovascular diseases (Barile et al., 2016), central nervous system disorders (Budnik et al., 2016), and many cancers (Becker et al., 2016). For example, in cancer patients, circulating exosomes often carry tumor-associated markers that are scarcely present in healthy individuals (Melo et al., 2014). The levels of exosomes can also be higher in disease states (Lässer et al., 2011). This specificity allows exosome analysis to distinguish patients from healthy controls and even provide prognostic information (Skog et al., 2008).
- **Exosomal DNA for Mutational Analysis:** Some studies indicate that exosomes contain small amounts of double-stranded DNA representing genomic fragments of the parent cell (Thakur et al., 2014). This has enabled detection of cancer mutations (e.g., KRAS or TP53 mutations) in exosomal DNA from patient blood samples (Kahlert et al., 2014). It is unclear how much DNA exosomes carry, but if exosomal DNA includes slightly larger fragments than cell-free DNA, it could improve mutation detection sensitivity for several cancers (K. S. Yang et al., 2017).
- **Exosomal microRNAs:** MicroRNAs in exosomes are particularly promising diagnostic biomarkers (O'Brien et al., 2020b). Many studies have identified oncogenic and tumor-suppressor miRNAs that are differentially enriched in exosomes from cancer patients. For example, exosomal miR-21 is observed to be dysregulated in ovarian, bladder, and prostate cancer (Foj et al., 2017; Rhim et al., 2022; Samsonov et al., 2016; F.-K. Yang et al., 2024). Other oncogenic miRNAs frequently found at higher levels in tumor-derived exosomes include miR-155, the miR-17-92 cluster, and miR-1246,

observed across multiple malignancies (Otmani et al., 2024)(W. (Jess) Li et al., 2021; Salehi et al., 2024a).

- **Exosomal Protein Markers:** The surface and internal proteins can also be used for diagnostics. One notable example is exosomal Glypican-1 (GPC1), a cell surface proteoglycan reported to be enriched on cancer-cell-derived exosomes (Buscail et al., 2019; Melo et al., 2015). Multiple researchers have found that GPC1-positive exosomes are elevated in patients with pancreatic cancer (and to some extent breast and colon cancer), distinguishing them from healthy individuals (Frampton et al., 2018; K. S. Yang et al., 2017). High GPC1 exosome levels may also correlate with tumor burden and drop after therapy, suggesting their use in monitoring treatment response (Buscail et al., 2019; Melo et al., 2015). Similarly, exosomal proteins such as heat-shock proteins or tumor antigens have been explored (Clayton et al., 2005; Whiteside, 2016). In breast cancer, exosome phosphoproteins have been detected and might carry diagnostic information about tumor signaling activity (I.-H. Chen et al., 2017). Another strategy has been immunocapture of disease-specific exosomes, e.g., using antibodies against a colorectal cancer antigen (CD147) to extract tumor exosomes from blood (Tian et al., 2018; Yoshioka et al., 2014). These approaches leverage the idea that if an exosome carries a unique surface marker of diseased tissue, it can be isolated and quantified (L. Doyle & Wang, 2019; Whiteside, 2016).
- **Combined Multi-Marker Approaches:** The multi-omic data from exosomes can be extracted to create better diagnostic tests. To capture a holistic picture of disease, a multi-component exosomal assay might measure a set of miRNAs together with a set of proteins and possibly lipid or metabolite content. This could enhance specificity and sensitivity, as suggested by initial studies (Kalluri & LeBleu, 2020). The protective lipid bilayer of exosomes also preserves enzyme-sensitive biomarkers (like specific peptides or RNAs) during transit, meaning those markers can be reliably measured.

In summary, exosomes provide a rich, multitarget platform for diagnostics. Their diagnostic potential has been demonstrated in many contexts. Whether it is identifying a cancer early through a panel of exosomal miRNAs, or tracking neurodegenerative disease progression via exosomal protein changes in cerebrospinal fluid. Ongoing research and clinical trials are translating these findings into practical diagnostic tests. It is important to note that, as of now, there are no fully FDA-approved exosome-based diagnostic tests. The FDA has granted Breakthrough Device Designation to certain exosome-based diagnostics, acknowledging their

potential clinical benefits and facilitating expedited review processes. The list of these diagnostic tests is given in Table 2.3.

Table 2.3: Exosome-based diagnostic tests under clinical trials

Biofluid (Exosome Source)	Target Disease/ Condition	Biomarkers	Detection Method	Stage	Company/ Developer	References
Urine	Prostate cancer (high-grade)	miR-21, PCA3, TMPRSS2:ERG	RT-qPCR, NanoString	FDA Breakthrough Device; CLIA-approved (ExoDx Prostate Test)	Exosome Diagnostics (Bio-Techne)	(Mitchell et al., 2008; Yekula et al., 2020)
Blood (Plasma)	Pancreatic cancer (early detection)	Glypican-1 (GPC1)	Flow cytometry, immunoblotting	FDA Breakthrough Device designated (Exo-PDAC)	Biological Dynamics	(Melo et al., 2015)
Blood (Plasma)	Colorectal cancer	miR-21, miR-1246, CD147	RT-qPCR, immunoaffinity capture	Clinical validation study	Exosome Diagnostics (Bio-Techne)	(W. Li et al., 2017)
Urine	Bladder cancer	miR-21	RT-qPCR	Clinical validation study	Exosome Diagnostics (Bio-Techne)	(F. Yang et al., 2024)
Urine	Kidney transplant rejection	IL32, B2M, CXCL11, PGK1)	RT-qPCR	Phase 1 clinical study (pilot) – ExoTRU	Brigham & Women's/Exosome Diagnostics	(El Fekih et al., 2021)

2.3.2 Therapeutic Potential of Exosomes

Beyond diagnostics, exosomes also hold promise as therapeutic agents or delivery vehicles in a variety of diseases. They naturally facilitate cell communication and are distributed well in the body, making them promising for therapy. Several aspects of the therapeutic potential of exosomes are noteworthy:

- **Intrinsic Therapeutic Effects:** Some exosomes, by themselves, can have therapeutic effects. For example, exosomes derived from mesenchymal stem cells (MSCs) have been shown to have anti-inflammatory and regenerative properties (Arslan et al., 2013; Kordelas et al., 2014). In one clinical case, repeated infusions of MSC exosomes were given to a patient with graft-versus-host disease and resulted in improvement without significant side effects (Kordelas et al., 2014). Exosomes are small and membrane-bound, so they typically evade immediate immune clearance and do not trigger strong immune reactions like whole cells or viral vectors (Arslan et al., 2013).
- **Drug Delivery Vehicles:** Exosomes are being actively explored as natural nanocarriers for drug delivery. In contrast to artificial liposomes or nanoparticles, exosomes have an innate ability to be taken up efficiently by recipient cells and can circulate with minimal immune clearance (Kamerkar et al., 2017; Wiklander et al., 2015). Studies in mice have shown that injected exosomes can deliver functional cargo (like drugs or nucleic acids) to target tissues with greater efficiency and stability than comparable synthetic liposomes (Kamerkar et al., 2017; Tian et al., 2014).
- **Nucleic Acid Delivery (siRNA/miRNA):** Exosomal transfer of RNA is a natural mechanism, and scientists are opting for it to deliver therapeutic RNA molecules like siRNA or miRNA mimics. One advantage is that RNA inside exosomes is protected from degradation by RNases in the blood (Hung & Leonard, 2015). Exosomes also exhibit prolonged circulation times relative to many liposomal formulations, aiding delivery to distant sites (Kalluri & LeBleu, 2020). Preclinical studies have reported promising results: for instance, exosomes loaded with a let-7a microRNA mimic and modified with an anti-EGFR peptide successfully targeted breast cancer cells and suppressed tumor growth in mice (Ohno et al., 2013). Another study used exosomes from MSCs to deliver miR-146b to glioma cells in the brain by engineering the exosomes with a targeting peptide. They crossed the blood-brain barrier and reduced tumor growth in a rat glioma model (Katakowski et al., 2013). A landmark effort from Kalluri et al. was made where they developed iExosomes which are exosomes loaded with a KRASG12D siRNA (Kamerkar et al., 2017). These exosomes are designed to treat pancreatic cancer carrying this specific mutation. In multiple animal models, systemic administration of these siRNA-loaded exosomes led to significant tumor shrinkage and improved survival without toxicity. Mechanistically, part of their success was attributed to exosomal CD47 (which provides a "don't eat me" signal to phagocytes, allowing exosomes to avoid macrophage clearance) and the propensity of cancer cells

to take up exosomes by macropinocytosis (Kamerkar et al., 2017). This work has progressed to a Phase I clinical trial testing exosome-based siRNA therapy in pancreatic cancer patients, underscoring the translational potential (Kamerkar et al., 2022).

- **Immune Modulation and Vaccines:** Exosomes can also carry antigens and immune modulators, suggesting uses in immunotherapy or as vaccines. For example, dendritic cell-derived exosomes naturally carry MHC-peptide complexes that can stimulate T-cells. Clinical trials have tested dendritic cell exosomes pulsed with tumor antigens as cancer vaccines to provoke an anti-tumor immune response (Besse et al., 2016; Pitt et al., 2014). Additional studies have shown that exosomes from tumor cells can be engineered to enhance their immunogenicity, potentially serving as personalized cancer vaccines (Dai et al., 2008). The ability of exosomes to transfer tumor antigens and immune-stimulating molecules between cells makes them particularly promising for immunotherapy applications (Robbins & Morelli, 2014).
- **Tissue Targeting and Engineering:** Scientists can direct exosomes to particular cell types or organs by engineering exosomal surface proteins. Peptides such as RGD (which binds specific integrins) or RVG (binds neuronal acetylcholine receptors) can be displayed on exosome surfaces to guide them to tumors or brain tissue, respectively (Alvarez-Erviti et al., 2011; Tian et al., 2014). There is also exploration into ligand enrichment on exosomes, e.g., loading exosomal membranes with antibodies or aptamers that recognize a tissue-specific marker to achieve targeted therapy (Kooijmans et al., 2016). The natural membrane composition of exosomes can fuse with target cell membranes or undergo endocytosis via various pathways (clathrin-mediated, caveolae, macropinocytosis, etc.), ensuring that the therapeutic cargo is efficiently delivered inside cells (Mulcahy et al., 2014). This multifaceted uptake capability gives exosomes an edge over many synthetic vectors (Luan et al., 2017).

A comprehensive list of the exosome-based therapeutics that are being developed and are under clinical trials has been given in Table 2.4.

Table 2.4: Exosome-based therapeutics under clinical trials

Disease/ Condition	Exosome Source	Therapeutic Mechanism	Trial Stage	Company/ Developer	References
Acute Respiratory Distress Syndrome (ARDS)	Bone marrow MSC-derived exosomes	Natural EV cargo downregulates inflammation and promotes repair	Phase III (EXTINGUISH trial, NCT05354141); RMAT & FDA Fast Track	Direct Biologics (ExoFlo)	(Lightner et al., 2023)
Dystrophic Epidermolysis Bullosa (DEB)	Bone marrow MSC-derived EVs (AGLE-102)	Delivers type VII collagen (COL7 protein + COL7A1 mRNA) to patient cells, promoting wound healing and reducing inflammation	Phase 1/2a trial (NCT04173650)	Aegle Therapeutics	(Aegle Therapeutics, 2024)
Duchenne Muscular Dystrophy (DMD)	Cardiosphere-derived cells (secreting exosomes)	Exosomes from infused cells modulate immunity and promote muscle regeneration	Phase III (HOPE-3 trial, NCT05126758)	Capricor Therapeutics (CAP-1002)	(McDonald et al., 2022)
Atopic Dermatitis	Stimulated stem cell-derived exosomes (BRE-AD01)	Multi-modal: suppresses Th2 inflammation, modulates IL-31 receptor, and repairs skin barrier	Phase 1 trial initiated (IND cleared 2022, trial pending)	Brexogen (BRE-AD01)	(Bader, 2022)
Inflammatory Diseases (acute & chronic)	Engineered exosomes carrying super-repressor (ILB-202)	Intracellular NF-κB inhibition – exosomal delivery of IκBα prevents inflammatory signal transduction	Phase 1 completed (healthy volunteers)	ILIAS Biologics (ILB-202)	(Hyun et al., 2025)

Acute Ischemic Stroke	Neural stem cell-derived exosomes (AB126)	Crosses blood–brain barrier; modulates CNS inflammasome with anti-inflammatory and neuroprotective effects (promotes neuroregeneration)	IND cleared - Phase 1b/2a starting 2024 (NCT06303375)	Aruna Bio (AB126)	(Aruna Bio, 2024)

2.3.3 Exosomal Biomarkers Databases

Several specialized databases have been established that compile identified exosomal components (proteins, mRNA, miRNA, etc) from numerous studies (See Table 3.5). The key databases include:

- **ExoCarta:** Exocarta is a manually curated source of exosomal proteins, RNAs, and lipids identified in various experiments (Mathivanan et al., 2012). ExoCarta has data from published proteomics and genomics studies of exosomes.
- **Vesiclepedia:** Vesiclepedia expands the ExoCarta concept to include all extracellular vesicles (not just exosomes) (Pathan et al., 2019). It is a community-driven database that merges information on exosomes, microvesicles, apoptotic bodies, etc. Vesiclepedia contains entries for tens of thousands of proteins, RNAs, and lipids detected in extracellular vesicles, along with the source information (organism, cell type, vesicle isolation method).
- **EVpedia:** EVpedia is another database introduced to host extracellular vesicle data (D. Kim et al., 2013). It provides an integrated platform for EV proteins, lipids, mRNA, miRNA, and metabolites from multiple studies, along with tools for functional enrichment analysis.
- **ExoRBase:** ExoRBase is a database of circulating exosomal RNA (including mRNAs, circRNAs, and miRNAs). It compiles RNA profiles from healthy individuals and cancer patients, allowing users to query if a certain RNA is present in exosomes and whether its levels differ in cancer (H. Lai et al., 2022)
- **miRandola:** miRandola is a database specializing in extracellular circulating RNAs (Russo et al., 2018). It includes miRNAs found in various extracellular forms, such as

being carried by vesicles, associated with RNA-binding proteins, or lipoproteins. Users can filter for exosome-associated miRNAs in miRandola.

- **EVmiRNA:** EVmiRNA lists all microRNAs identified in EVs from different studies, often annotating the source and method (T. Liu et al., 2019). These can be useful for identifying candidate miRNA biomarkers and comparing their prevalence across conditions.
- **RNALocate:** RNALocate database is a repository for subcellular locations of different types of RNA, like mRNA, miRNA, lncRNA, tRNA, snoRNA, snRNA, rRNA, ncRNA, circRNA, vRNA, and piRNA (Cui et al., 2022; "RNALocate: A Resource for RNA Subcellular Localizations," 2016; L. Wu et al., 2025). It includes information from about 177 subcellular locations in about 242 species, with one of the locations being "exosome".
- **ExoBCD:** It contains information on exosomal mRNA, miRNA, and lncRNA linked to Breast Cancer (X. Wang et al., 2021). It includes information on experimental biology for biomarkers, gene expression patterns, tumour stage, overall survival, functional evidence, and clinical use.
- **EV-ADD:** EV-ADD (Extracellular Vesicle-Associated DNA Database) is a comprehensive, manually curated repository that compiles data on extracellular vesicle-associated DNA derived from bodily fluids. It contains information on EV isolation methods, characterization techniques, DNA isolation procedures, fragment sizes, starting material volumes, gene names, and disease contexts (Tsering et al., 2022).

Table 2.5: Major databases containing information for extracellular vesicles (exosomes) biomarkers

Database	Contents	Website	Update
Exocarta	mRNA, miRNA, lncRNA, tRNA, snoRNA, snRNA, rRNA, ncRNA, Protein, Lipid	http://www.exocarta.org/	2015
EVPedia	Protein, mRNA, miRNA, Lipid, Metabolite	https://evpedia.info/evpedia2_xe	2015
miRandola	miRNA, lncRNA, circRNA	http://mirandola.iit.cnr.it/	2017
EVmiRNA	miRNA	https://guolab.wchscu.cn/EVmiRNA/	2019
ExoBCD	mRNA, miRNA, lncRNA in Breast Cancer	https://exobcd.liumwei.org/	2021

Exorbase 2.0	mRNA, miRNA, circRNA	https://www.exorbase.org/	2022
EV-ADD	DNA	https://www.evdnadatabase.com/	2022
Vesiclepedia	miRNA, DNA, mRNA, Protein, Lipid, Metabolite	http://www.microvesicles.org/	2024
RNAlocate V3.0	mRNA, miRNA, lncRNA, tRNA, snoRNA, snRNA, rRNA, ncRNA, circRNA, vRNA, piRNA	http://www.rnalocate.org/	2024

2.3.4. Exosomal Cargo Prediction Tools

The experimental methods for identifying exosomal cargo are complemented by *in silico* tools that aim to predict or prioritize which proteins and RNAs are likely to be sorted into exosomes. Several notable tools and approaches include:

Proteins: For proteins, ExoPred is a bioinformatic tool specifically developed to predict proteins secreted via exosomes (Ras-Carmona et al., 2021). ExoPred achieved an AUROC of 0.73 for training and 0.84 for the independent validation set for distinguishing exosomal vs non-exosomal proteins. Apart from ExoPred, there are a few general tools that predict non-classical protein secretion which comprises exosomal protein prediction among others. These include SecretomeP (Bendtsen et al., 2004), SPRED (Kandaswamy et al., 2010a), SecretP (L. Yu et al., 2010), OutCyte (Zhao et al., 2019), and ASPIRER (X. Wang et al., 2022). These tools were not specifically trained on exosomes, but on proteins known to be secreted unconventionally (via vesicles or direct translocation).

miRNA: For miRNA, there are no specific tools for the prediction of exosomal miRNA. However, some tools predict the subcellular location of miRNA, and "exosome" is one of the locations in most of them. These tools include MiRLoc (Xu et al., 2022), LocmiRNA (H. Wang et al., 2021), MiRLocPredictor (Asim et al., 2020), EL-RMLocnet (Asim et al., 2022), DAmiRLocGNet (Bai et al., 2023), MGFmiRNAloc (takes miRNA SMILES) (Y. Liang et al., 2024).

mRNA: Similar to miRNA, for mRNA, there are no tools that particularly predict exosomal mRNA. The available ones only predict mRNA's subcellular location, including "exosome" as one of the locations. These include DM3Loc (D. Wang et al., 2021), mRNALoc (Garg et al., 2020), LocmRNA (H. Wang et al., 2021), mLoc-mRNA(Meher, Rai, et al., 2021), Mulstack (Z. Liu et al., 2024), Clarion (Bi et al., 2022), MRSLPred (Choudhury et al., 2024), etc.

lncRNA: There are some tools also developed for predicting subcellular location of lncRNA - lncLocator 2.0 (Lin et al., 2021), GM-lncLoc (Cai et al., 2023), LncLocation (Feng et al., 2020), LncLocFormer (Zeng et al., 2023), SGCL-LncLoc (M. Li et al., 2024), iLoc-lncRNA-BERT (Z.-Y. Zhang et al., 2024), that predict "Exosome" as one of the locations. The details of the mentioned tools are given in Table 2.6.

Table 2.6: Major prediction tools for predicting exosomal molecules

S. No.	Methods	Biomolecule	Model	Locations	Website	YOP
1	**ExoPred**	Protein	ML	Exosome	http://imath.med.ucm.es/exopred/	2021
2	**SecretomeP 2.0**	Protein	ML	Non-classical Secreted Proteins	https://services.healthtech.dtu.dk/services/SecretomeP-2.0/	2005
3	**SPRED**	Protein	ML	Non-classical Secreted Proteins	https://www.inb.uni-luebeck.de/tools-demos/spred/spred	2010
4	**SecretP**	Protein	ML	Non-classical Secreted Proteins	http://cic.scu.edu.cn/bioinformatics/secretp/index.html	2010
5	**Outcyte**	Protein	DL	Unconventional Secreted Proteins	http://www.outcyte.com/	2019
6	**Aspirer**	Protein	DL	Non-classical secreted proteins	https://github.com/yanwu20/ASPIRER	2022
7	**DM3Loc**	mRNA	DL	Nucleus, Exosome, Cytosol, Ribosome, Membrane, ER	http://dm3loc.lin-group.cn/	2021
8	**mRNALoc**	mRNA	ML	Extracellular Region, ER, Cytoplasm, Mitochondria, Nucleus	http://proteininformatics.org/mkumar/mrnaloc	2020
9	**mLoc-mRNA**	mRNA	ML	cytoplasm, cytosol, endoplasmic reticulum, exosome, mitochondrion, nucleus, pseudopodium, posterior and ribosome	http://cabgrid.res.in:8080/mlocmrna/	2021
10	**MulStack**	mRNA	ML/DL	Nucleus, Exosome, Cytosol, Ribosome, Membrane, ER	http://bliulab.net/MulStack	2024

11	**Clarion**	mRNA	ML	chromatin, cytoplasm, cytosol, exosome, membrane, nucleolus, nucleoplasm, nucleus and ribosome	https://monash.bioweb.cloud.edu.au/Clarion/	2022
12	**MRSLPred**	mRNA	ML + motif	ribosome, cytosol, endoplasmic reticulum (ER), membrane, nucleus, and exosome	https://webs.iiitd.edu.in/raghava/mrslpred/	2024
13	**miRNALoc**	miRNA	ML	Axon, Circulating, Cytoplasm, Exosome, Extracellular Vesicle, Microvesicle, Mitochondria, Nucleus	http://cabgrid.res.in:8080/mirnaloc/	2020
14	**MiRLocPredictor**	miRNA	DL	Exosome, Cytoplasm, Mitochondria, Microvesicle, Circulating, Nucleus	https://github.com/muas16/MirLocPredictor	2020
15	**MiRLoc**	miRNA	Network based	Exosome, Cytoplasm, Mitochondria, Microvesicle, Nucleolus, Nucleus, Extracellular Vesicle	https://github.com/xumingmin/MiRLoc	2022
16	**iLoclncRNA**	lncRNA	ML	Nucleus, Cytosol, Cytoplasm, Exosome, Nucleolus	http://lin-group.cn/server/iLoc-LncRNA/	2018
17	**lncLocation**	lncRNA	ML + DL	Cytoplasm, Ribosome, Exosome, Nucleus	https://github.com/FengSY-JLU/Core-lncLocation/	2020
18	**lncLocator**	lncRNA	NLP + ML	Cytoplasm, Nucleus, Cytosol, Ribosome and Exosome	http://www.csbio.sjtu.edu.cn/bioinf/lncLocator/	2018
19	**Locate-R**	lncRNA	ML	Cytoplasm, Ribosome, Exosome, Nucleus	http://locate-r.azurewebsites.net/	2020

20	**GM-lncLoc**	lncRNA	Graph NN + Meta learning	Nucleus, Cytoplasm, Cytosol, Ribosome, Exosome	https://github.com/JunzheCai/GM-lncLoc	2023
21	**SGCL-LncLoc**	lncRNA	DL	Nucleus, Cytosol, Cytoplasm, Exosome	http://csuligroup.com:8000/SGCL-LncLoc	2024
22	**iLoc-lncRNA-BERT**	lncRNA	DL	7 classes including Exosome	https://github.com/ZhaoyueZhang/iLoc-lncRNA-BERT	2024

Chapter 3

Development of Database for Salivary Biomarkers

3.1 Introduction

Patients often undergo numerous invasive and painful procedures to diagnose a range of medical conditions. Techniques such as repeated blood draws, biopsies, and lumbar punctures add to the stress and discomfort of the diagnostic experience (S. V. Ahmed et al., 2006; Poggio et al., 2020). This has created an urgent need for diagnostic methods that are non-invasive, cost-effective, accurate, and time-efficient. Saliva is a bodily fluid that can serve as a non-invasive source of biomarkers for detecting numerous diseases (Malamud, 2011). Human saliva is a clear liquid with a pH of approximately 6.0–7.0, which comprises about 99% water and 1% organic and inorganic constituents (Kubala et al., 2018). Approximately 97% of saliva volume is produced by the major salivary glands (submandibular, sublingual, and parotid), with the rest of ~3% coming from the minor glands (lingual, labial, buccal, and palatal) (Iorgulescu, 2009; Llena-Puy, 2006). Saliva has many functions, including enabling taste, aiding swallowing and digestion, lubricating the oral cavity, and protecting against pathogens such as bacteria (Vila et al., 2019). The salivary glands are surrounded by capillaries, which permit exchange of molecules from blood into the acinar cells that produce saliva (See Figure 3.1) (M. G. Lee et al., 2012).

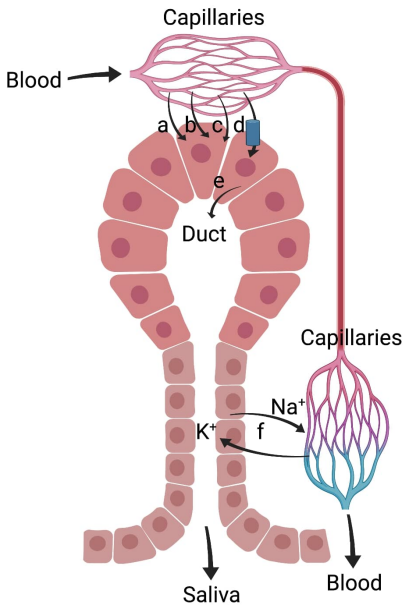


Figure 3.1: Acinar cells and mechanism of transport of molecules from blood to salivary gland. a: Passive diffusion of fat-soluble compounds. b: Transport of molecules by simple diffusion. c: Entry of water and active pumping of sodium ions. d: Active transport. e: Secretion by salivary glands into the duct. f: Pumping of sodium ions into blood

In recent years, studies have demonstrated that saliva contains a diverse array of biomolecules that can serve as biomarkers for disease (Roi et al., 2019; Yoshizawa et al., 2013a). Various blood-borne molecules such as antibodies, enzymes, growth factors, and hormones can enter the saliva through transcellular (passive or active transport) or paracellular (extracellular ultrafiltration) pathways (Yoshizawa et al., 2013b). These salivary constituents can act as circulating disease biomarkers and often carry enough information to determine a patient's disease status (Condrat et al., 2020). Saliva-based biomarkers offer several advantages over blood, which is traditionally the most common biofluid used for disease diagnosis. Collecting saliva samples is non-invasive and low-cost, carries a lower risk of pathogen transmission, and is generally easier to store than blood (Campo et al., 2006; Vaz et al., 2020).

Furthermore, saliva collection does not typically require highly trained personnel and can even be self-administered by patients. Numerous studies have demonstrated that salivary biomarkers can accurately detect a variety of diseases (F. Cheng et al., 2015; Sugimoto et al., 2010). For example, a saliva-based enzyme-linked immunosorbent assay (ELISA) can be utilized to test patients for human immunodeficiency virus types 1 and 2 (HIV-1 and HIV-2) (Yoshizawa et al., 2013a). Another study investigating breast cancer salivary biomarkers found that a panel of eight messenger RNA (mRNA) biomarkers combined with one protein biomarker achieved a diagnostic accuracy of 92%, with 97% specificity and 83% sensitivity (L. Zhang et al., 2010). Overall, using saliva as a diagnostic and prognostic medium can save millions of lives and significantly reduce healthcare costs.

In recent years, multiple efforts have been made to aggregate information related to saliva-based biomarkers. Several major resources currently provide information on salivary components, including the SalivaTecDB, Human Salivary Proteome Wiki, CancerPDF, and The Human Metabolome Database (HMDB) (Arrais et al., 2013b; Bhalla, Verma, et al., 2017b; W. W. Lau et al., 2021b; Rosa et al., 2012; Wishart et al., 2018b). Each of these resources focuses on specific types of data. However, none of the existing databases offer up-to-date information encompassing all five categories of saliva-based disease biomarkers. To address this gap and complement the existing resources, we developed a new SalivaDB database, which compiles updated information on all types of saliva-based biomarkers. SalivaDB offers detailed, manually curated entries from published research papers, SalivaTecDB, and HMDB databases.

3.2 Methods

3.2.1 Data collection

To gather all relevant literature on salivary biomarkers, we systematically searched research articles in PubMed using various keyword combinations. We used specific search terms for each biomarker category. For example, we used “Salivary Biomarker AND Protein” to find protein-related articles, “Salivary Biomarker AND miRNA” for microRNAs, and “Salivary Biomarker AND (microbe OR microbiota OR microorganism)” for microbial biomarkers. We limited the searches for proteins, miRNAs, and microbes to publication years 2017 through 2022, as SalivaTecDB contains information for these biomarkers until 2017. However, it does not contain any information on genes and metabolites. Hence, the queries “Salivary Biomarker AND gene” for genes and “Salivary Biomarker AND metabolite” for metabolites were applied to all years up to February 3, 2022. All searches were performed across all fields in PubMed. These queries returned 1,423 publications for proteins, 187 for miRNAs, 111 for microbes, 1,274 for genes, and 170 for metabolites.

In addition to the direct results from PubMed, we examined the reference lists of each relevant paper to identify any additional studies, thereby ensuring comprehensive coverage of salivary biomarker literature. After excluding review articles, non-English papers, and other irrelevant records, we identified a final set of 478 publications that contained data on salivary biomarkers. Additionally, we incorporated the data from existing databases such as SalivaTecDB which contains information for protein, miRNA and microbes until the year 2017 and HMDB which contains information about metabolites. The entries from these databases without experimental validation or a missing PubMed reference were removed. This integration ensured that our compiled dataset was comprehensive and up-to-date.

3.2.2 Database architecture

The SalivaDB web application is hosted on an Apache HTTP Server (v2.4.7) with a MySQL database (v5.5.62) at the back end for data management. The front-end interface was developed using HTML5, PHP5, CSS3, and JavaScript, allowing a responsive design compatible with desktops, tablets, and mobile devices. We implemented the web interface using a combination of Perl and PHP. The overall content and architecture of SalivaDB are given in Figure 3.2.

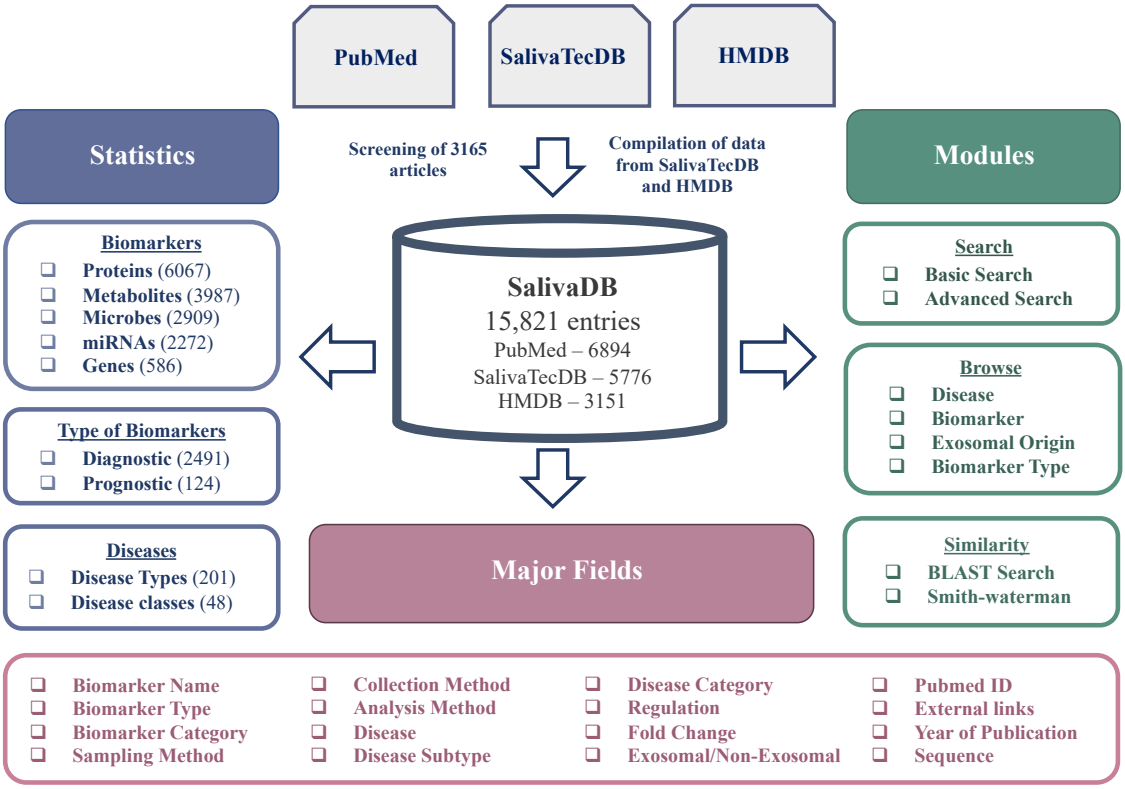


Figure 3.2: The content and architecture of SalivaDB

3.2.3 Database content

SalivaDB encompasses experimentally validated information on proteomic, genetic, transcriptomic, microbial, and metabolomic biomarkers across a broad range of diseases. The information for each biomarker entry in SalivaDB is organized into several key fields. The key fields include:

1. **Biomarker Name** – The identifier or name of the salivary biomarker as reported in the literature.
2. **Biomarker Type** – The category of the biomarker (e.g., metabolite, protein, microbe, gene, or miRNA).
3. **Sampling Method** – Details about the study's sample cohort, including the number of diseased and control samples.
4. **Collection Method** – The procedure used to collect the saliva sample from subjects.
5. **Analysis Method** – The experimental techniques used to analyze and validate the biomarker (for example, ELISA or liquid chromatography–tandem mass spectrometry).
6. **Disease Type** – The disease or condition in which the biomarker was identified.

7. **Disease Subtype** – A more specific subtype of the disease, if provided in the source.
8. **Disease Category** – The broader disease category (for instance, neurological disorder, cancer, etc.).
9. **Fold Change** – The magnitude of change (fold increase or decrease) in the biomarker's level reported in the condition vs. control.
10. **Regulation Status** – Indicates whether the biomarker was found to be upregulated or downregulated in the disease condition.
11. **External Links** – Hyperlinks to relevant external databases (such as UniProt for proteins, PubChem for metabolites, miRBase for miRNA, NCBI Taxonomy for microbes, and NCBI Gene for genes) for additional information on the biomarker.
12. **Sequence Information** – Sequence data for the biomarker: amino acid or nucleotide sequences for proteins, genes, and miRNAs; taxonomic identifiers for microbes; and SMILES strings for metabolite structures.

Each SalivaDB entry is also linked to its source publication via a PubMed ID, and includes references to external database identifiers for cross-reference, like UniProt, PubChem, NCBI Gene, NCBI Taxonomy, and miRBase for proteins, metabolites, genes, microbes, and miRNA, respectively (Brown et al., 2015; S. Kim et al., 2016; Kozomara et al., 2019; Schoch et al., 2020; UniProt Consortium, 2021).

3.3 Results

3.3.1 Data analysis

SalivaDB currently holds a total of 15,821 biomarker entries. Out of these, 6,894 entries were manually curated from 478 PubMed-indexed articles, 5,776 entries were imported from SalivaTecDB, and 3,151 entries were obtained from HMDB. SalivaTecDB's dataset covers salivary biomarkers (proteins, miRNAs, and microbes) up to 2017. To update beyond 2017, we collected additional research articles on these three biomarker types published between 1 January 2017 and 3 February 2022, as SalivaTecDB did not specify the exact cutoff date in 2017. The entries from SalivaTecDB were manually checked and the records with a lot of missing information were removed. Furthermore, information on metabolites and gene biomarkers was compiled from all relevant papers available up to 3 February 2022.

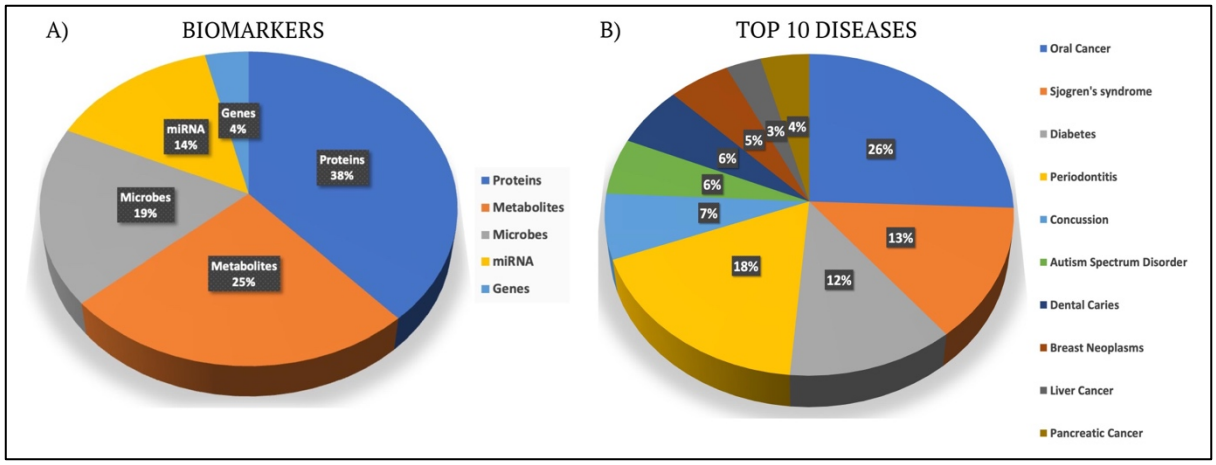


Figure 3.3: Distribution of: A: Types of biomarkers in SalivaDB; B: Top 10 diseases in SalivaDB

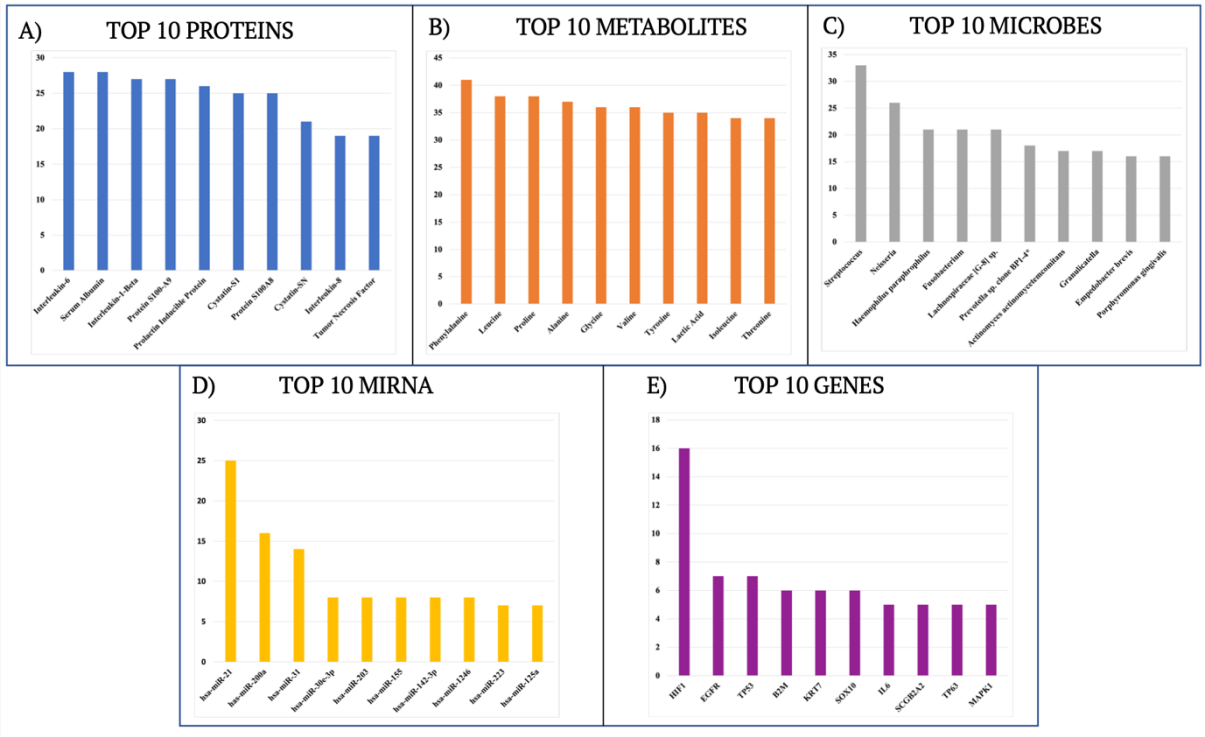


Figure 3.4: Statistics of the top 10 entries present in SalivaDB for each category – A) Proteins, B) Metabolites, C) Microbes, D) miRNA, and E) Genes

In summary, SalivaDB comprises 6,067 protein entries (38% of the database), 3,987 metabolite entries (25%), 2,909 microbial entries (19%), 2,272 miRNA entries (14%), and 586 gene entries (4%). These biomarkers are associated with 201 distinct diseases, grouped into 48 broader disease categories. The entries correspond to 7,729 unique salivary biomarkers. The database includes 2,491 biomarkers classified as diagnostic, 124 as prognostic, and 159 that serve as both diagnostic and prognostic. Additionally, SalivaDB catalogs 742 salivary

biomarkers that have been identified in exosomes. A large proportion of the salivary biomarkers in the database have been reported in conditions such as periodontitis, oral cancer, Sjögren's syndrome, diabetes, and concussion, among others. Figure 3.3 presents the breakdown of entries by biomarker type and the top 10 diseases in the database. Figure 3.4 shows the top 10 biomarkers of each type represented in SalivaDB (by number of entries).

3.3.2 Web interface

The SalivaDB web interface provides three major modules for users to access and analyze the data. These modules are discussed below:

a) Browse

The users can browse the database using the following options:

- **Biomarker Category:** Select a biomarker category (gene, protein, miRNA, metabolite, or microbe).
- **Biomarker Type:** Filter by biomarker designation as diagnostic, prognostic, or both.
- **Disease Category:** Choose from various disease categories to find associated salivary biomarkers.
- **Exosomal Origin:** See which biomarkers were derived from exosomes in saliva

These browsing options generate a list of salivary biomarkers with detailed information for each biomarker (including PubMed IDs, experimental details, etc.).

b) Search

SalivaDB also offers both basic and advanced search modules to locate specific information. In the basic search, users can query the database by fields such as biomarker category, biomarker name, biomarker type, disease name, disease category, disease sub-category, and exosomal origin. The basic search results can be tailored by adjusting the query parameters. The advanced search allows combining multiple search criteria using Boolean operators (AND, OR, NOT), enabling more complex and refined queries.

c) Similarity search

A sequence similarity search feature is provided to find database entries similar to a query sequence. SalivaDB integrates the Basic Local Alignment Search Tool (BLAST) and the Smith–Waterman algorithm for this purpose (Altschul et al., 1990; Smith & Waterman, 1981). The users can submit a protein, gene, or miRNA sequence in FASTA format (with either default or custom parameters), and the server will perform a BLAST search against the stored sequences. The Smith–Waterman algorithm can likewise be applied to find local sequence alignments, offering an alternative method to identify similar proteins, genes, or miRNAs in the database.

3.3.3 Utility of SalivaDB

SalivaDB's utility lies in being an integrated one-stop platform for obtaining detailed information on salivary biomarkers. For example, a researcher interested in the microRNA miR-1246 can enter "miR-1246" in the search bar and select the desired fields to display. Upon searching, the interface will return all entries for miR-1246; the user can then click on a specific entry's identifier (Sal_ID) to view comprehensive details about that entry. The detail page includes information about the experimental context and hyperlinks to external resources (e.g., PubMed, PubChem, UniProt, NCBI Gene, miRBase) relevant to that biomarker. The sequence data for the biomarker is also provided when applicable.

The users can further refine their searches using advanced options—applying Boolean operators (AND, OR, NOT) to combine criteria—and they can take advantage of the browsing filters (by disease, biomarker category, exosomal origin, etc.) described above. Additionally, the similarity search feature allows users to find database entries similar to a query protein, miRNA, or gene sequence, facilitating comparative analyses.

3.3.4 Comparison with existing resources

There are various existing databases for salivary biomarkers, such as the Human Salivary Proteome Wiki, HMDB, SalivaTecDB, and CancerPDF. However, none of these provide complete information on all types of salivary biomarkers. The Human Salivary Proteome Wiki is a database containing information on salivary proteins curated from UniProt and is maintained by the U.S. National Institutes of Health (NIH). SalivaTecDB provides manually curated data from journal articles only up to 2017 for proteins, miRNAs, and microbes. HMDB contains information on salivary metabolites. CancerPDF contains information on cancer-

associated peptides found in biofluids, including saliva. SalivaDB distinguishes itself by covering all five major categories of salivary biomarkers (genes, proteins, miRNAs, metabolites, and microbes), filling the gaps left by these earlier resources. Along with this, each entry in our resource is mapped to disease context, biomarker type (diagnostic/prognostic, exosomal/non-exosomal), experimental technique, and validation status. This level of disease-centric annotation is less consistently available in existing repositories. Table 3.1 and Figure 3.5 compare SalivaDB with the existing salivary biomarker databases, highlighting features such as the number of entries per biomarker type, last update year, and use of manual curation. SalivaDB is a freely accessible, comprehensive repository of salivary biomarker literature, which can help accelerate the discovery and development of non-invasive diagnostic methods for human diseases.

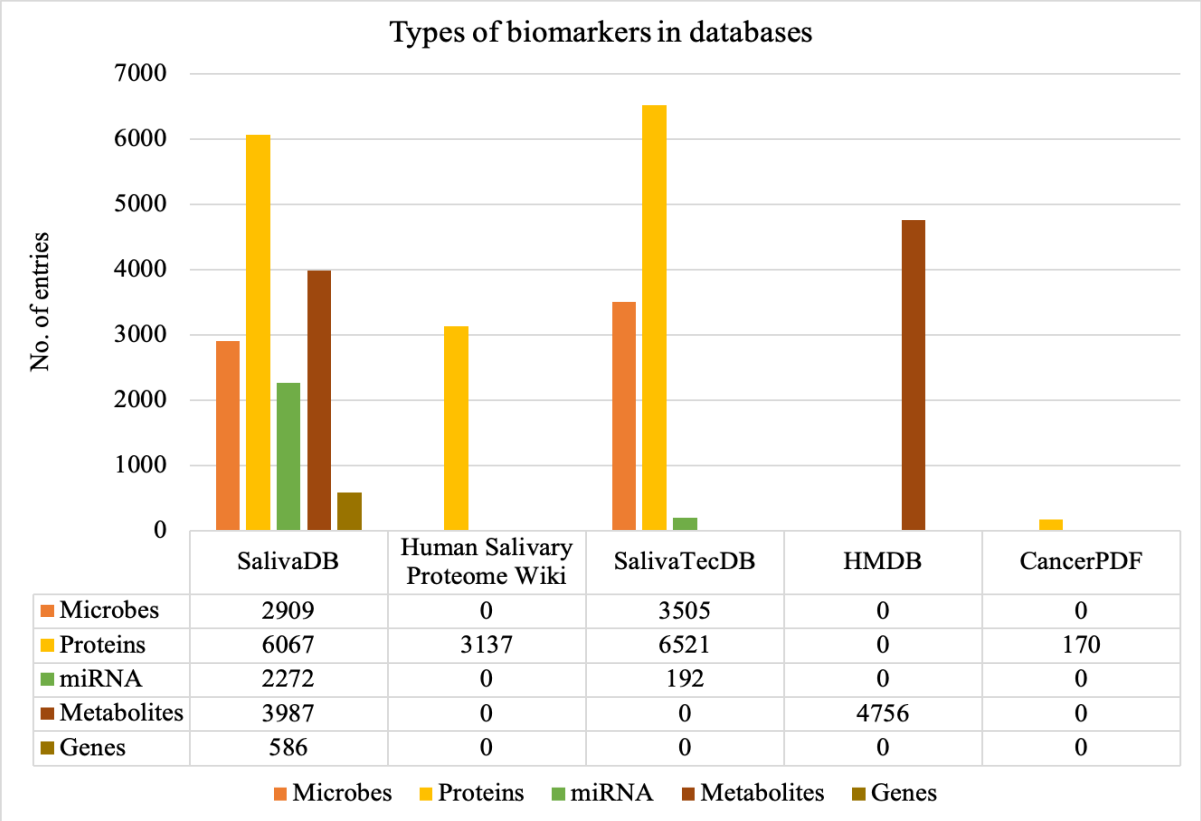


Figure 3.5: Comparison of entries present in SalivaDB with existing resources for different types of biomarkers

Table 3.1: The comparison of existing resources with SalivaDB

Database	Biomarker Type		Last Update	Manual Curation
	Diagnostic	Prognostic		

SalivaDB	**2491**	**124**	**2022**	**Yes**
Human Salivary Proteome Wiki	-	-	2022	No
SalivaTecDB	193	-	2017	Yes
HMDB	-	-	2022	Yes
CancerPDF	-	-	2017	Yes

3.4 Discussion

SalivaDB brings together five distinct categories of salivary biomarkers in one place. All data gathered from the literature were manually curated before entry to ensure accuracy and consistency. The database provides users with details for each biomarker, encompassing aspects such as patient cohorts, sample collection and analysis methodologies, and the regulatory status (upregulation or downregulation) of the biomarker in disease. We provide this information in a structured table on a free, easy-to-use web platform, making it simple for researchers and clinicians to access and understand salivary biomarker data. SalivaDB can be leveraged for numerous purposes in research and clinical contexts:

1. **Comprehensive Biomarker Resource:** To our knowledge, SalivaDB is unique in aggregating salivary biomarker information across the genomics, proteomics, transcriptomics, metabolomics, and microbiomics domains, all in one database.
2. **User-Friendly Interface:** The platform a user-friendly interface to make it accessible to users with minimal bioinformatics experience.
3. **Cross-Reference Linking:** Each entry in SalivaDB is cross-linked to external databases (PubMed, UniProt, NCBI Gene, miRBase, NCBI Taxonomy, PubChem), allowing users to seamlessly navigate and gather information from multiple sources on a single platform.
4. **Advanced Query Capability:** The advanced search functionality allows users to perform complex queries, helping them filter and pinpoint data relevant to their research.
5. **Literature-Based Evidence:** SalivaDB facilitates the retrieval of complete supporting information from the literature, aiding researchers in selecting specific biomarkers for further investigation in a given disease.

6. **Broad Utility:** The database is valuable to a wide range of fields, including disease diagnostics and prognostics, drug development, bioinformatics, and various “omics” research areas.

3.4.1 Limitation

While the SalivaDB data have been carefully curated and verified to minimize errors, it is possible that some mistakes remain due to human error. Therefore, absolute accuracy cannot be guaranteed.

Chapter 4

Prediction of Exosomal Proteins

4.1 Introduction

Protein secretion is essential for a wide range of biological functions (Kravchenko-Balasha et al., 2016). In eukaryotes, most secreted proteins follow the classical endoplasmic reticulum (ER)-Golgi pathway(Gomez-Navarro & Miller, 2016). In this pathway, a signal peptide at the protein's N-terminus directs the nascent protein through the ER and Golgi, which is packaged into vesicles and transported to the cell surface (Lopez-Verrilli & Court, 2013). In addition to this classical route, some proteins are secreted via unconventional pathways that do not require an ER signal peptide. Unconventional secretion can occur through direct translocation across the plasma membrane or via release in extracellular vesicles. Exosomes are particularly notable among the various classes of extracellular vesicles (Kuo et al., 2022; Meldolesi, 2022).

Exosomes are small membrane vesicles (approximately 30–150 nm in diameter) of endosomal origin (L. M. Doyle & Wang, 2019). They are formed inside the cell when multivesicular bodies (MVBs) containing intraluminal vesicles (ILVs) fuse with the plasma membrane, releasing the ILVs as exosomes into the extracellular environment (Abels & Breakefield, 2016; Bellingham et al., 2012). Exosomes carry a complex cargo derived from their cell of origin, including DNA, lipids, proteins, microRNA, and mRNA (Figure 4.1) (Kalluri & LeBleu, 2020). The molecular content of exosomes can change under diseased conditions, making exosomal cargo a useful source of biomarkers (Huda et al., 2021). They are found abundantly in body fluids such as saliva, urine, blood, bile, cerebrospinal fluid, breast milk, semen, amniotic fluid, and sputum (Han et al., 2018). Exosomes are highly stable in these biofluids and protect their cargo from degradation, which makes exosomal biomarkers more sensitive and specific than biomarkers extracted directly from whole fluids (Sun et al., 2019; B. Zhou et al., 2020). Additionally, exosomal markers are readily accessible from most biofluids, making exosome-based diagnostics cost-effective and convenient (Théry et al., 2006; W. Yu et al., 2021). Since proteins and peptides are among the most widely studied biomolecules for biomarkers, characterizing exosomal proteins could enable the development of minimally invasive diagnostic methods and novel therapies for a variety of diseases (Hu et al., 2022; Jeppesen et al., 2014; Poersch et al., 2016). For example, the protein cargo of exosomes shed by distant tumors can carry diagnostic information about those tumors which is otherwise difficult to obtain involving complex and invasive procedures like tissue biopsy (Hu et al., 2022). Isolating proteins from exosomes is also more efficient than directly from blood, since blood contains many interfering substances (Yi et al., 2022).

The identification of proteins secreted via exosomes presents certain challenges. The cells produce a wide array of proteins, many of which are highly similar to each other. In addition, determining the origin of exosomes is also extremely difficult unless they produce a highly specific cargo (X. Li et al., 2019). Thus, there is a need for computational methods that can reliably predict whether a protein is secreted in exosomes. In this direction, several predictors have been developed for classical and unconventional secretory proteins. These include tools such as SRTpred, SPRED, SecretP, OutCyte, and SecretomeP 2.0 (Bendtsen et al., 2004; Garg & Raghava, 2008; Kandaswamy et al., 2010b; L. Yu et al., 2010; Zhao et al., 2019). However, none of these methods was specifically trained on exosomally secreted proteins, nor do they identify motifs characteristic of exosomal proteins. To date, the only method focused on exosomal protein prediction is ExoPred, which was trained on exosomal proteins from vertebrates (Ras-Carmona et al., 2021). However, ExoPred has limited accuracy and does not provide insight into sequence motifs that might be involved in exosomal secretion.

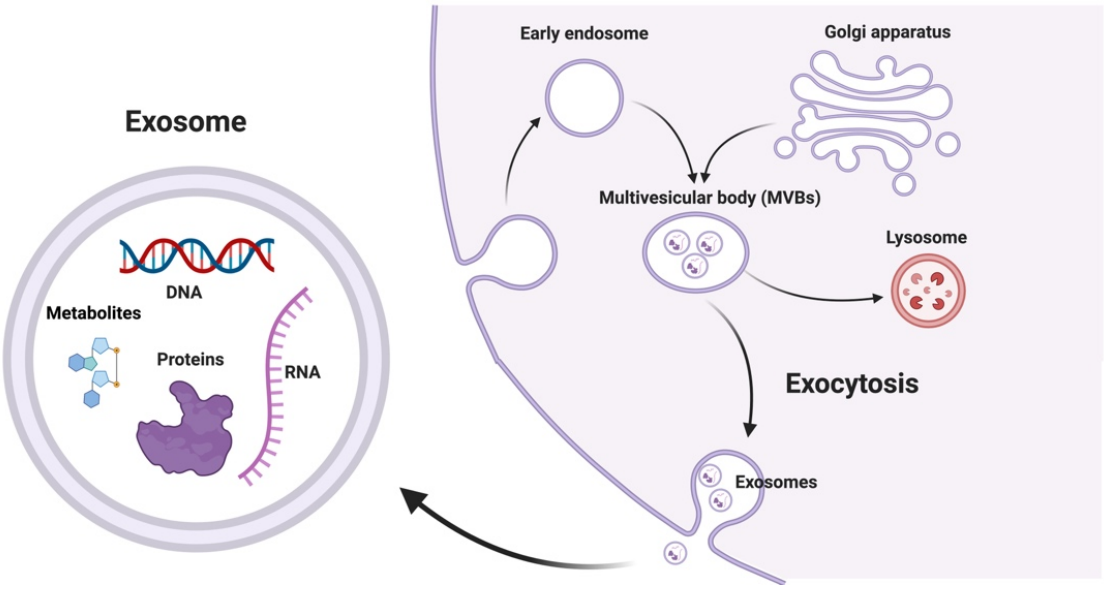


Figure 4.1: Mechanism of formation of exosomes

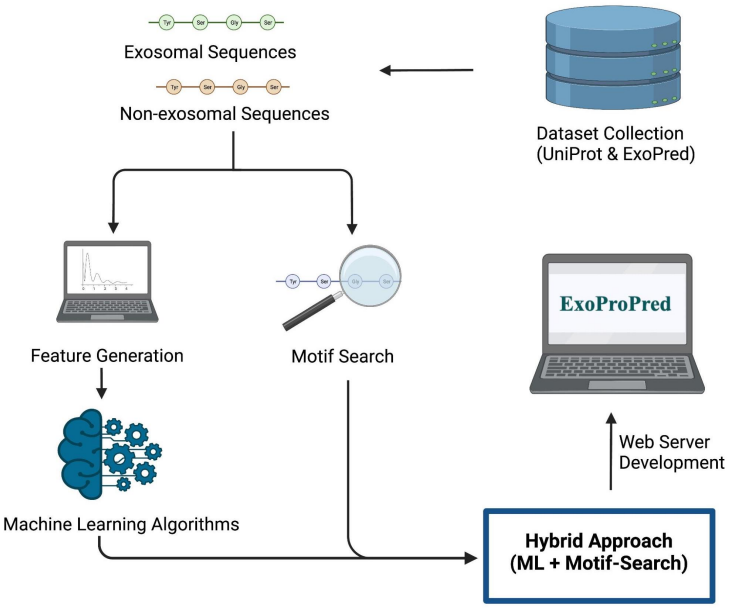


Figure 4.2: A brief overview of the methodology followed in the study

To complement existing approaches, we developed a dedicated classifier for human exosomal proteins. We employed several model-building techniques and incorporated diverse types of protein features, including sequence composition descriptors and a motif-based approach (see Figure 4.2). In addition, we also provide users with a new module to detect exosomal motifs within protein sequences. The availability of such a motif analysis can help researchers design new protein sequences for targeted exosomal delivery. It can improve our understanding of the mechanisms by which proteins are sorted into exosomes.

4.2 Materials and Methods

4.2.1 Compilation and processing of the dataset

The dataset used in this study was collected from UniProt (release 2022_02, May 25, 2022) and from the ExoPred dataset (Ras-Carmona et al., 2021; UniProt Consortium, 2021). We retrieved 2,178 human exosomal protein entries from UniProt by using the following queries: (i) go:0070062 AND reviewed:true AND organism_id:9606 (i.e., proteins annotated with gene ontology term “extracellular exosome”), (ii) the keyword “extracellular exosome” with reviewed:true AND organism_id:9606, and (iii) the term “exosome” with reviewed:true AND organism_id:9606. In addition, we obtained 2,551 reviewed human exosomal proteins from the ExoPred dataset (the positive set used by ExoPred). After merging the data from UniProt and ExoPred and removing duplicates, we got 3,915 unique exosomal protein sequences.

For the negative (non-exosomal) class, we retrieved 18,207 human protein entries from UniProt that were annotated as not being associated with exosomes. Specifically, we used a query excluding the GO term and keywords above (e.g., NOT go:0070062 NOT "Extracellular exosome" NOT Exosome AND reviewed:true AND organism_id:9606). After removing the redundant sequences, we combined these with the non-exosomal proteins from the ExoPred dataset, yielding 20,330 total negative sequences. We further filtered out any protein sequences containing non-standard amino acid letters (B, J, O, U, X, or Z) and any proteins shorter than 55 residues or longer than 1500 residues.

Finally, we applied CD-HIT clustering to eliminate redundancy in both classes, ensuring that no two sequences had >40% pairwise identity (Fu et al., 2012). After that, we obtained 2,831 non-redundant exosomal and 10,680 non-exosomal proteins. To maintain the balance within the dataset, we randomly selected 2,831 non-exosomal proteins from 10,680 non-exosomal proteins to use as the negative set. Thus, the final dataset for model training and evaluation consists of 2,831 exosomal and 2,831 non-exosomal protein sequences. A detailed flowchart of the methodology followed in this study is given in Figure 4.3.

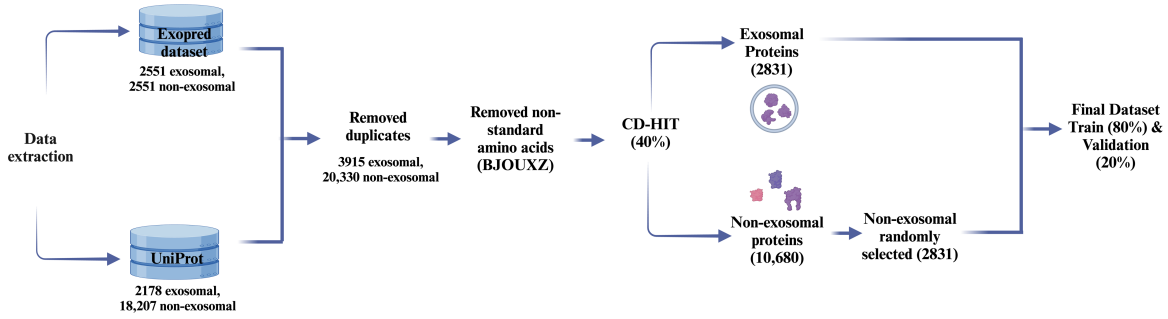


Figure 4.3: A flowchart of data collection and preprocessing in this study

4.2.2 Feature generation

We computed a broad set of sequence-derived features for every protein to develop a prediction model that can classify proteins as exosomal and non-exosomal. Several protein feature encoding techniques have been described in previous studies (Aggarwal et al., 2023; Hasan et al., 2020; Kaur et al., 2024; Mathur et al., 2021). In this work, we used the Pfeature tool to generate a comprehensive collection of descriptors for each protein, including composition-based and evolutionary features (Pande et al., 2022).

A) Composition-based features

The composition-based feature module of Pfeature produces a vector of up to 9,163 features for each protein. These features encompass a wide range of sequence composition measures and physicochemical properties. For example, they include basic measures such as the amino acid composition (AAC), tripeptide composition (TPC), and dipeptide composition (DPC) as well as more complex descriptors like pseudo-amino acid composition (PAAC), amino acid index features (AAI), atomic composition (ATC), Shannon entropy of the sequence, and many more. Each protein in the positive (exosomal) and negative (non-exosomal) datasets was encoded by this high-dimensional feature vector (Pande et al., 2022). The composition-based features computed in this study using Pfeature, along with their vector sizes, are given in Table 4.1.

Table 4.1: List of all composition-based features computed in this study, along with their vector length

Feature Name	Feature Vector Length
Tripeptide Composition (TPC)	8000
Dipeptide Composition (DPC)	400
Conjoint Triad Calculation (CTC)	343
Composition enhanced Transition Distribution (CTD)	189
Quasi-Sequence Order (QSO)	42
Physicochemical Properties Composition (PCP)	30
Shannon Entropy of Physicochemical Property (SPC)	25
Amphiphilic Pseudo Amino Acid Composition (APAAC)	23
Pseudo Amino Acid Composition (PAAC)	21
Amino Acid Composition (AAC)	20
Distance Distribution of Residue (DDOR)	20
Residue Repeat Information (RRI)	20
Shannon Entropy of Residues (SER)	20
Atom Composition (ATC)	5
Bond Composition (BTC)	4
Shannon Entropy of Protein (SEP)	1

B) Evolutionary features

Evolutionary features can provide information beyond a protein's primary sequence features (Kumar et al., 2007; Sharma et al., 2021). To capture evolutionary information, we generated a position-specific scoring matrix (PSSM) for each protein using PSI-BLAST (Position-Specific Iterated BLAST) (Altschul et al., 1997). Each PSSM is a matrix of size $20 \times L$ (where L is the sequence length) that quantifies the evolutionary conservation of each of the 20 standard amino acids at each position in the sequence. However, a fixed-length feature vector is needed for every protein to use these profiles in machine learning models. We therefore computed PSSM-400 composition features, a condensed representation of the PSSM used in earlier studies (Sharma et al., 2021). The PSSM-400 is a 400-dimensional vector (20×20) that summarizes the PSSM by measuring the occurrence of each of 20 amino acids in a sequence (Pande et al., 2022).

4.2.3 Feature selection

It is known that not all extracted features are expected to be informative for distinguishing two classes, here, exosomal from non-exosomal proteins (Abraham et al., 2014; Ren et al., 2022). So, we performed feature selection to identify the most relevant subset of features from the big set of features (9163 composition-based and 400 evolutionary information-based features). We applied a recursive feature elimination (RFE) technique using a logistic regression (LR) classifier as the estimator. In RFE, features are ranked by importance, and the weakest features are iteratively removed until a desired number of features remains. We used RFE to select the top 20 most relevant features from the composition-based feature set and the top 50 features from the evolutionary feature set.

Feature selection was performed on the standardized feature values (we applied a z-score normalization using the StandardScaler from scikit-learn before RFE) (Pedregosa et al., 2011). The RFE procedure eliminated features one by one until only the specified number remained. The selected top features (50 evolutionary features and 20 composition features) were then used to train various machine learning models for evaluation.

4.2.4 Similarity search using BLAST

Basic Local Alignment Search Tool (BLAST) version 2.2.29+ is a broadly used tool to identify and annotate protein sequences based on sequence similarity (Altschul et al., 1990). In this study, we evaluated a BLAST-based approach for identifying exosomal proteins. We

constructed a BLAST database containing all proteins from our exosomal and non-exosomal datasets and then performed BLAST searches for each query protein against this database. In the training set, the query protein's self-hit was excluded. We experimented with three different decision strategies for predicting class labels from BLAST results, considering up to the top five hits for each query (Sharma et al., 2021, 2022). These strategies were evaluated across various E-value cutoffs:

- **Top-hit only:** In this strategy, we consider only the best BLAST hit for each query. If the top hit of a query protein was an exosomal protein, the query was predicted to be exosomal. However, if the top hit was non-exosomal, the query was predicted as non-exosomal.
- **Top-3 voting:** In this strategy, we consider the query's top three BLAST hits. We assign the query to the class (exosomal or non-exosomal) that constitutes the majority of these three hits. This voting strategy could not be applied if fewer than three hits were returned for a query.
- **Top-5 voting:** Similarly, we consider the top five BLAST hits and use a majority vote to assign the query's class. A minimum of five hits is required for this strategy.

4.2.5 Motif search

The identification of functional motifs in protein sequences can be important for annotation and for distinguishing between classes. In this study, we used the MERCI (Motif Emerging and Classes Identification) program to discover sequence motifs characteristic of exosomal and non-exosomal proteins (Vens et al., 2011). MERCI identifies motifs enriched in one set of sequences (the positive set) compared to another set (the negative set). We ran MERCI in two modes to identify both exosomal-specific motifs and non-exosomal-specific motifs:

- In the first run, we provided the exosomal protein sequences as positive and the non-exosomal sequences as negative sets. This finds motifs that occur frequently in exosomal proteins and are not present in non-exosomal proteins (i.e., motifs enriched in exosomal proteins).
- In the second run, we reversed the sets (non-exosomal as positive, exosomal as negative) to find motifs enriched in non-exosomal proteins.

We used MERCI's options to extract exclusive motifs (present in one class and completely absent in the other) and inclusive motifs (present in both classes but at different frequencies). By default, MERCI sets the maximal allowed frequency of a motif in the negative set (parameter fn) to 0, which yields only exclusive motifs. We increased this value to fn = 8 to allow motifs that appear a limited number of times in the negative set, thereby obtaining inclusive motifs.

We explored different gap settings and amino acid grouping options for motif discovery within the exclusive and inclusive motif searches. In particular, we ran MERCI while allowing: (a) no gaps, (b) a single gap, (c) two gaps. In addition, we also used the Koolman–Rohm amino acid class grouping (which treats amino acids with similar properties as equivalent) to identify additional motifs. These different settings enabled MERCI to identify a variety of motif types. After that, we identified the unique motifs to calculate their coverage within the protein sequences.

4.2.6 ML classifiers

We evaluated several machine-learning algorithms to distinguish exosomal from non-exosomal proteins. The classifiers include GNB, DT, KNN, LR, XGB, RF, and SVC. These models were implemented using the scikit-learn machine learning library in Python. For each algorithm, we performed hyperparameter tuning (via grid search) to optimize its performance on the training data. The optimal hyperparameters were determined by an exhaustive search over parameter combinations using five-fold cross-validation on the training set (Pedregosa et al., 2011). We selected the hyperparameter set that yielded the best average performance across the cross-validation folds for each classifier.

4.2.7 Performance metrics calculation and cross-validation

We split the data into training and validation sets in an 80:20 ratio for model training and evaluation. The 80% of the data was used for training the model, and their final performance was evaluated on the 20% independent validation set, which was kept unseen while training the data. We used five-fold cross-validation on the 80% training portion during training to tune hyperparameters and assess the model's robustness. In each cross-validation fold, the training data were further split: 4/5 of the data were used to train the models, and the rest 1/5 was used as an internal test, and this process was iterated five times so that each fold served as the test

once. This approach ensures that the performance reported is not dependent on any particular subset of data and helps prevent overfitting.

We evaluated the classification performance using both threshold-independent and threshold-dependent metrics. In particular, we computed the AUROC, which is threshold-independent, and other standard metrics at a chosen probability cutoff, including specificity, sensitivity, accuracy, and MCC. Here, sensitivity is the proportion of exosomal proteins correctly identified (true positive rate), while specificity is the proportion of non-exosomal proteins correctly identified (true negative rate). Accuracy is the overall percentage of correct predictions, and MCC is a balanced measure that accounts for true and false positives and negatives.

An optimal threshold was chosen for the threshold-dependent metrics (sensitivity, specificity, and MCC) to produce balanced sensitivity and specificity. By contrast, AUROC measures performance across all possible thresholds and thus does not require a specific cutoff. These metrics have been widely used in prior studies to assess predictive models (Dhall et al., 2021; Jain et al., 2022; Jiao et al., 2021).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \quad (1)$$

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

$$\text{MCC} = \frac{(\text{TN} \times \text{TP}) - (\text{FN} \times \text{FP})}{\sqrt{(\text{FP}+\text{TP})\,(\text{FN}+\text{TP})\,(\text{FP}+\text{TN})\,(\text{FN}+\text{TN})}} \quad (3)$$

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

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

4.2.8 Hybrid model

To improve the performance of our ML models, we implemented a hybrid prediction approach that integrates the results from the motif analysis with the ML classifier outputs. In the hybrid approach, we used a weighted scoring scheme that combined two components: (i) a motif-based score and (ii) the ML model's probability score. In this hybrid model, each protein's final score (S') is calculated by adjusting the ML model's predicted probability (S) with a

contribution from the presence or absence of motifs. Specifically, we defined a motif contribution of +0.5 if a protein sequence contains any exosomal motif (as identified by MERCI), and a contribution of -0.5 if it contains any non-exosomal motif. If no motifs from our lists are found in the sequence, the motif contribution is 0. This contribution score is added to the ML model's probability score (0 to 1). Thus, the combined hybrid score S' can range from –0.5 to +1.5.

In summary, the hybrid scoring function can be written as:

- **S' = S + 0.5**, if the sequence has an exosomal motif,
- **S' = S – 0.5**, if the sequence has a non-exosomal motif,
- **S' = S** if no relevant motif is found,

Where S is the probability from the ML model (e.g., the RF classifier) and S' is the resulting hybrid score. This approach effectively boosts the prediction score for proteins that contain exosomal motifs and reduces the score for those with non-exosomal motifs. We then classify sequences as exosomal or not based on the hybrid score S' and an optimized threshold. Several previous studies have employed such hybrid approaches (Dhanda et al., 2013; Sharma et al., 2021). We aimed to strengthen our ML model by adding motif analysis, using sequence patterns that might be missed by global features alone.

4.3 Results

4.3.1 Amino acid composition (AAC) analysis

We first analyzed AAC of exosomal versus non-exosomal proteins. Overall, the average AAC values of the two classes were similar, with only slight differences observed for most amino acids. However, using a Mann–Whitney U test to compare the distributions, we found that about 15 amino acids showed statistically significant differences between exosomal and non-exosomal proteins. (Detailed *p*-values for each amino acid are provided in Table 4.2, and the composition profiles are visualized in Figure 4.4.) The most significant differences in mean composition were observed for a few residues: serine had the most significant difference (exosomal proteins had a 0.93 higher average fractional content of serine compared to non-exosomal proteins), followed by leucine (0.76 higher) and proline (0.64 higher). All of these differences were significant with $p < 0.05$. This suggests that exosomal proteins may be slightly

enriched in specific amino acids such as serine, although the compositional differences were minor for most residues.

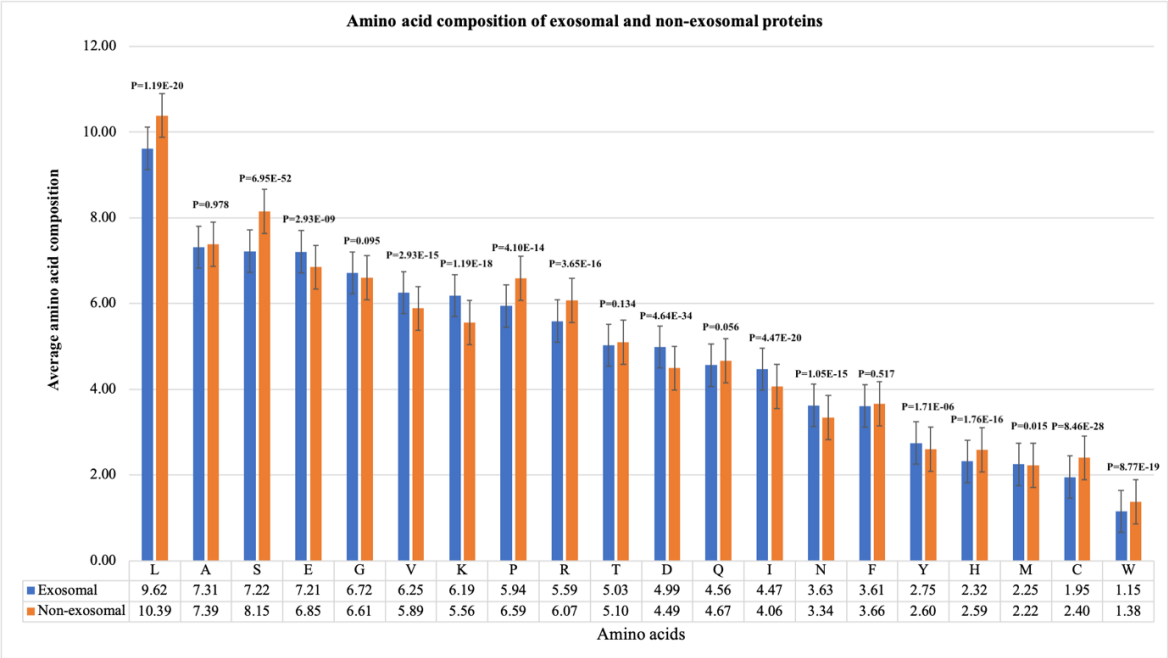


Figure 4.4: Analysis of AAC for exosomal and non-exosomal protein sequences [https://doi.org/10.1002/pmic.202300231]

Table 4.2: Mann-Whitney U test results for Amino Acid Composition (AAC), here, E = Exosomal, N = Non-exosomal

Feature	Mean_E	Mean_N	Mean Difference	Var_E	Var_N	p-value	Significant
AAC_S	7.221	8.153	-0.932	5.294	6.918	6.95E-52	yes
AAC_L	9.620	10.388	-0.768	7.210	9.085	1.19E-20	yes
AAC_P	5.944	6.585	-0.641	7.701	10.099	4.10E-14	yes
AAC_R	5.590	6.070	-0.481	4.655	5.823	3.65E-16	yes
AAC_C	1.950	2.399	-0.450	1.957	3.471	8.46E-28	yes
AAC_H	2.316	2.587	-0.271	1.103	1.543	1.76E-16	yes
AAC_W	1.153	1.377	-0.224	0.653	0.901	8.77E-19	yes
AAC_Q	4.563	4.668	-0.105	3.020	3.605	0.056	no
AAC_A	7.314	7.385	-0.071	6.372	7.985	0.978	no
AAC_T	5.030	5.099	-0.068	2.250	2.898	0.134	no
AAC_F	3.611	3.664	-0.053	1.997	2.976	0.517	no
AAC_M	2.247	2.222	0.024	0.911	1.108	0.015	yes

AAC_G	6.715	6.605	0.110	6.413	6.672	0.095	no
AAC_Y	2.745	2.604	0.141	1.571	1.882	1.71E-06	yes
AAC_N	3.625	3.340	0.286	1.983	2.538	1.05E-15	yes
AAC_E	7.207	6.849	0.357	6.692	8.277	2.93E-09	yes
AAC_V	6.253	5.887	0.366	2.903	3.567	2.93E-15	yes
AAC_I	4.472	4.065	0.407	3.183	3.830	4.47E-20	yes
AAC_D	4.986	4.492	0.494	2.369	2.860	4.64E-34	yes
AAC_K	6.186	5.560	0.626	7.788	7.720	1.19E-18	yes

4.3.2 BLAST performance

We used BLAST-based sequence similarity to classify the proteins as exosomal or non-exosomal, but this approach yielded low predictive performance. In five-fold cross-validation on the training set, the simple top-hit BLAST strategy correctly identified only ~18% of exosomal proteins (sensitivity ~18%) when using E-value cutoff (e.g., 1e-6). The voting strategies that consider the top 3 or top 5 hits were even more conservative, achieving sensitivity around 11% and 8% respectively on the training data. On the independent validation set, BLAST showed a similarly poor sensitivity: approximately 18% for top-hit, ~11% for top-3, and ~9% for top-5. These low sensitivities indicate that BLAST frequently fails to find a closely homologous exosomal protein in the database for a given query. Increasing the E-value threshold increased the number of BLAST hits, but it also increased the misclassification rate (more non-exosomal hits being wrongly considered as matches for exosomal proteins, and vice versa). In summary, the BLAST approach produced a high rate of false negatives (many exosomal proteins were not recognized as such) and also some false positives. Even under the best conditions, BLAST's sensitivity was below 20% on both training and validation sets, with overall accuracies around 60–65% (see Table 4.3). These results demonstrate that a purely sequence similarity-based method is not sufficient for this prediction task, likely due to high sequence similarity among exosomal and non-exosomal proteins.

Table 4.3: The results for top 1, 3, and 5 BLAST hits in the independent validation set searched against the database created using training dataset (Here, sens = sensitivity and spec = specificity) [https://doi.org/10.1002/pmic.202300231]

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	10.80	11.28	18.45	10.41	14.03	13.06
	10^{-4}	19.17	10.11	11.30	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	4.50	6.35	7.41
	10^{-5}	11.99	5.03	6.49	6.98	12.00	4.94	6.80	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}	15.90	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	3.20	4.22	5.12	8.65	3.35	4.94	6.00
	10^{-5}	8.90	3.44	4.70	5.50	8.91	3.80	5.12	6.62
	10^{-4}	9.43	3.73	4.86	5.90	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}	12.10	5.06	7.00	7.33	12.53	5.74	7.68	8.38

4.3.3 ML models

A total of 9163 composition-based features and 400 evolutionary features were calculated for every sequence in the training and validation sets. We first developed machine learning models using only the basic composition features (20-dimensional AAC) and only the PSSM-based evolutionary features (400-dimensional PSSM composition) to establish a baseline performance. The random forest (RF) model trained on just the 20 AAC features achieved an AUROC of about 0.70 on the independent validation set. In comparison, the logistic regression (LR) model trained on the 400 PSSM-composition features achieved an AUROC of 0.72 on the validation set. These results are summarized in Table 4.4.

In order to improve the model's performance, we applied feature selection to the complete feature set. We identified the top 20 composition-based features and the top 50 PSSM-based (evolutionary) features. The best-performing model with the reduced feature set (70 features) was an RF classifier, which achieved an AUROC of 0.75 and 0.73 on the training and independent validation sets, respectively. The detailed results for all ML models developed on selected features are given in Table 4.5.

A) Compositional features

Using only the amino acid composition (AAC) features, the random forest model achieved an AUROC of 0.71 on the training set and 0.70 on the independent validation set. The other classifiers (SVC, XGB, LR, etc.) yielded slightly lower performance with AAC alone (in the high 0.60s to 0.70 AUROC range). These results (see Table 4.4) indicate that overall amino acid composition provides some discriminatory signal, but AAC is not sufficient for differentiating exosomal and non-exosomal proteins.

B) Evolutionary features

We also trained models using only the evolutionary features derived from PSSMs. The LR classifier performed the best and achieved an AUROC of about 0.73 on the training data and 0.72 on the validation set. The LR, a simple model, slightly outperformed more complex classifiers like RF and SVC on these features (RF obtained ~0.71 AUROC on validation, SVC ~0.71 as well; see Table 2). One possible reason for this result is that the PSSM features may have a roughly linear relationship with the class label that LR can capture effectively. The PSSM composition quantifies how conserved each amino acid is at positions throughout the sequence; this representation might align well with a linear decision boundary between exosomal and non-exosomal classes. In contrast, non-linear models like SVM or RF could overfit noise in the 400-dimensional feature space if some positions' scores are not informative or are noisy. Overall, including evolutionary information via PSSMs improved the model's performance over using simple AAC, confirming that evolutionary profiles contribute a useful signal for this classification task.

Table 4.4: Results for ML models developed for AAC and PSSM composition features [https://doi.org/10.1002/pmic.202300231]

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

4.3.4 Feature selection

We applied RFE-based feature selection to identify the most informative features. Using RFE, we obtained 20 top-performing features from the 9163 composition-based features. After training the ML model on these 20 features, we found that an SVC classifier achieved an AUROC of 0.71 on both the training and validation sets. We similarly selected the top 50 features from the 400 evolutionary (PSSM) features. An RF model using only these 50 features reached AUROCs of 0.74 (training) and 0.71 (validation).

Next, we combined the two selected subsets to form a 70-feature input (20 composition + 50 evolutionary features). The models performed better using this combined feature set than using either set alone. The best result was obtained with the RF model, which achieved AUROCs of 0.75 on the training data and 0.73 on the independent validation set. The complete results for the features selected in this stuyd are shown in Table 4.5.

Table 4.5: Results for the various ML models developed on features including composition-based features (n=20), evolutionary features (n=50), and a combination of evolutionary and composition-based features (n=70) [https://doi.org/10.1002/pmic.202300231]

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

4.3.5 Top selected features

The top 20 composition-based features included descriptors such as amino acid indices (AAI), atom composition (ATC, which counts the number of atoms like C, H, N, O, S in the protein), PAAC, SEP, QSO, SER. Among these, we observed that three of the highest-ranking features were specifically related to the amino acid tryptophan (W): the SER, QSO, and PAAC values for tryptophan were all selected as important features. This suggests that the abundance and distribution of tryptophan residues might be a distinguishing factor for exosomal proteins. Tryptophan is a bulky aromatic amino acid, and its presence could affect protein structure or interactions relevant to exosomal packaging.

In addition to tryptophan-related features, the feature selection indicated that certain atomic composition features were significant. In particular, the ATC features corresponding to the counts of nitrogen and sulfur atoms in the protein sequence were among the top features. These features reflect the overall composition of amino acids containing nitrogen (which is essentially all amino acids) and sulfur (primarily cysteine and methionine). The prominence of nitrogen and sulfur counts might point to differences in amino acid composition, such as slightly higher methionine/cysteine content in one class. Other selected features (SEP, QSO, etc.) capture aspects of sequence complexity and order, implying that exosomal proteins might differ in how their amino acids are ordered or how composition varies along the sequence.

4.3.6 Motif search

We identified motifs that were either exclusive to or significantly enriched in exosomal and non-exosomal proteins using the MERCI algorithm. We obtained a rich set of motifs by applying various gap settings and the Koolman–Rohm amino acid class grouping. Overall, MERCI found 89 distinct motifs significantly associated with exosomal proteins and 130 motifs associated with non-exosomal proteins. These numbers include exclusive motifs (completely absent from the opposite class) and inclusive motifs (present in both classes but more frequent in one).

The identified exosomal motifs were present in 1,441 of the 2,831 exosomal proteins in our dataset (coverage ~50.9%), while the identified non-exosomal motifs were present in 1,373 of the 2,831 non-exosomal proteins (~48.5% coverage). In other words, roughly half of the proteins in each class contain at least one of the motifs. This relatively high coverage indicates that many exosomal proteins share common short sequence patterns. The top five motifs for each class, in both the exclusive and inclusive categories, along with the number of sequences

in which each motif occurs, are given in Table 4.6. Upon examining the sequences of the top motifs, we observed a trend: many discriminative motifs were enriched in aliphatic (hydrophobic) amino acids. Hydrophobic motifs could be involved in interactions with the exosome membrane or sorting receptors.

Table 4.6: The top 5 exclusive and inclusive motifs found in exosomal sequences and their coverage in sequences for different settings: a) No gap, b) Gap=1, c) Gap=2, and d) Class=Koolman Rohm (here, pos = occurrence in positive sequences, neg = occurrence in negative sequences, fn = maximal frequency in negative sequences) [https://doi.org/10.1002/pmic.202300231]

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

4.3.7 Hybrid approach

As we got a high sequence coverage using the motif-based approach, we integrated the motif-based predictions with our machine learning model to develop a hybrid approach. As detailed in the Methods section, we computed a hybrid score for each protein by adding a motif-based score or penalty to the ML model's probability. When we combined motif information with the RF model trained on AAC features alone, the performance on the training set increased from an AUROC of 0.71 to 0.84 on the independent validation set. We then applied the hybrid approach to the RF model developed on 70 (20 composition and 50 evolutionary) selected features. This resulted in an AUROC of 0.87 on the training data and 0.85 on the validation set. In both cases, the hybrid model showed a clear improvement over the corresponding ML-only model. The performance metrics of various hybrid combinations are summarized in Table 4.7, and AUROC plots are given in Figure 4.5.

Table 4.7: The Results for the ensemble approach applied in the study comprising (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) [https://doi.org/10.1002/pmic.202300231]

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

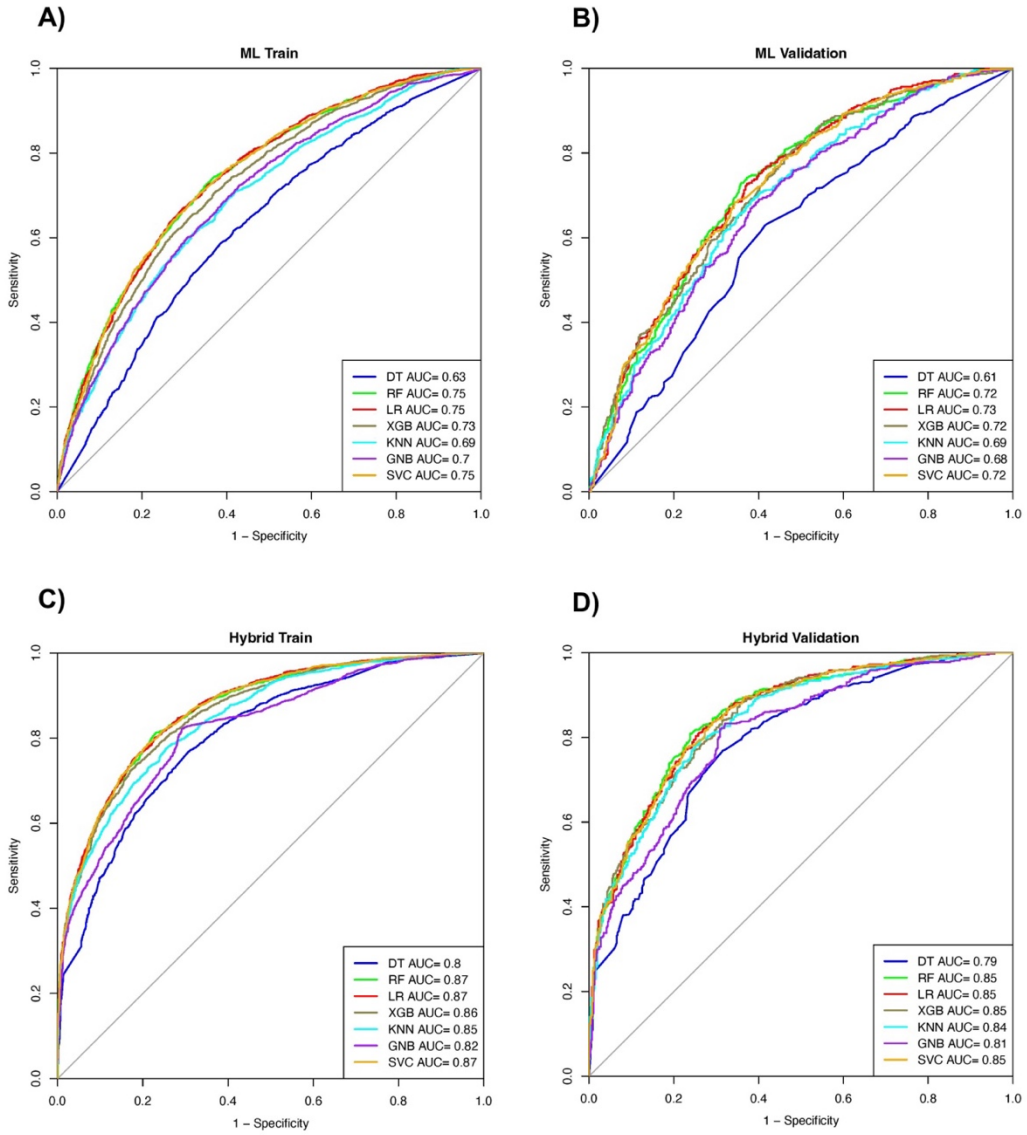


Figure 4.5: The AUROC plots depicting: (A) Training set performance using ML models, (B) Validation set performance using ML models, (C) Training set performance using the hybrid model, and (D) Validation set performance using the hybrid model

[https://doi.org/10.1002/pmic.202300231]

4.3.8 Web server development

To facilitate wider use of our prediction method, we developed a user-friendly web server called ExoProPred (https://webs.iiitd.edu.in/raghava/exopropred/). The web server implements our two best-performing hybrid models: (a) AAC features with motifs, and (b) Top 70 selected features with motifs. The users can choose either model to make predictions for their protein

sequences. The web interface is built on a modern HTML framework, making it accessible on various devices and screen sizes. The ExoProPred web server consists of three main modules:

- **Prediction module:** This module allows users to submit one or multiple protein sequences in FASTA format. The sequences are run through our prediction pipeline, and the server returns a prediction for each sequence (exosomal or non-exosomal) along with the probability score. This module uses the selected hybrid model (as chosen by the user) to make the predictions.
- **Motif scan module:** This module enables users to analyze a protein sequence for the presence of exosomal or non-exosomal motifs. When a sequence is submitted, the server scans it using the set of motifs identified by MERCI and highlights any motifs found. The output specifies whether each detected motif is associated with exosomal or non-exosomal proteins.
- **Download module:** This provides links to download the standalone version of ExoProPred and related data. We developed a Python-based standalone package for users who wish to run bulk predictions locally or integrate our model into their own pipelines. The standalone software, the dataset, and a help file can be downloaded via this module.

4.3.9 Comparison with other prediction tools

We compared ExoProPred with existing tools for the prediction of exosomal proteins. To our knowledge, ExoPred is the only prior method specifically trained for exosomal protein prediction (Ras-Carmona et al., 2021). Other methods like SecretomeP 2.0 and OutCyte predict unconventional secretion but do not focus exclusively on exosomes (Bendtsen et al., 2004; L. Yu et al., 2010; Zhao et al., 2019). We evaluated ExoPred, SecretomeP 2.0, and OutCyte on the same independent test set of 569 proteins that we used for our model comparison. We ensured that none of these sequences were used in training ExoPred by removing any overlap with ExoPred's dataset. ExoPred achieved an accuracy of 66.08% on our validation set. We noted that ExoPred tends to predict most proteins as non-exosomal, which means it had a high specificity but very low sensitivity. In other words, while it correctly identified many non-exosomal proteins, it missed a significant fraction of the exosomal proteins. For SecretomeP 2.0, which predicts if a protein is secreted unconventionally (without a signal peptide), we used the recommended threshold of NN-score > 0.6 to call a protein "secreted via

non-classical pathway". On our dataset, SecretomeP 2.0 identified exosomally secreted proteins with an accuracy of 54.83%. For OutCyte, we ran OutCyte in "UPS" mode (Unconventional Protein Secretion) on our validation dataset proteins. OutCyte obtained an accuracy of 61.16% on the validation set (Zhao et al., 2019). Similar to ExoPred, OutCyte's predictions showed low sensitivity. Finally, we evaluated our ExoProPred hybrid model on this test set. ExoProPred achieved an accuracy of 79.4%, substantially higher than the other methods. In terms of the balance between sensitivity and specificity, ExoProPred could identify exosomal proteins at a much higher true positive rate while maintaining a high actual negative rate. ExoProPred's sensitivity and specificity were around the 79–80% range, indicating a well-balanced predictor. In contrast, ExoPred, SecretomeP, and OutCyte all skewed towards high specificity but low sensitivity. Table 4.8 provides a detailed comparison of the predictions by each tool.

Table 4.8: Comparison of prediction by existing web servers, Outcyte, SecretomeP 2.0, and ExoPred with ExoProPred on a validation dataset [https://doi.org/10.1002/pmic.202300231]

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

4.4 Discussion

There is a growing need for non-invasive therapies and diagnostic methods to spare patients from undergoing painful and stressful medical procedures. Exosome-based biomarkers have emerged as promising candidates in this regard. Exosomes are abundant in body fluids like saliva, blood, and urine and can be harnessed for disease detection or treatment development (Huda et al., 2021). These biomarkers originate from parent cells and often reflect the condition of those cells. Exosomes are highly stable in circulation and protect their cargo, making exosomal biomarkers more reliable than those obtained from direct fluid sampling (Hu et al., 2022; Jeppesen et al., 2014; Poersch et al., 2016). For example, exosomal proteins are shielded from proteases in the blood, and exosomal RNAs are protected from RNases. Moreover, exosomes are generally non-immunogenic, meaning they can circulate without provoking an

immune response, which further contributes to the stability of their cargo as biomarkers (Boukouris & Mathivanan, 2015).

Among the various types of molecules carried by exosomes, proteins are widely studied and used for the diagnosis, prognosis, and therapy of different diseases. However, identifying which proteins are secreted via exosomes is challenging. One of the reasons is that exosomal proteins are highly similar to other cellular proteins as they are eventually secreted to exosomes via the cell (X. Li et al., 2019). Given this challenge, computational methods become crucial for predicting exosomal proteins. In this work, we developed a systematic computational approach to predict exosomal proteins and to discover sequence motifs associated with exosomal secretion. We assembled a dataset of experimentally validated 2831 exosomal and 2831 non-exosomal proteins found in humans from UniProt and ExoPred (Ras-Carmona et al., 2021; UniProt Consortium, 2021). This dataset was combined and divided into an 80:20 ratio, where 80% accounted for training and 20% for independent validation sets. A number of composition-based (n=9163) and evolutionary (n=400) features were generated for each sequence in both sets. Firstly, we developed ML models on simple amino acid features (n=20) and evolutionary features (n=400), which resulted in an AUROC of 0.70 on RF and 0.72 on the LR model, respectively, on an independent validation set. Interestingly, a simple logistic regression outperformed more complex classifiers like SVM and random forest on the evolutionary feature set. One possible reason is the nature of the PSSM composition features. Logistic regression assumes a roughly linear relationship between input features and the output; the PSSM composition is a kind of linear summary of evolutionary conservation across the sequence, which might relate linearly to a protein's propensity to be exosomal. Complex non-linear models can potentially capture interactions between features, but they also risk overfitting, especially if many features are noisy or irrelevant. In our case, the PSSM features likely contained some noisy dimensions (positions that do not contribute to secretion via exosomes), and a linear model like LR might have been more robust to that noise than an SVM trying to maximize margins in a high-dimensional space.

Next, we tried to increase the performance and address feature noise and redundancy using the Recursive Feature Elimination method for feature selection from 9163 composition-based and 400 evolutionary features. After feature selection, we got the top-performing 20 compositions and 40 PSSM (evolutionary) features. The top 20 compositional features included AAI, QSO, PAAC, ATC, SEP, and SER. It was also observed that amino acids like Serine, Leucine, and

Proline showed the maximum difference in average amino acid compositions between exosomal and non-exosomal proteins. We built ML models using these features, and the highest performing model was RF with an AUROC of 0.73 on an independent validation set.

Besides building Machine Learning models, we tried alignment-based approaches to differentiate between exosomal and non-exosomal proteins. We used the conventional sequenced alignment tool – BLAST- to identify exosomal proteins (Altschul et al., 1990). However, BLAST could not work well on our data with low sensitivity and high error rate, suggesting that the exosomal and non-exosomal proteins might have high similarity. Next, we explored motif-search using MERCI, where we discovered various motifs exclusively and inclusively present in exosomal and non-exosomal datasets. We used different methods within MERCI to identify a diverse set of motifs, eg, N, Gap, Gap=1, Gap=2, Class=Koolman-Rohm. Using these approaches, we discovered 89 exosomal and 130 non-exosomal motifs covering about 1373 non-exosomal and 1441 exosomal sequences (~50% of the dataset) (Vens et al., 2011). It was observed that motif-approach has high precision but low coverage. In addition, we found that many of these motifs had a prevalence of hydrophobic aliphatic amino acids, which could reflect that exosomal proteins often interact with lipid membranes. The presence of aliphatic amino acids helps stabilize these interactions, making these motifs more likely to be involved in exosome formation or function. In addition, Many exosomal cargo are sorted via the ESCRT (Endosomal Sorting Complex Required for Transport) machinery, which recognizes ubiquitinated proteins (Moreno-Gonzalo et al., 2014) Notably, several of exosomal motifs contain Lysine (K). Lysines are key sites for ubiquitin attachment, and ubiquitination of a cytosolic tail is a well-known signal for inclusion into ILVs. The enrichment of lysine within acidic/basic clusters might indicate potential ubiquitin tag sites or recognition sequences for E3 ubiquitin ligases.

Hence, we developed a hybrid method combining both ML and motif-based classification, which significantly boosted the prediction AUROC from 0.73 to 0.85 and accuracy from 66.55% to 78% with a balanced sensitivity and specificity in an independent validation set. We also compared the performance of our model with the existing ones. Hence, we created a validation set by removing the sequences in the ExoPred training dataset, leading to 569 sequences. These sequences were input for existing methods like Outcyte, ExoPred, and SecretomeP 2.0. The accuracy obtained by Outcyte, ExoPred, SecretomeP 2.0, and ExoProPred is 61.16%, 66.08%, 54.83%, and 79.40%, respectively. ExoProPred outperformed all other

existing models for predicting exosomal proteins or predicting proteins secreted via unconventional pathways.

Finally, we built a tool that helps users to classify exosomal and non-exosomal protein sequences, where we implemented our best-performing hybrid model. In addition to the prediction of exosomal proteins, we also incorporated a motif-search module in our web server to help users discover exosomal and non-exosomal motifs in their query protein sequences to aid in the prediction and identification of new exosomal proteins.

4.5 Conclusion

Exosomal proteins have diverse applications in healthcare, particularly in developing non-invasive biomarkers for disease. As exosomes can be obtained via liquid biopsy (drawing body fluids such as blood or saliva), they enable disease detection and monitoring without requiring invasive tissue biopsies. In this work, we present a highly accurate computational method for predicting exosomal proteins. Our method integrates machine learning models with motif analysis, achieving substantially higher accuracy than previous approaches. In addition to the "black-box" machine learning models, our predictor also incorporates sequence similarity and motif-based reasoning.

Additionally, through our analysis, we discovered several sequence motifs significantly associated with exosomal proteins. These motifs provide clues to the potential mechanisms guiding proteins into exosomes. We developed an online web server and an offline standalone tool for ExoProPred. The tool is freely available and allows scientists worldwide to easily predict whether a protein is exosomal. We believe our study will be helpful for those researching on peptide and proteins-based diagnostics and therapeutics.

Chapter 5

Prediction of exosomal miRNA

5.1 Introduction

Liquid biopsy is a ground-breaking diagnostic approach that enables disease detection and monitoring through the analysis of biofluids (Arora, Kaur, et al., 2023; Shegekar et al., 2023; Swarup et al., 2023). This minimally invasive method allows repeated sampling throughout disease progression, reducing the risks associated with traditional tissue biopsies (Armakolas et al., 2023). Common molecules examined in liquid biopsies include cell-free RNA (cfRNA), circulating tumor DNA (ctDNA), cell-free DNA (cfDNA), and circulating tumor cells (CTCs). However, these biomolecules have challenges, such as low abundance of the target nucleic acids and lack of specific surface markers (Yadav et al., 2024). In recent years, exosomal biomarkers have gained attention as potential alternatives due to their enhanced stability and crucial role in cell-to-cell communication. Exosomes are small extracellular vesicles released by cells, carrying cargoes of lipids, proteins, DNA, miRNA, mRNA, and metabolites (Kalluri & LeBleu, 2020). Exosomal biomarkers are more stable than other cell-free biomolecules and can provide insights into dynamic changes in the tumor microenvironment (Arora et al., 2024). They often can be detected earlier in disease progression, making them suitable for early diagnosis and disease monitoring.

Exosomal miRNAs are small non-coding RNA molecules, about 19–22 nucleotides long, that are important for gene regulation and cell communication (Salehi et al., 2024b; Zheng et al., 2021). The biogenesis of exosomal miRNAs starts in the nucleus, where pri-miRNAs (primary miRNAs) are transcribed by RNA polymerase to form hairpin structures of about 70–100 nucleotides. These pri-miRNAs undergo initial processing (Fan et al., 2020). Exportin-5 then transports the hairpin precursors to the cytoplasm, where Dicer further processes them. Once matured, the double-stranded miRNAs are converted into single strands and sorted into exosomes (See Figure 5.1) (F. Ahmed et al., 2009, 2013). Notably, the loading of miRNAs into exosomes is a non-random and regulated process (W. Li et al., 2017; Wozniak et al., 2020). A number of miRNA biomarkers have been discovered, for example, miR-155, miR-21, and miR-1246 for lung cancer, miR-200c and miR-200b for ovarian cancer, and miR-92a-3p and miR-17-5p for colorectal cancer (Bakhsh et al., 2024; X. Liang et al., 2022; M. Wang et al., 2019). In addition to cancer, exosomal miRNAs are being explored as potential biomarkers for other diseases, including neurological and cardiovascular disorders, underscoring their potential in personalized medicine (Bellingham et al., 2012; Zheng et al., 2021).

In the present study, we made an effort to identify miRNA found in exosomes to leverage their diagnostic and therapeutic potential. First, we curated a a dataset using experimentally validated exosomal and non-exosomal miRNA. Next, we split this dataset, using 80% of the sequences for training and 20% for an independent validation set. We developed predictive models on the training set using five-fold cross-validation and then evaluated their performance on the independent validation set. We explored different strategies to achieve high-accuracy prediction of exosomal miRNAs. Initially, we applied alignment-based techniques based on motif discovery and sequence similarity. These alignment-based methods succeed only if a query sequence contains a known motif or significant similarity to annotated sequences; they fail without such features. To overcome these challenges, we developed AI-based models encompassing ML, DL, and LLM approaches. Finally, we constructed an ensemble model that integrates the strengths of the alignment-based and AI-based models. This ensemble strategy is intended to accurately predict exosomal miRNAs, facilitating their broader use in biomedical studies.

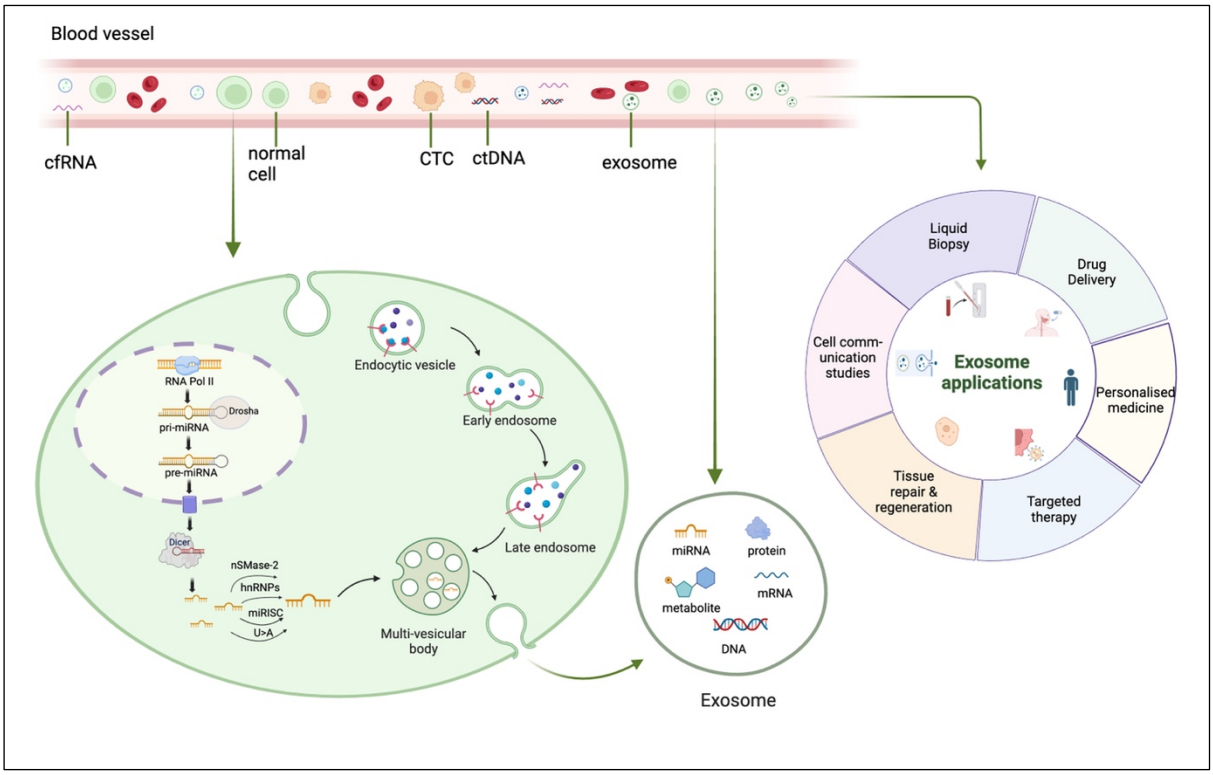


Figure 5.1: The commonly used biomolecules in liquid biopsy found in body fluids and the mechanism of secretion of miRNA from cells to exosomes
[https://doi.org/10.1101/2024.06.20.599824]

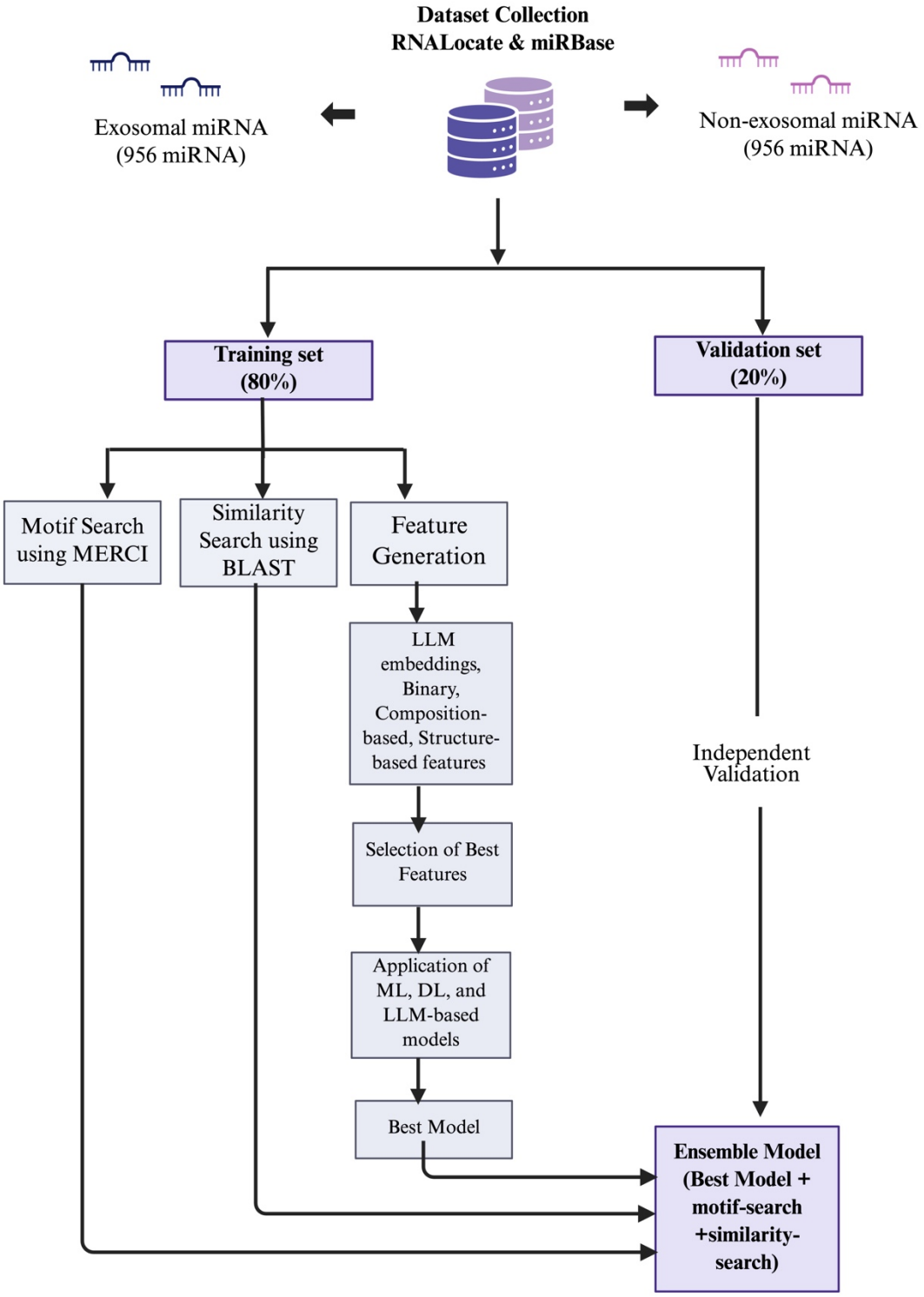


Figure 5.2: The architecture of the algorithm used in EmiRPred

5.2 Methods

5.2.1 Data Collection and Pre-processing

In this study, we created datasets using the miRBase and RNALocate databases (Cui et al., 2022; Kozomara et al., 2019). We collected human exosomal miRNA sequences from RNALocate, which were experimentally validated, yielding 1195 unique exosomal miRNAs. Out of these, we extracted 956 mature miRNA sequences. In the same way, we extracted experimentally validated miRNAs found in humans from RNALocate and miRBase that were not detected in exosomes. This process resulted in 1,694 unique non-exosomal miRNAs. Out of these, we randomly selected 956 mature sequences to balance the data. The final dataset thus contains about 956 non-exosomal and 956 exosomal miRNAs. The lengths of the sequences range from 16 to 26 nucleotides, with maximum sequences having lengths between 21 and 24 nucleotides. The architecture of EmiRPred is illustrated in Figure 5.2.

5.2.2 Alignment-based approaches

5.2.2.1 Motif-Search

We used the MERCI (Motif EmeRging with Classes Identification) tool to identify sequence motifs enriched in exosomal miRNAs (Vens et al., 2011). Motif discovery was performed on the training set of exosomal sequences, and the identified motifs were then searched in the independent validation set. MERCI allows selection of exclusive motifs that occur in the positive class (exosomal) and are absent in the negative class (non-exosomal) in the training data. The software also provides options to specify motif frequency thresholds and whether to allow gaps in motifs.

5.2.2.2 Similarity Search

We employed an alignment-based annotation method called BLAST (blastn-short) to classify miRNA sequences using similarity to known exosomal or non-exosomal sequences (Altschul et al., 1990). We first built a BLAST database from the miRNA sequences in the training set. For evaluating training-set performance, we excluded self-hits and took the top BLAST hit after removing any self-match. For each sequence in the independent validation set, we considered the top BLAST hit to determine classification at a number of E-value thresholds.

This similarity-based annotation approach has been used in previous studies (Kaur et al., 2024; Sharma et al., 2021).

5.2.3 AI-based classification methods

5.2.3.1 Feature Generation

We extracted a broad range of sequence features to develop predictive models for classifying exosomal vs non-exosomal miRNAs. These features, described below, include compositional properties, binary profiles, structural attributes, and embeddings from pre-trained language models.

A) Composition-based Features

We utilized the Nfeature tool to compute numerous composition-based features for each miRNA sequence (Mathur et al., 2021). These included simple nucleotide composition, reverse-complement nucleotide composition, correlation, entropy, nucleotide repeats index, distance distribution of nucleotides, etc. We also calculated term frequency–inverse document frequency (TF–IDF) and used them as features for our AI models. In this context, "terms" are k-mers (sub-sequences of length k), and each miRNA sequence is treated as a document. Term Frequency (TF) is the k-mer count in a sequence which is normalized by the total number of k-mers in that sequence. Inverse Document Frequency (IDF) is the log-scaled inverse of the fraction of sequences in the dataset that contain the k-mer. Thus, TF–IDF highlights k-mers that are frequent in a given sequence but are uncommon across the dataset. We computed TF–IDF features for k-mers in both the sequences and their reverse complements, for k ranging from 1 to 7. Additionally, we experimented with different weighting schemes for TF–IDF, including logarithmic term count (LTC) weighting, term frequency count (TFC) weighting, and entropy weighting.

B) Binary features

We computed binary profile features (one-hot encodings) to represent each miRNA sequence as a binary vector. Since miRNA sequences vary in length (up to 26 nucleotides), we applied padding with a dummy nucleotide (denoted "X") to standardize all sequences to length 26. For example, a sequence "AUTGGCUCUGTCGCCUAAUCU" of length 21 would be padded to "AUTGGCUCUGTCGCCUAAUCUXXXXX". After padding, each nucleotide position was encoded

as a binary indicator across the four possible nucleotides (with "X" for padding as needed). We generated binary profiles for mononucleotides, dinucleotides, and trinucleotides. These one-hot encoded vectors served as input features for certain machine learning models.

C) Structure-based features

We computed RNA secondary structural features for each miRNA using the RNAfold tool from the ViennaRNA Package 2.0 (Lorenz et al., 2011). RNAfold predicts minimum free energy (MFE) secondary structure of an RNA sequence. For each miRNA, we obtained features including the ensemble free energy, minimum free energy, centroid structure diversity, centroid free energy, ensemble diversity, and MFE structure frequency in the ensemble. In this context, an "ensemble" means collecting all possible secondary structures the RNA can form at equilibrium. We generated these features for all miRNAs in both the exosomal and non-exosomal categories. Additionally, we saved the predicted secondary structures as images (JPEG format) for use in deep learning models.

D) Large Language Models (LLM) embeddings as features

We computed feature embeddings using two pre-trained BERT-based LLMs. BERT is a transformer-based model originally developed for natural language processing. Embeddings generated from BERT models are continuous, high-dimensional feature vectors from the final layers of the fine-tuned models. We used the models: (a) BERT-Base Uncased, where all text is lowercased during training (Devlin et al., 2018). We fine-tuned this model on our miRNA training sequences and then extracted vector embeddings for each sequence. (b) DNABERT, a BERT variant specialized for DNA sequences (Ji et al., 2021). As DNABERT expects DNA inputs (with T instead of U), we converted all "U" nucleotides to "T" in our sequences. We fine-tuned DNABERT on our dataset and then obtained embeddings for each miRNA. These embeddings were used as features for classification.

5.2.3.2 Prediction Models

A) ML Models

We implemented several machine learning algorithms to distinguish exosomal from non-exosomal miRNAs. The algorithms included k-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), Extreme Gradient Boosting (XGB), Support Vector Classifier (SVC), Logistic

Regression (LR), Extra Trees (ET), Decision Tree (DT), and Random Forest (RF) (Breiman, 2001; Bulac & Bulac, 2016; T. Chen & Guestrin, 2016; Cristianini & Ricci, 2008; Geurts et al., 2006; Joshi et al., 2012; Stoltzfus, 2011; Y. Wu et al., 2002). We trained models using these ML algorithms on feature sets defined above. We evaluated a diverse set of algorithms to identify which classifiers and features yield the best predictive performance.

B) DL Models

We applied deep learning approaches, specifically convolutional neural networks (CNNs), for classifying exosomal vs non-exosomal miRNA sequences. We used the one-hot encoded sequence vectors for sequence-based classification as input to a CNN. Additionally, we trained CNN and ResNet models on the RNA secondary structure images of miRNAs generated in Section 5.2.3.1. ResNet (Residual Network) employs "shortcut" connections to facilitate training very deep networks; we used the 50-layer ResNet50 variant, which balances depth with manageable complexity (Chauhan et al., 2018; He et al., 2016; Kalchbrenner et al., 2014; LeCun et al., 1989; Lecun et al., 2015). In summary, we trained: (i) a CNN on one-hot sequence data, (ii) a CNN on structure images, and (iii) a ResNet50 on structure images.

C) LLM Models

Large language models (LLMs) are powerful in capturing sequence patterns due to deep contextual learning and have been successfully fine-tuned for sequence classification tasks (Rathore et al., n.d.). We fine-tuned pre-trained LLMs, specifically BERT-based models such as BERT-base uncased and DNABERT, for miRNA sequence classification. After fine-tuning, these models can serve as classifiers to predict if a given miRNA is exosomal, based on the learned contextual embeddings.

5.2.3.3 Cross-validation and performance metrics

We split the complete dataset containing 1912 miRNA sequences into training and validation sets in an 80:20 ratio. We employed five-fold cross-validation as explained in the methods section of Chapter 4. We evaluated model performance using both threshold-independent and dependent metrics as described previously in Chapter 4. (sensitivity, specificity, accuracy, and MCC) and threshold-independent metrics (AUROC).

5.2.4 Ensemble method

We attempted to improve AI-based model performance based on the best set of features by developing an ensemble prediction method. This ensemble approach combines three approaches: (i) motif-search based approach, (ii) similarity-search approach using BLAST, and (iii) predictions from the best-performing ML/DL model. We developed a weighted scoring system for this approach. We assigned a score of +0.5 to an miRNA sequence if it contains an exosomal motif and a score of 0 if no exosomal motif is found. Similarly, we add +0.5 if a sequence is similar to a known exosomal miRNA via BLAST (Section 2.2); otherwise, a 0 score is added. We then add these motifs and similarity scores to the prediction score from the best ML/DL model (between 0 and 1). In this way, the ensemble score ranges from 0 to 2, integrating scores from motif-search, sequence similarity, and machine learning prediction (Equations 8 and 9). This ensemble approach leverages the high precision of motif/similarity methods when known patterns are present, while relying on AI-based predictions to capture sequences lacking known motifs or homologs.

$$E' = \begin{cases} E + 0.5 & If\ exosomal\ motif\ is\ present \\ E & if\ no\ exosomal\ motif\ present \end{cases}$$ - Eq. 8

Here, E is the probability score given by the best performing AI model and E' is the score after adding motif-based approach scores to AI prediction probability score.

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

Here, E'' is the final score obtained by adding the prediction probability of the best performing ML/DL model with motif-search based approach score, and BLAST-based approach score, which ranges from 0 to 2

5.3 Results

In this study, we used multiple approaches to predict exosomal miRNAs, which can be grouped into three classes: (i) alignment-based methods, (ii) AI-based methods, and (iii) an ensemble method. The alignment-based methods include similarity-search using BLAST and motif-search using MERCI. The AI-based approaches involve deep learning, machine learning, and large language model-based models using various features and feature combinations. Finally, we devised an ensemble method that integrates the strengths of both alignment-based and AI-based techniques.

5.3.1 Alignment-based classification methods

5.3.1.1 Motif-Search

In our analysis, we discovered sequence motifs enriched in exosomal miRNAs using different parameters in MERCI software. At gap=0, about 11 unique motifs were identified, covering 26 and 8 exosomal sequences in the training and validation sets, respectively. Allowing a single gap (gap=1) in motifs also yielded 11 motifs, covering 30 exosomal sequences in training set exosomal sequences and 10 sequences in validation set. Table 5.1 presents the results of motif discovery under various gap settings, along with the number of sequences in the validation set containing each motif.

Table 5.1: The identified exosomal motifs in the training set and their distribution in the validation dataset (Here, E and NE stand for exosomal and non-exosomal, respectively) [https://doi.org/10.1101/2024.06.20.599824]

Motifs (Sequence)	No of sequences covered in validation set	Motifs	No of sequences covered in validation set
AGGGAAGC	2E	AGGGAAgapGC	2E
GAAGCAC	2E, 1NE	AgapGGGAAGC	2E
GGAAGCAC	1E	CCCACCgapC	2E
ACAAAA	1E	CCCgapACCC	2E
CCCACCC	2E	GgapAAGCAC	4E, 1NE
AGGGAAgapC	2E, 1NE	GGgapAGCAC	1E
AGGGAAGC	2E	GGGAAGCgapC	2E
AGGGAAGCgapC	2E	AgapAGGGAAG	2E
AGGGAAGgapC	2E	CAAAGgapGC	3E
AGGgapGAAGC	2E	GGCgapGGCG	1E

5.3.1.2 Similarity-search using BLAST

We next evaluated a similarity-search based approach using BLAST. Exosomal and non-exosomal miRNA sequences in the training set were used as a database for running BLAST. While assessing the training set performance, we excluded the self-hits and considered only the top BLAST hit for each query. We considered top hits for the independent validation set to evaluate the results. We measured performance across E-value thresholds from 10^{-6} to 10^{6}.

Table 5.2 summarizes the number of correct and incorrect classifications (hits) for exosomal miRNAs in both training and validation sets at various E-value cutoffs. As expected, very stringent E-values (e.g., 10^{-6}) yield fewer hits but high precision, whereas more permissive E-values (up to 10^{6}) increase coverage but include more false hits. This method yielded the best performance at an e-value of 10^{-2}, resulting in 233 true positives and 85 false positives for exosomal miRNA sequences in the training dataset, and 66 true positives alongside 24 false positives in the validation dataset. Lower e-values ($<10^{-2}$) failed to provide sufficient coverage, while higher e-values ($>10^{-2}$) led to increased error rates.

Table 5.2: The results for similarity-search using BLAST for different e-values ranging from 10^{-6} to 10^{6} [https://doi.org/10.1101/2024.06.20.599824]

	Training Dataset				Validation set			
e-value	**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

5.3.2 AI-based classification methods

5.3.2.1 ML Models

We trained various machine learning classifiers using the features generated in this study, as described in the methods section. These ML models include DT, XGB, SVC, KNN, LR, ET, and RF classifiers.

A) Composition-based Features

We examined a variety of composition features, including nucleotide compositions, correlation, entropy, pseudo composition, etc. The results for composition-based features are given in Tables 5.3 and 5.4. We computed the nucleotide composition of sequences and their reverse complements for different k-mers. The best performing models included - a random forest model using reverse-complement k-mer composition features for k=3 and k=4 (denoted RDK-3 & RDK-4), which achieved an AUROC of 0.677, and a KNN classifier using TF–IDF features (term frequencies of k-mers), which achieved an AUROC of 0.656 on the independent validation set.

Table 5.3: Results for composition-based features: composition of nucleotides of sequences (CDK), their reverse complement (RDK), Shannon Entropy, Pseudo dinucleotide composition (PDNC, PC_PDNC, SC_PDNC), Autocorrelation (DAC, DCC, DACC)

CDK-1 = (composition of nucleotides, k-mer = 1)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.50	54.994	54.617	54.807	0.565	0.096	55.135	51.010	53.003	0.556	0.061
RF	0.51	54.604	56.201	55.396	0.585	0.108	56.216	53.535	54.830	0.556	0.097
LR	0.50	55.253	54.749	55.003	0.544	0.100	56.757	52.525	54.569	0.544	0.093
XGB	0.51	52.140	56.728	54.415	0.570	0.089	63.243	45.455	54.047	0.577	0.088
KN	0.50	57.458	53.166	55.330	0.582	0.106	61.081	43.939	52.219	0.550	0.051
GNB	0.51	55.383	53.562	54.480	0.533	0.089	55.135	49.495	52.219	0.537	0.046
SVC	0.51	53.826	59.367	56.573	0.581	0.132	55.676	55.051	55.352	0.583	0.107
ET	0.50	54.734	58.179	56.442	0.589	0.129	57.838	53.535	55.614	0.563	0.114
CDK-2 = (composition of nucleotides, k-mer = 2)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.49	54.215	53.430	53.826	0.561	0.076	65.405	40.404	52.480	0.541	0.060
RF	0.51	57.328	55.937	56.638	0.592	0.133	61.622	47.475	54.308	0.596	0.092
LR	0.50	52.918	51.055	51.995	0.537	0.040	61.622	51.010	56.136	0.561	0.127
XGB	0.52	55.383	57.916	56.638	0.587	0.133	58.919	47.980	53.264	0.562	0.069
KN	0.50	57.977	56.596	57.292	0.585	0.146	62.703	51.515	56.919	0.599	0.143
GNB	0.51	50.713	54.090	52.387	0.544	0.048	59.459	46.465	52.742	0.572	0.060
SVC	0.50	77.173	23.351	50.491	0.495	0.006	55.070	48.790	48.303	0.520	0.050
ET	0.50	57.198	53.958	55.592	0.583	0.112	67.027	46.465	56.397	0.592	0.138

CDK-3 = (composition of nucleotides, k-mer = 3)											
	Training Set					Independent Validation Set					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.49	55.123	54.881	55.003	0.572	0.100	52.973	52.020	52.480	0.558	0.050
RF	0.51	58.885	57.256	58.077	0.626	0.161	60.541	59.596	60.052	0.629	0.201
LR	0.50	58.495	56.992	57.750	0.596	0.155	54.595	52.525	53.525	0.561	0.071
XGB	0.51	57.847	56.992	57.423	0.606	0.148	57.297	61.111	59.269	0.620	0.184
KN	0.49	59.533	59.631	59.581	0.630	0.192	61.081	58.081	59.530	0.630	0.192
GNB	0.37	56.809	56.728	56.769	0.584	0.135	55.135	50.000	52.480	0.560	0.051
SVC	0.51	57.069	61.741	59.385	0.628	0.188	56.757	57.576	57.180	0.611	0.143
ET	0.50	60.700	59.367	60.039	0.638	0.201	60.541	54.545	57.441	0.619	0.151
CDK-4 = (composition of nucleotides, k-mer = 4)											
	Training Set					Independent Validation Set					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.50	51.621	55.805	53.695	0.537	0.074	42.703	62.626	53.003	0.541	0.054
RF	0.50	59.014	59.631	59.320	0.626	0.186	61.081	55.556	58.225	0.634	0.166
LR	0.50	57.458	58.839	58.143	0.605	0.163	59.459	57.071	58.225	0.607	0.165
XGB	0.51	57.588	58.707	58.143	0.612	0.163	63.243	55.051	59.008	0.628	0.183
KN	0.49	60.311	55.805	58.077	0.615	0.161	63.243	54.545	58.747	0.632	0.178
GNB	0.03	57.328	58.047	57.685	0.604	0.154	51.892	57.071	54.569	0.564	0.090
SVC	0.51	56.939	61.346	59.124	0.622	0.183	61.081	60.101	60.574	0.630	0.212
ET	0.50	58.885	60.422	59.647	0.629	0.193	55.676	58.081	56.919	0.611	0.138
RDK-1 = (composition of nucleotides in reverse complementary sequence, k-mer = 1)											
	Training Set					Independent Validation Set					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	52.918	51.055	51.995	0.527	0.040	63.243	39.394	50.914	0.516	0.027
RF	0.50	52.918	50.528	51.733	0.528	0.034	58.378	44.444	51.175	0.515	0.028
LR	0.50	43.969	54.485	49.182	0.497	-0.016	49.189	56.566	53.003	0.526	0.058
XGB	0.50	55.383	49.077	52.256	0.528	0.045	56.757	40.909	48.564	0.506	-0.024
KN	0.51	51.881	53.958	52.910	0.521	0.058	55.135	45.455	50.131	0.506	0.006
GNB	0.51	40.856	54.222	47.482	0.488	-0.050	57.838	47.980	52.742	0.514	0.058
SVC	0.51	54.864	54.090	54.480	0.549	0.090	57.297	48.485	52.742	0.533	0.058
ET	0.50	51.751	52.902	52.322	0.534	0.047	56.757	46.465	51.436	0.516	0.032
RDK-2 = (composition of nucleotides in reverse complementary sequence, k-mer = 2)											
	Training Set					Independent Validation Set					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.49	56.809	55.937	56.377	0.567	0.127	62.703	51.515	56.919	0.600	0.143
RF	0.49	57.458	55.409	56.442	0.593	0.129	59.459	48.990	54.047	0.603	0.085
LR	0.49	57.069	51.715	54.415	0.557	0.088	58.919	47.475	53.003	0.582	0.064

XGB	0.49	57.717	56.596	57.162	0.593	0.143	60.541	52.020	56.136	0.588	0.126
KN	0.48	57.328	57.256	57.292	0.602	0.146	60.541	55.051	57.702	0.605	0.156
GNB	0.46	56.420	55.277	55.853	0.578	0.117	57.297	55.051	56.136	0.601	0.123
SVC	0.50	58.106	56.596	57.358	0.609	0.147	65.405	57.071	61.097	0.639	0.225
ET	0.49	57.717	55.937	56.835	0.604	0.137	61.622	51.515	56.397	0.611	0.132
RDK-3 = (composition of nucleotides in reverse complementary sequence, k-mer = 3)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.50	52.789	51.979	52.387	0.542	0.048	51.351	57.576	54.569	0.537	0.089
RF	0.51	58.495	60.554	59.516	0.632	0.191	60.541	61.616	61.097	0.654	0.221
LR	0.50	54.215	53.694	53.957	0.576	0.079	57.297	55.051	56.136	0.588	0.123
XGB	0.51	58.885	57.916	58.404	0.616	0.168	57.297	55.556	56.397	0.602	0.128
KN	0.50	58.236	58.839	58.535	0.623	0.171	65.405	58.586	61.88	0.639	0.24
GNB	0.42	56.55	55.409	55.984	0.571	0.12	56.757	53.535	55.091	0.586	0.103
SVC	0.50	60.83	56.596	58.731	0.627	0.174	67.027	57.071	61.88	0.642	0.242
ET	0.50	60.83	57.784	59.32	0.63	0.186	62.162	58.586	60.313	0.663	0.207
RDK-4 = (composition of nucleotides in reverse complementary sequence, k-mer = 4)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.49	51.232	57.652	54.415	0.552	0.089	45.405	65.152	55.614	0.540	0.108
RF	0.50	58.755	59.499	59.124	0.631	0.183	58.378	56.566	57.441	0.636	0.149
LR	0.50	54.734	57.784	56.246	0.585	0.125	57.297	58.081	57.702	0.611	0.154
XGB	0.49	55.772	54.749	55.265	0.593	0.105	62.162	48.990	55.352	0.584	0.112
KN	0.49	59.274	56.464	57.881	0.620	0.157	65.405	54.040	59.530	0.649	0.195
GNB	0.11	56.161	56.464	56.311	0.595	0.126	54.054	54.545	54.308	0.571	0.086
SVC	0.50	58.655	58.499	58.124	0.627	0.170	57.378	57.566	56.441	0.630	0.140
ET	0.50	60.571	58.575	59.581	0.636	0.191	64.324	55.051	59.530	0.643	0.194
RDK-3 + RDK-4											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.53	54.864	49.208	52.060	0.529	0.041	58.919	52.020	55.352	0.559	0.110
RF	0.50	58.885	56.992	57.946	0.627	0.159	61.622	63.636	62.663	0.677	0.253
LR	0.50	57.069	56.596	56.835	0.596	0.137	56.216	60.101	58.225	0.602	0.163
XGB	0.50	58.495	56.728	57.619	0.601	0.152	58.378	58.586	58.486	0.620	0.170
KN	0.48	60.311	57.520	58.927	0.624	0.178	65.405	53.535	59.269	0.644	0.191
GNB	0.08	55.901	56.464	56.181	0.594	0.124	52.973	55.051	54.047	0.568	0.080
SVC	0.50	60.655	59.499	59.124	0.637	0.170	57.578	57.576	56.541	0.630	0.142
ET	0.50	59.663	59.367	59.516	0.639	0.190	62.703	59.091	60.836	0.662	0.218
Shannon Entropy											

	Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.5	55.642	56.728	56.181	0.567	0.124	51.892	61.111	56.658	0.558	0.131
RF	0.49	56.291	55.541	55.919	0.577	0.118	56.216	55.051	55.614	0.555	0.113
LR	0.5	68.093	40.501	54.415	0.546	0.089	56.757	50.505	53.525	0.522	0.073
XGB	0.5	56.161	56.464	56.311	0.59	0.126	63.784	47.98	55.614	0.585	0.119
KN	0.5	57.588	53.694	55.657	0.586	0.113	58.919	48.485	53.525	0.58	0.074
GNB	0.48	55.772	51.319	53.564	0.542	0.071	38.378	66.162	52.742	0.53	0.047
SVC	0.51	44.266	62.629	53.022	0.554	0.082	49.108	56.078	53.744	0.556	0.051
ET	0.5	58.495	59.631	59.058	0.632	0.181	62.703	53.535	57.963	0.598	0.163
PDNC (Pseudo dinucleotide composition)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	53.437	55.145	54.284	0.537	0.086	61.081	49.495	55.091	0.54	0.106
RF	0.51	57.069	58.839	57.946	0.614	0.159	61.081	55.556	58.225	0.599	0.166
LR	0.5	58.236	54.354	56.311	0.574	0.126	55.676	53.03	54.308	0.56	0.087
XGB	0.49	57.458	56.86	57.162	0.592	0.143	58.378	52.02	55.091	0.574	0.104
KN	0.49	58.885	59.894	59.385	0.623	0.188	64.324	55.051	59.53	0.594	0.194
GNB	0.49	56.68	56.201	56.442	0.578	0.129	58.919	49.495	54.047	0.568	0.084
SVC	0.51	48.174	55.311	51.181	0.521	0.042	45.865	66.192	59.441	0.587	0.15
ET	0.5	58.625	58.707	58.666	0.625	0.173	62.162	51.01	56.397	0.58	0.132
PC_PDNC (parallel correlation pseudo dinucleotide composition)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.47	54.475	54.881	54.676	0.556	0.094	58.378	46.465	52.219	0.531	0.049
RF	0.5	55.772	55.541	55.657	0.599	0.113	59.459	55.556	57.441	0.584	0.15
LR	0.5	57.328	58.311	57.816	0.588	0.156	52.973	50.505	51.697	0.539	0.035
XGB	0.5	55.642	54.354	55.003	0.577	0.1	57.297	52.02	54.569	0.564	0.093
KN	0.49	58.106	59.499	58.797	0.618	0.176	59.459	53.535	56.397	0.606	0.13
GNB	0.47	56.68	56.069	56.377	0.577	0.127	56.757	51.515	54.047	0.563	0.083
SVC	0.51	45.266	62.929	54.022	0.549	0.083	48.108	57.071	52.742	0.555	0.052
ET	0.5	58.366	58.179	58.273	0.613	0.165	57.838	52.525	55.091	0.578	0.104
SC-PDNC (Serial correlation pseudo dinucleotide composition)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.49	54.734	53.826	54.284	0.570	0.086	49.73	53.03	51.436	0.524	0.028
RF	0.51	58.495	61.346	59.908	0.631	0.198	59.459	55.051	57.18	0.602	0.145
LR	0.50	55.772	56.069	55.919	0.573	0.118	55.676	52.525	54.047	0.562	0.082
XGB	0.50	57.328	56.992	57.162	0.599	0.143	60.541	59.091	59.791	0.61	0.196

KN	0.49	57.458	59.367	58.404	0.622	0.168	61.081	54.545	57.702	0.601	0.156
GNB	0.49	56.42	56.069	56.246	0.580	0.125	59.459	51.515	55.352	0.573	0.110
SVC	0.51	57.495	60.346	58.908	0.630	0.188	58.459	53.051	56.18	0.600	0.140
ET	0.50	58.495	59.631	59.058	0.632	0.181	62.703	53.535	57.963	0.598	0.163
DAC (Dinucleotide based auto correlation)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.5	55.123	54.222	54.676	0.558	0.093	36.216	67.172	52.219	0.536	0.036
RF	0.5	57.717	56.992	57.358	0.617	0.147	54.054	50.505	52.219	0.564	0.046
LR	0.5	55.383	55.277	55.33	0.567	0.107	57.297	54.545	55.875	0.577	0.118
XGB	0.5	55.772	56.069	55.919	0.581	0.118	51.892	54.04	53.003	0.541	0.059
KN	0.51	50.454	54.354	52.387	0.548	0.048	48.649	48.485	48.564	0.512	-0.029
GNB	0.34	52.14	52.375	52.256	0.549	0.045	52.973	54.545	53.786	0.549	0.075
SVC	0.51	46.174	58.311	52.191	0.527	0.045	44.865	69.192	57.441	0.597	0.145
ET	0.5	55.512	59.103	57.292	0.622	0.146	53.514	51.01	52.219	0.553	0.045
DCC (Dinucleotide based cross correlation)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.47	53.178	51.715	52.453	0.536	0.049	51.351	56.061	53.786	0.530	0.074
RF	0.50	59.274	57.520	58.404	0.645	0.168	57.838	51.010	54.308	0.562	0.089
LR	0.49	59.144	58.311	58.731	0.616	0.175	63.784	54.545	59.008	0.603	0.184
XGB	0.51	59.274	59.763	59.516	0.635	0.190	54.595	59.091	56.919	0.600	0.137
KN	0.51	55.901	57.256	56.573	0.593	0.132	53.514	50.000	51.697	0.529	0.035
GNB	0.02	58.106	57.388	57.750	0.590	0.155	58.378	45.455	51.697	0.529	0.039
SVC	0.51	55.512	59.367	57.423	0.606	0.149	54.054	55.556	54.830	0.556	0.096
ET	0.50	59.403	60.686	60.039	0.634	0.201	61.081	51.010	55.875	0.569	0.121
DACC (Dinucleotide based auto cross correlation)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.48	54.086	53.562	53.826	0.544	0.076	51.892	47.475	49.608	0.488	-0.006
RF	0.50	61.997	58.047	60.039	0.646	0.201	57.297	54.040	55.614	0.568	0.113
LR	0.50	58.755	58.047	58.404	0.622	0.168	65.405	53.535	59.269	0.626	0.191
XGB	0.50	58.885	59.235	59.058	0.634	0.181	55.676	54.545	55.091	0.584	0.102
KN	0.51	58.625	54.881	56.769	0.606	0.135	57.838	49.495	53.525	0.541	0.074
GNB	0.02	56.550	58.443	57.489	0.590	0.150	55.676	46.970	51.175	0.532	0.027
SVC	0.51	56.161	61.873	58.993	0.632	0.181	58.919	54.545	56.658	0.568	0.135
ET	0.50	58.106	59.631	58.862	0.634	0.177	57.838	56.566	57.180	0.575	0.144

Table 5.4: Results for composition-based features: TFIDF for miRNA sequences and their reverse complementary sequences from kmer 1 to 7

TFIDF (1,1)											
	Training Set						Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.51	53.437	56.332	54.872	0.540	0.098	57.297	53.030	55.091	0.552	0.103
SVC	0.50	56.420	56.069	56.246	0.576	0.125	62.162	48.485	55.091	0.592	0.107
KNN	0.41	54.604	51.847	53.237	0.536	0.065	55.135	59.091	57.180	0.587	0.142
DT	0.51	47.471	53.166	50.294	0.495	0.006	44.865	59.091	52.219	0.523	0.040
RF	0.51	52.918	53.034	52.976	0.540	0.060	49.730	56.566	53.264	0.546	0.063
GNB	0.51	53.696	55.145	54.415	0.534	0.088	55.676	53.535	54.569	0.549	0.092
XGB	0.52	52.270	53.034	52.649	0.539	0.053	51.351	55.556	53.525	0.551	0.069
ET	0.51	49.935	56.464	53.172	0.064	0.543	48.649	58.586	53.786	0.542	0.073
TFIDF (1,2)											
	Training Set						Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.51	52.399	57.388	54.872	0.570	0.098	54.054	54.040	54.047	0.570	0.081
SVC	0.51	56.291	61.346	58.797	0.619	0.177	60.000	55.556	57.702	0.617	0.156
KNN	0.41	54.864	58.575	56.704	0.592	0.134	58.378	48.485	53.264	0.567	0.069
DT	0.01	53.307	53.034	53.172	0.532	0.063	52.432	52.020	52.219	0.522	0.045
RF	0.50	58.106	56.992	57.554	0.609	0.151	63.243	50.505	56.658	0.607	0.138
GNB	0.51	54.475	56.596	55.526	0.575	0.111	57.297	52.020	54.569	0.577	0.093
XGB	0.50	54.734	55.409	55.069	0.579	0.101	64.324	51.010	57.441	0.598	0.155
ET	0.51	59.274	57.652	58.470	0.169	0.625	60.000	51.010	55.352	0.592	0.110
TFIDF (1,3)											
	Training Set						Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.51	54.864	56.860	55.853	0.594	0.117	51.892	53.030	52.480	0.565	0.049
SVC	0.51	57.847	62.005	59.908	0.635	0.199	61.081	61.111	61.097	0.634	0.222
KNN	0.41	56.939	59.367	58.143	0.602	0.163	71.351	58.586	64.752	0.656	0.301
DT	0.01	55.772	54.485	55.134	0.551	0.103	54.054	58.586	56.397	0.563	0.126
RF	0.52	57.328	60.290	58.797	0.624	0.176	57.297	62.121	59.791	0.632	0.194
GNB	0.46	56.809	56.728	56.769	0.588	0.135	52.432	51.515	51.958	0.564	0.039
XGB	0.51	57.198	57.520	57.358	0.605	0.147	62.162	54.545	58.225	0.611	0.167
ET	0.51	60.311	59.894	60.105	0.202	0.635	56.757	63.131	60.052	0.638	0.199
TFIDF (1,4)											
	Training Set						Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**

LR	0.51	57.198	58.443	57.816	0.611	0.156	60.541	59.596	60.052	0.606	0.201
SVC	0.50	61.868	58.047	59.974	0.641	0.199	64.324	56.061	60.052	0.638	0.204
KNN	0.41	56.809	56.596	56.704	0.604	0.134	64.324	54.040	59.008	0.622	0.184
DT	0.01	52.270	53.826	53.041	0.530	0.061	56.757	45.960	51.175	0.514	0.027
RF	0.51	59.274	57.124	58.208	0.631	0.164	60.541	54.040	57.180	0.629	0.146
GNB	0.09	57.717	58.047	57.881	0.604	0.158	56.216	55.556	55.875	0.582	0.118
XGB	0.54	56.031	56.728	56.377	0.601	0.128	56.216	55.051	55.614	0.606	0.113
ET	0.51	58.495	59.631	59.058	0.181	0.633	55.676	60.101	57.963	0.637	0.158
TFIDF (1,5)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.51	57.847	60.158	58.993	0.618	0.180	57.297	60.606	59.008	0.613	0.179
SVC	0.50	62.257	57.256	59.778	0.631	0.195	60.541	54.545	57.441	0.634	0.151
KNN	0.41	58.885	56.332	57.619	0.602	0.152	66.486	52.020	59.008	0.627	0.187
DT	0.01	58.495	53.166	55.853	0.558	0.117	54.595	51.010	52.742	0.528	0.056
RF	0.51	58.106	58.971	58.535	0.631	0.171	63.243	59.091	61.097	0.649	0.223
GNB	0.01	33.722	77.836	55.592	0.588	0.129	31.351	75.758	54.308	0.592	0.079
XGB	0.51	57.847	56.992	57.423	0.603	0.148	64.324	50.505	57.180	0.631	0.150
ET	0.50	58.885	60.290	59.581	0.192	0.617	59.459	57.071	58.225	0.628	0.165
TFIDF (1,6)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.51	57.328	60.686	58.993	0.620	0.180	56.757	60.101	58.486	0.620	0.169
SVC	0.50	62.516	56.332	59.451	0.625	0.189	62.162	56.566	59.269	0.636	0.187
KNN	0.41	59.144	54.749	56.965	0.590	0.139	63.784	54.040	58.747	0.613	0.179
DT	0.01	55.253	53.826	54.545	0.545	0.091	61.081	55.556	58.225	0.583	0.166
RF	0.51	57.847	60.554	59.189	0.627	0.184	58.919	56.566	57.702	0.629	0.155
GNB	0.01	49.676	58.443	54.022	0.539	0.081	44.865	62.626	54.047	0.551	0.076
XGB	0.52	57.977	56.860	57.423	0.603	0.148	59.459	56.566	57.963	0.617	0.160
ET	0.50	58.755	57.388	58.077	0.161	0.611	62.162	56.061	59.008	0.633	0.182
TFIDF (1,7)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.50	61.479	57.784	59.647	0.623	0.193	60.541	57.576	59.008	0.630	0.181
SVC	0.50	61.479	58.839	60.170	0.627	0.203	61.622	58.586	60.052	0.645	0.202
KNN	0.41	59.403	54.354	56.900	0.593	0.138	68.649	56.566	62.402	0.625	0.254
DT	0.01	55.901	53.430	54.676	0.547	0.093	58.378	49.495	53.786	0.539	0.079
RF	0.51	59.144	59.367	59.254	0.629	0.185	58.919	58.586	58.747	0.628	0.175
GNB	0.01	54.864	53.166	54.022	0.540	0.080	58.378	53.535	55.875	0.559	0.119

XGB	0.52	57.588	57.124	57.358	0.597	0.147	58.378	58.586	58.486	0.604	0.170
ET	0.50	57.588	57.652	57.619	0.152	0.610	60.541	55.051	57.702	0.634	0.156
TFIDF RC – (1,1)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.51	53.437	56.332	54.872	0.540	0.098	57.297	53.030	55.091	0.552	0.103
SVC	0.50	56.420	56.069	56.246	0.576	0.125	62.162	48.485	55.091	0.592	0.107
KNN	0.41	54.604	51.847	53.237	0.536	0.065	55.135	59.091	57.180	0.587	0.142
DT	0.51	47.990	53.034	50.491	0.498	0.010	41.622	60.101	51.175	0.512	0.018
RF	0.51	53.437	53.166	53.303	0.540	0.066	48.649	55.556	52.219	0.549	0.042
GNB	0.51	53.696	55.145	54.415	0.534	0.088	55.676	53.535	54.569	0.549	0.092
XGB	0.50	53.826	53.694	53.761	0.544	0.075	54.595	51.515	53.003	0.550	0.061
ET	0.51	49.157	56.069	52.583	0.052	0.540	47.568	60.606	54.308	0.548	0.082
TFIDF RC – (1,2)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.51	52.399	57.388	54.872	0.570	0.098	54.054	54.040	54.047	0.570	0.081
SVC	0.51	56.291	61.346	58.797	0.619	0.177	60.000	55.556	57.702	0.617	0.156
KNN	0.41	54.864	58.575	56.704	0.592	0.134	58.378	48.485	53.264	0.567	0.069
DT	0.01	52.659	53.166	52.910	0.529	0.058	54.595	51.010	52.742	0.528	0.056
RF	0.51	57.977	57.124	57.554	0.613	0.151	62.703	53.030	57.702	0.617	0.158
GNB	0.51	54.475	56.596	55.526	0.575	0.111	57.297	52.020	54.569	0.577	0.093
XGB	0.49	55.642	55.541	55.592	0.591	0.112	62.162	52.020	56.919	0.611	0.142
ET	0.51	58.625	57.124	57.881	0.158	0.618	63.784	54.545	59.008	0.623	0.184
TFIDF RC – (1,3)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.51	54.864	56.860	55.853	0.594	0.117	51.892	53.030	52.480	0.565	0.049
SVC	0.51	57.847	62.005	59.908	0.635	0.199	61.081	61.111	61.097	0.634	0.222
KNN	0.41	56.939	59.367	58.143	0.602	0.163	71.351	58.586	64.752	0.656	0.301
DT	0.01	55.253	55.277	55.265	0.553	0.105	52.432	60.101	56.397	0.563	0.126
RF	0.51	59.533	57.784	58.666	0.622	0.173	63.784	56.566	60.052	0.632	0.204
GNB	0.46	56.809	56.728	56.769	0.588	0.135	52.432	51.515	51.958	0.564	0.039
XGB	0.51	58.755	58.575	58.666	0.605	0.173	62.703	53.030	57.702	0.620	0.158
ET	0.51	61.608	62.005	61.805	0.236	0.652	58.919	58.586	58.747	0.629	0.175
TFIDF RC – (1,4)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.51	57.198	58.443	57.816	0.611	0.156	60.541	59.596	60.052	0.606	0.201

SVC	0.50	61.868	58.047	59.974	0.641	0.199	64.324	56.061	60.052	0.638	0.204
KNN	0.41	56.809	56.596	56.704	0.604	0.134	64.324	54.040	59.008	0.622	0.184
DT	0.01	52.529	52.902	52.714	0.527	0.054	54.595	44.949	49.608	0.498	-0.005
RF	0.51	58.625	59.894	59.254	0.629	0.185	64.865	58.586	61.619	0.648	0.235
GNB	0.09	57.717	58.047	57.881	0.604	0.158	56.216	55.556	55.875	0.582	0.118
XGB	0.51	56.939	57.388	57.162	0.609	0.143	63.243	55.556	59.269	0.637	0.188
ET	0.50	59.533	57.520	58.535	0.171	0.620	60.000	52.525	56.136	0.619	0.125
TFIDF RC – (1,5)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.51	57.847	60.158	58.993	0.618	0.180	57.297	60.606	59.008	0.613	0.179
SVC	0.50	62.257	57.256	59.778	0.631	0.195	60.541	54.545	57.441	0.634	0.151
KNN	0.41	58.885	56.332	57.619	0.602	0.152	66.486	52.020	59.008	0.627	0.187
DT	0.01	56.550	52.507	54.545	0.545	0.091	60.000	51.010	55.352	0.555	0.110
RF	0.51	58.106	59.235	58.666	0.627	0.173	61.081	58.586	59.791	0.635	0.197
GNB	0.01	33.722	77.836	55.592	0.588	0.129	31.351	75.758	54.308	0.592	0.079
XGB	0.51	56.031	56.860	56.442	0.594	0.129	61.081	53.030	56.919	0.621	0.141
ET	0.50	59.533	58.575	59.058	0.181	0.626	58.919	56.566	57.702	0.630	0.155
TFIDF RC – (1,6)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.51	57.328	60.686	58.993	0.620	0.180	56.757	60.101	58.486	0.620	0.169
SVC	0.50	62.516	56.332	59.451	0.625	0.189	62.162	56.566	59.269	0.636	0.187
KNN	0.41	59.144	54.749	56.965	0.590	0.139	63.784	54.040	58.747	0.613	0.179
DT	0.01	58.236	51.583	54.938	0.549	0.098	58.378	56.061	57.180	0.572	0.144
RF	0.51	60.311	59.367	59.843	0.633	0.197	63.784	56.061	59.791	0.646	0.199
GNB	0.01	49.676	58.443	54.022	0.539	0.081	44.865	62.626	54.047	0.551	0.076
XGB	0.51	57.977	57.520	57.750	0.609	0.155	63.784	52.020	57.702	0.630	0.159
ET	0.51	56.809	58.575	57.685	0.154	0.606	61.081	57.576	59.269	0.633	0.187
TFIDF RC – (1,7)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
LR	0.5	61.479	57.784	59.647	0.623	0.193	60.541	57.576	59.008	0.630	0.181
SVC	0.5	61.479	58.839	60.170	0.627	0.203	61.622	58.586	60.052	0.645	0.202
KNN	0.41	59.403	54.354	56.900	0.593	0.138	68.649	56.566	62.402	0.625	0.254
DT	0.01	53.956	53.298	53.630	0.536	0.073	57.838	45.960	51.697	0.519	0.038
RF	0.51	57.977	59.367	58.666	0.625	0.173	64.324	55.556	59.791	0.640	0.199
GNB	0.01	54.864	53.166	54.022	0.540	0.080	58.378	53.535	55.875	0.559	0.119
XGB	0.53	56.809	57.652	57.227	0.602	0.145	54.595	56.566	55.614	0.592	0.112

ET	0.5	57.069	58.311	57.685	0.154	0.605	56.757	54.545	55.614	0.624	0.113

B) Binary profile

We computed binary profiles for each miRNA sequence using one-hot encoding. We generated binary profiles for mononucleotides, dinucleotides, and trinucleotides. We used one-hot encoded sequence vectors as features and developed several ML models on them. An SVM-based classifier performed the best on these binary features, reaching an AUROC of 0.642 on the validation dataset. The results are given in Table 5.5.

Table 5.5: Results for binary-profile based features for mononucleotides, dinucleotides, trinucleotides, and a combination of all three (mono, di, and tri)

Binary features - mononucleotides (Padding with X)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	54.086	54.485	54.284	0.553	0.086	58.378	53.030	55.614	0.571	0.114
RF	0.50	61.608	59.631	60.628	0.640	0.212	58.919	63.636	61.358	0.634	0.226
LR	0.50	54.994	54.354	54.676	0.578	0.093	52.432	60.101	56.397	0.560	0.126
XGB	0.50	57.198	57.124	57.162	0.614	0.143	56.757	52.020	54.308	0.571	0.088
KN	0.51	55.512	59.763	57.619	0.613	0.153	58.919	60.101	59.530	0.614	0.190
GNB	0.84	49.935	49.472	49.706	0.489	-0.006	35.135	74.242	55.352	0.555	0.102
SVC	0.50	59.403	60.950	60.170	0.634	0.204	61.081	62.121	61.619	0.642	0.232
ET	0.50	56.420	59.499	57.946	0.628	0.159	58.378	65.657	62.141	0.633	0.241
Binary features - Dinucleotide (Padding with X)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.53	55.383	53.430	54.415	0.563	0.088	42.162	64.646	53.786	0.545	0.070
RF	0.50	58.366	62.005	60.170	0.643	0.204	58.378	57.576	57.963	0.625	0.159
LR	0.49	57.847	58.839	58.339	0.609	0.167	49.730	50.000	49.869	0.527	-0.003
XGB	0.50	58.755	59.631	59.189	0.628	0.184	54.054	56.566	55.352	0.581	0.106
KN	0.50	61.219	55.805	58.535	0.607	0.170	62.162	53.030	57.441	0.620	0.152
GNB	0.51	63.035	49.208	56.181	0.562	0.124	75.135	26.768	50.131	0.566	0.022
SVC	0.49	59.533	60.554	60.039	0.640	0.201	61.081	59.091	60.052	0.634	0.202
ET	0.50	59.792	59.499	59.647	0.634	0.193	50.811	55.556	53.264	0.585	0.064
Binary features - Trinucleotide (Padding with X)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**

DT	0.51	53.307	53.694	53.499	0.547	0.070	45.405	67.677	56.919	0.585	0.134
RF	0.50	60.700	58.707	59.712	0.625	0.194	63.243	57.071	60.052	0.634	0.203
LR	0.50	58.625	59.103	58.862	0.621	0.177	57.297	60.101	58.747	0.601	0.174
XGB	0.50	57.328	56.464	56.900	0.592	0.138	57.838	51.010	54.308	0.572	0.089
KN	0.51	53.826	58.971	56.377	0.603	0.128	51.892	64.646	58.486	0.622	0.167
GNB	0.50	60.960	53.958	57.489	0.580	0.150	44.865	66.667	56.136	0.541	0.118
SVC	0.49	59.274	59.499	59.385	0.630	0.188	62.703	62.121	62.402	0.644	0.248
ET	0.50	56.550	55.145	55.853	0.602	0.117	59.459	58.081	58.747	0.617	0.175
Binary features - Mononucleotides + Dinucleotides + Trinucleotides (Padding with X)											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.46	53.696	53.430	53.564	0.547	0.071	57.297	50.505	53.786	0.551	0.078
RF	0.50	58.885	60.158	59.516	0.639	0.190	59.459	57.576	58.486	0.635	0.170
LR	0.51	57.977	58.179	58.077	0.607	0.162	49.730	57.071	53.525	0.540	0.068
XGB	0.50	58.885	58.971	58.927	0.623	0.179	56.757	56.061	56.397	0.585	0.128
KN	0.50	60.311	55.805	58.077	0.609	0.161	59.459	56.061	57.702	0.617	0.155
GNB	0.50	61.608	52.902	57.292	0.577	0.146	47.027	60.101	53.786	0.530	0.072
SVC	0.49	59.274	59.499	59.385	0.630	0.188	62.703	62.121	62.402	0.644	0.248
ET	0.50	57.198	59.499	58.339	0.629	0.167	57.838	62.626	60.313	0.632	0.205

C) Secondary Structure-based features

We computed structural features for miRNA sequences using the RNAfold tool. These features included centroid diversity, minimum free energy, minimum free energy (MFE) frequency structure in the ensemble, centroid free energy, ensemble free energy, and ensemble diversity. An LR model trained on these structural features achieved the highest AUROC of 0.558 on the validation set, which is lower than what was achieved with sequence composition features. This suggests that while structural stability and related metrics differ somewhat between exosomal and non-exosomal miRNAs, they alone do not provide strong predictive power. The results for prediction using secondary structure-based features are given in Table 5.6.

Table 5.6: Results obtained for ML models developed on secondary structure-based descriptors

	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.49	49.676	48.417	49.052	0.497	-0.019	54.054	45.960	49.869	0.496	0.000
RF	0.49	50.195	50.660	50.425	0.507	0.009	57.838	50.505	54.047	0.535	0.084

LR	0.50	45.785	58.179	51.929	0.523	0.040	47.027	64.646	56.136	0.558	0.119
XGB	0.51	50.454	54.354	52.387	0.532	0.048	54.054	56.566	55.352	0.537	0.106
KN	0.51	46.952	53.562	50.229	0.502	0.005	51.351	56.061	53.786	0.542	0.074
GNB	0.47	50.973	54.485	52.714	0.530	0.055	42.162	65.657	54.308	0.547	0.080
SVC	0.51	18.547	81.794	49.902	0.515	0.004	0.000	100.000	51.697	0.499	0.000
ET	0.49	52.529	49.340	50.948	0.513	0.019	55.676	48.990	52.219	0.541	0.047

D) Embeddings from Large Language Models

Next, we utilized embeddings extracted from BERT models fine-tuned to our miRNA dataset. A random forest model trained on embeddings from the fine-tuned DNABERT model achieved the highest AUROC of 0.598 on the validation set. Similarly, for the BERT-base uncased model embeddings, the random forest model also performed the best, achieving an AUROC of 0.565. These performance levels show that LLM-derived features contain relevant information. However, the LLM embeddings alone did not outperform simpler composition or TF–IDF features in our study. The results for LLM embeddings used as features to train various ML classifiers are shown in Table 5.7.

Table 5.7: Results for LLM embeddings used as features: BERT base uncased, and DNABERT

DNA Bert embeddings											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.48	48.898	48.417	48.659	0.493	-0.027	52.973	51.010	51.958	0.537	0.040
RF	0.50	57.328	51.451	54.415	0.558	0.088	63.784	51.010	57.180	0.598	0.149
LR	0.51	54.604	56.201	55.396	0.570	0.108	52.432	54.040	53.264	0.540	0.065
XGB	0.51	53.307	53.562	53.434	0.538	0.069	54.054	55.556	54.830	0.550	0.096
KN	0.48	54.604	51.319	52.976	0.549	0.059	55.135	52.525	53.786	0.557	0.077
GNB	0.41	52.918	53.034	52.976	0.535	0.060	56.757	55.556	56.136	0.566	0.123
SVC	0.51	52.232	56.575	54.772	0.565	0.096	51.892	55.545	54.264	0.562	0.062
ET	0.51	50.843	56.464	53.630	0.554	0.073	51.351	61.111	56.397	0.556	0.125
BERT-base-uncased Finetuned - embeddings											
	Training Set					**Independent Validation Set**					
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.49	52.010	53.166	52.583	0.544	0.052	45.405	64.646	55.352	0.560	0.102
RF	0.51	54.864	58.839	56.835	0.582	0.137	61.622	48.990	55.091	0.565	0.107
LR	0.50	57.458	56.201	56.835	0.593	0.137	57.838	52.020	54.830	0.557	0.099
XGB	0.51	55.642	55.277	55.461	0.567	0.109	57.297	47.475	52.219	0.530	0.048

KN	0.52	54.864	53.034	53.957	0.557	0.079	58.378	44.444	51.175	0.516	0.028
GNB	0.57	53.307	53.430	53.368	0.548	0.067	60.000	50.505	55.091	0.564	0.105
SVC	0.51	54.086	58.575	56.311	0.588	0.127	58.378	50.000	54.047	0.549	0.084
ET	0.50	57.588	52.639	55.134	0.571	0.102	64.324	48.485	56.136	0.553	0.130

E) Best Features

We identified a set of "best" features by combining the most informative features from different categories: specifically, mononucleotide binary profile features, reverse-complement composition features for k-mers 3 and 4 (RDK-3 and RDK-4), and selected TF–IDF features. This combined feature set comprised 382 features. We normalized these features using standard z-score normalization. An Extra Trees classifier trained on this feature set achieved an AUROC of 0.707 on the validation set. We also conducted a Mann–Whitney U test on the 382 features to find those significantly differing between exosomal and non-exosomal classes. We found 75 features with p-value < 0.05, indicating significant differences. The results for the combined best features are provided in Table 5.8. The combined best-feature model outperformed models based on any single feature category. The Mann-Whitney test results performed on the best features are given in Table 5.9.

Table 5.8: Results for a) AI models developed on combined best-performing features, and b) ensemble model integrating Alignment-based and AI-based methods [https://doi.org/10.1101/2024.06.20.599824]

Best features combined – One hot encoding (mononucleotides) + RDK-3 + RDK-4 + Reverse complement TFIDF (1,3)											
		Training set					**Validation set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
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											
		Training set					**Validation set**				

Model	Thr	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
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 5.9: Results for Mann-Whitney U test for the best features – significantly different features (here, R_ stands for composition for reverse complementary strand features; Features like A_1, A_2,.. Stand for binary features, and features like A_rc, T_rc,.. Stand for TFIDF features for reverse complementary strand for k-mer range (1,3))

Features	Mean (E)	Mean (NE)	Mean Diff	p-Value	Features	Mean (E)	Mean (NE)	Mean Diff	p-Value
A_10	0.192	0.232	-0.040	0.0336	**R_AGA**	3.502	4.312	-0.810	0.0007
A_13	0.202	0.263	-0.061	0.0017	**R_AGAU**	0.529	0.848	-0.319	0.0005
A_20	0.168	0.204	-0.036	0.0458	**R_AGCA**	1.278	1.015	0.262	0.0480
A_21	0.161	0.217	-0.055	0.0020	**R_AGG**	4.311	4.850	-0.539	0.0238
A_22	0.133	0.180	-0.047	0.0046	**R_AGGC**	0.873	1.335	-0.462	9.33E-05
A_rc	0.298	0.286	0.012	0.0167	**R_AGUG**	1.436	1.164	0.272	0.0249
AC_rc	0.094	0.080	0.014	0.0001	**R_AUGA**	0.598	0.793	-0.195	0.0495
ACA_rc	0.051	0.044	0.008	0.0253	**R_CAAC**	0.800	0.543	0.257	0.0038
ACC_rc	0.043	0.036	0.007	0.0141	**R_CAC**	4.217	3.567	0.649	0.0002
ACG_rc	0.020	0.014	0.006	0.0195	**R_CACG**	0.463	0.289	0.174	0.0141
AGC_rc	0.046	0.041	0.005	0.0458	**R_CAG**	5.217	5.718	-0.501	0.0301
AUC_rc	0.029	0.035	-0.006	0.0431	**R_CAGA**	1.095	1.473	-0.378	0.0028
C_1	0.226	0.181	0.045	0.0146	**R_CAGG**	1.398	1.710	-0.312	0.0156
C_10	0.279	0.229	0.050	0.0117	**R_CCCG**	0.699	0.488	0.210	0.0102
C_11	0.260	0.204	0.056	0.0035	**R_CCG**	2.001	1.522	0.479	0.0016
C_19	0.207	0.258	-0.051	0.0080	**R_CGA**	1.066	0.798	0.268	0.0350
C_8	0.267	0.226	0.041	0.0385	**R_CGAA**	0.218	0.093	0.125	0.0014
CAC_rc	0.052	0.044	0.008	0.0088	**R_CGAC**	0.307	0.148	0.159	0.0118
CCG_rc	0.030	0.024	0.006	0.0351	**R_CUGA**	1.114	1.329	-0.215	0.0437
CCU_rc	0.044	0.054	-0.010	0.0129	**R_CUGC**	1.326	1.601	-0.274	0.0373

CG_rc	0.059	0.046	0.013	4.79E-05	**R_GAGA**	0.898	1.137	-0.239	0.0249
CGG_rc	0.024	0.019	0.005	0.0290	**R_GAUC**	0.149	0.321	-0.172	0.0004
CU_rc	0.095	0.105	-0.010	0.0086	**R_GCAC**	1.220	0.747	0.473	3.36E-06
G_15	0.327	0.285	0.043	0.0419	**R_GCC**	3.488	3.959	-0.471	0.0189
G_4	0.322	0.387	-0.065	0.0030	**R_GCCA**	0.789	1.112	-0.323	0.0038
GA_rc	0.086	0.081	0.005	0.0409	**R_UCGA**	0.128	0.055	0.074	0.0319
GAA_rc	0.041	0.035	0.006	0.0250	**R_UUAA**	0.102	0.272	-0.171	0.0003
GAC_rc	0.034	0.026	0.008	0.0048	**U_15**	0.225	0.266	-0.041	0.0382
GGA_rc	0.041	0.034	0.007	0.0012	**U_18**	0.312	0.268	0.044	0.0343
GGC_rc	0.030	0.041	-0.011	0.0002	**U_21**	0.276	0.209	0.067	0.0006
GUC_rc	0.027	0.037	-0.010	0.0002	**U_5**	0.247	0.199	0.048	0.0115
GUG_rc	0.040	0.034	0.006	0.0437	**U_9**	0.313	0.254	0.059	0.0045
RDK_AACG	0.231	0.129	0.102	0.0116	**U_rc**	0.251	0.269	-0.018	0.0027
RDK_AAGC	0.939	0.592	0.347	0.0007	**UC_rc**	0.082	0.098	-0.017	2.21E-05
RDK_ACC	3.253	2.809	0.444	0.0123	**UCC_rc**	0.042	0.052	-0.010	0.0233
RDK_ACCC	0.944	0.732	0.212	0.0356	**UCG_rc**	0.016	0.011	0.005	0.0422
RDK_ACG	1.201	0.862	0.339	0.0023	**UCU_rc**	0.034	0.050	-0.016	4.65E-05
RDK_ACGG	0.479	0.252	0.227	0.0004					

F) Feature Importance

We examined the top 20 features from the best-performing feature set identified using feature importance in the Extra Trees model. Notably, 17 of these top 20 features were among those with $p < 0.05$ by the Mann–Whitney test, confirming they significantly differentiate the two classes. For example, in the binary profile features, the presence of cytosine at the first position (feature C_1) and uracil at the 21st position (U_21) were among the most important features and showed significant enrichment or depletion in exosomal miRNAs compared to non-exosomal. In the reverse-complement k-mer composition features, certain k-mers were markedly more frequent in exosomal miRNAs. From the TF–IDF features (for k-mers in reverse complements), we observed that nucleotide “U” was more abundant in exosomal sequences. Specific di-nucleotides like "AC" and "UC", and tri-nucleotides "GGA", "GGC", "GUC" appeared more often in exosomal miRNAs than in non-exosomal miRNAs. Figure 5.4

illustrates the 20 most important features for distinguishing exosomal vs non-exosomal miRNAs. A complete list of features and their importance is provided in Table 5.10.

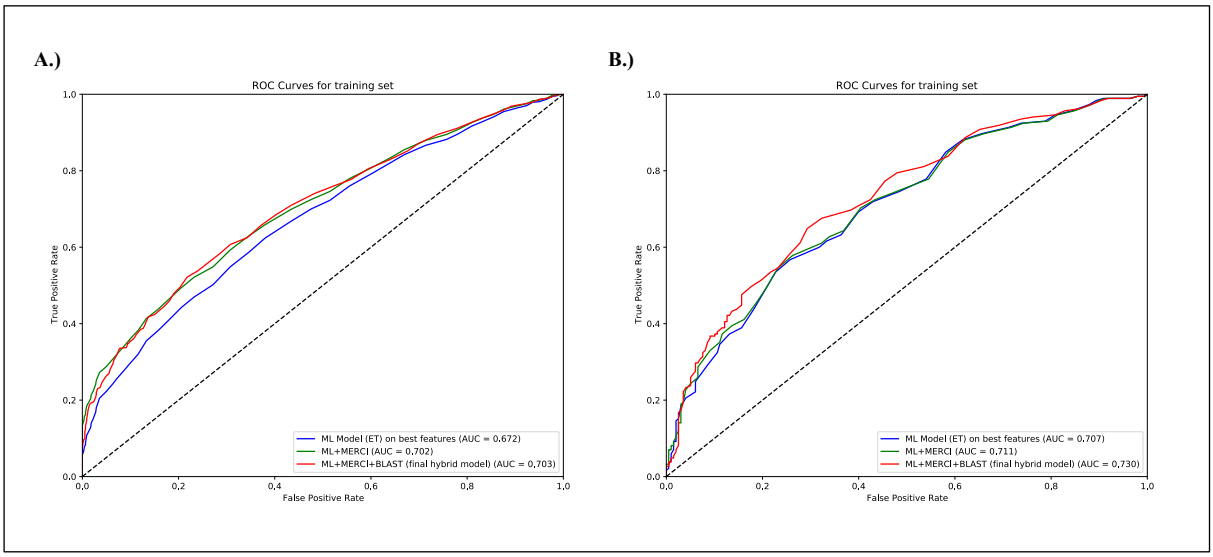


Figure 5.3: The AUROC plots depicting the performance of the ensemble model (AI-based model and alignment-based method) for A) Training set and B) Independent Validation set [https://doi.org/10.1101/2024.06.20.599824]

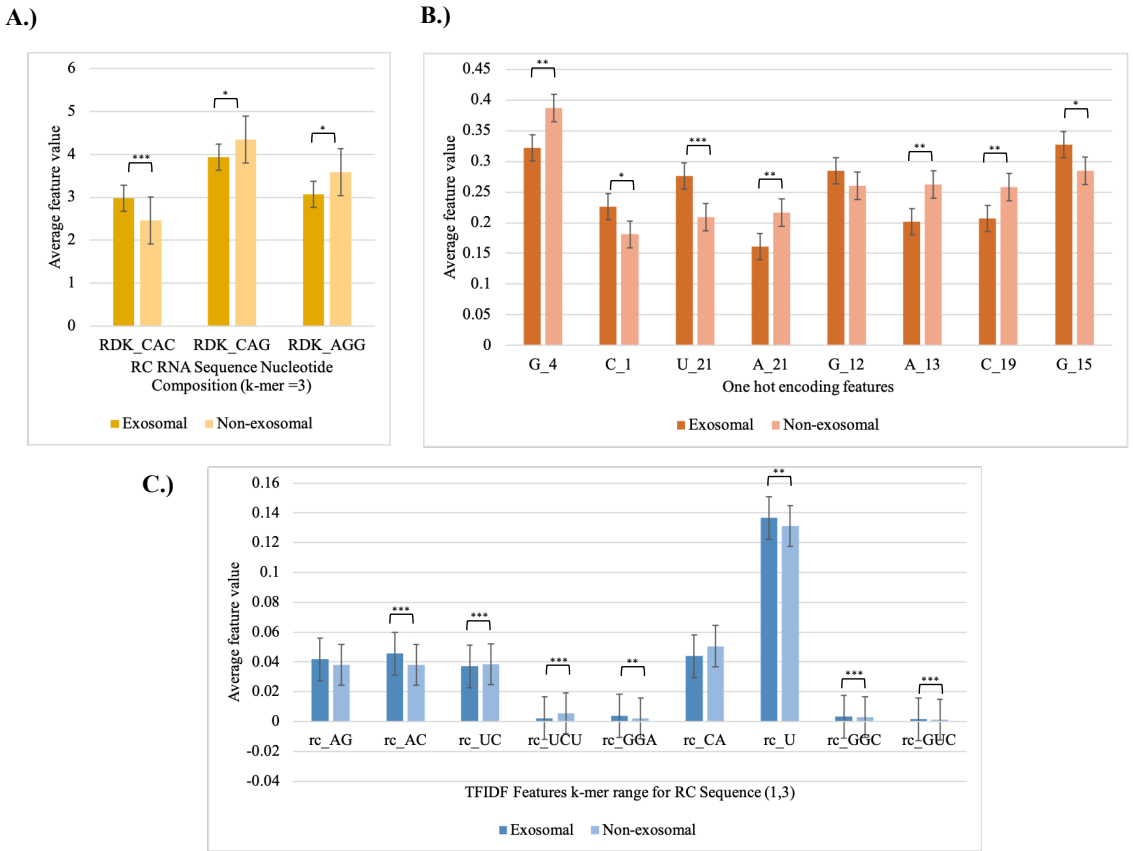


Figure 5.4: The comparison of the top 20 important features for distinguishing exosomal and non-exosomal classes [https://doi.org/10.1101/2024.06.20.599824]

Table 5.10: Feature importance of the selected features using Extra Tree Classifier

Feature	Importance	Feature	Importance	Feature	Importance	Feature	Importance
A_1	0.00399172	**G_11**	0.00394648	**R_AGGA**	0.00261773	**R_GGAC**	0.00236174
A_10	0.00341059	**G_12**	0.00425297	**R_AGGC**	0.00375953	**R_GGCA**	0.00312434
A_11	0.00341207	**G_13**	0.00349591	**R_AGGG**	0.00380984	**R_GGCC**	0.00173627
A_12	0.00388907	**G_14**	0.00340207	**R_AGUA**	0.00192297	**R_GGGA**	0.00297184
A_13	0.00421233	**G_15**	0.00416182	**R_AGUC**	0.00186995	**R_GGUA**	0.00198468
A_14	0.00285332	**G_16**	0.00321676	**R_AGUG**	0.00344138	**R_GUA**	0.00340754
A_15	0.00269059	**G_17**	0.00373743	**R_AUA**	0.00259519	**R_GUAA**	0.00199873
A_16	0.00266655	**G_18**	0.00319047	**R_AUAA**	0.00158491	**R_GUAC**	0.00134345
A_17	0.00310757	**G_19**	0.00375706	**R_AUAC**	0.00154717	**R_GUCA**	0.00218114
A_18	0.00310532	**G_2**	0.00386101	**R_AUAG**	0.00179907	**R_GUGA**	0.00302649
A_19	0.0026608	**G_20**	0.00400638	**R_AUAU**	0.00109669	**R_GUUA**	0.00149183
A_2	0.00371118	**G_21**	0.00359712	**R_AUC**	0.00288653	**R_UAA**	0.00335938
A_20	0.00317313	**G_22**	0.00365574	**R_AUCA**	0.00177004	**R_UAAA**	0.00199777
A_21	0.00427718	**G_23**	0.00185117	**R_AUCC**	0.00258889	**R_UACA**	0.00223948
A_22	0.00330643	**G_24**	0.00090095	**R_AUCG**	0.00061823	**R_UAGA**	0.00215711
A_23	0.00094039	**G_25**	0.00022469	**R_AUG**	0.00315252	**R_UAUA**	0.00133506
A_24	0.00039086	**G_26**	3.79E-05	**R_AUGA**	0.00189484	**R_UCA**	0.00344112
A_25	8.75E-05	**G_3**	0.0031244	**R_AUGC**	0.00186458	**R_UCAA**	0.00226255
A_26	1.96E-05	**G_4**	0.00581187	**R_AUGG**	0.00239213	**R_UCCA**	0.00256843
A_3	0.0035623	**G_5**	0.00334583	**R_AUUA**	0.00153421	**R_UCGA**	0.00089941
A_4	0.00328034	**G_6**	0.00327323	**R_AUUC**	0.00156483	**R_UGAA**	0.00271097
A_5	0.00300501	**G_7**	0.0032546	**R_AUUG**	0.0024475	**R_UGCA**	0.00195621
A_6	0.00313525	**G_8**	0.0030653	**R_CAA**	0.00357447	**R_UUAA**	0.00208828
A_7	0.00372421	**G_9**	0.00328128	**R_CAAA**	0.00230496	**U_1**	0.00364039
A_8	0.00261907	**G_rc**	0.00387447	**R_CAAC**	0.00268186	**U_10**	0.00330965
A_9	0.00261424	**GA_rc**	0.00383716	**R_CAAG**	0.00264839	**U_11**	0.00384095
A_rc	0.00399365	**GAA_rc**	0.00280309	**R_CAC**	0.00446877	**U_12**	0.00321291
AA_rc	0.00354497	**GAC_rc**	0.00303058	**R_CACA**	0.00244914	**U_13**	0.00352525
AAA_rc	0.00279006	**GAG_rc**	0.00266854	**R_CACC**	0.00309941	**U_14**	0.00331285
AAC_rc	0.00321902	**GAU_rc**	0.00271377	**R_CACG**	0.00146756	**U_15**	0.00349181
AAG_rc	0.0028866	**GC_rc**	0.00396389	**R_CAG**	0.00439409	**U_16**	0.00330264
AAU_rc	0.00292631	**GCA_rc**	0.003563	**R_CAGA**	0.00320505	**U_17**	0.00281353
AC_rc	0.00505434	**GCC_rc**	0.00314855	**R_CAGC**	0.00277907	**U_18**	0.00403927

ACA_rc	0.00298039	**GCG_rc**	0.00199742	**R_CAGG**	0.00393949	**U_19**	0.00317805
ACC_rc	0.0034129	**GCU_rc**	0.00333948	**R_CAUA**	0.00192866	**U_2**	0.0030566
ACG_rc	0.00134923	**GG_rc**	0.00362573	**R_CAUC**	0.00220718	**U_20**	0.0028204
ACU_rc	0.00326243	**GGA_rc**	0.00454274	**R_CAUG**	0.00166301	**U_21**	0.00436984
AG_rc	0.00411415	**GGC_rc**	0.00428143	**R_CCA**	0.00407846	**U_22**	0.00359829
AGA_rc	0.00279495	**GGG_rc**	0.00232933	**R_CCAA**	0.00284056	**U_23**	0.00186774
AGC_rc	0.00287649	**GGU_rc**	0.00284617	**R_CCAC**	0.00281145	**U_24**	0.00069782
AGG_rc	0.0036679	**GU_rc**	0.00341759	**R_CCAG**	0.00323279	**U_25**	0.00032045
AGU_rc	0.00300615	**GUA_rc**	0.00257633	**R_CCC**	0.00346897	**U_26**	0
AU_rc	0.00357341	**GUC_rc**	0.00420536	**R_CCCA**	0.00322593	**U_3**	0.00340367
AUA_rc	0.00189938	**GUG_rc**	0.00287514	**R_CCCC**	0.00291418	**U_4**	0.00343859
AUC_rc	0.00230496	**GUU_rc**	0.00247766	**R_CCCG**	0.0024791	**U_5**	0.00344603
AUG_rc	0.00272643	**R_AAA**	0.00348523	**R_CCG**	0.00248863	**U_6**	0.00290544
AUU_rc	0.00249417	**R_AAAA**	0.00248008	**R_CCGA**	0.00151445	**U_7**	0.00307198
C_1	0.00475205	**R_AAAC**	0.00241904	**R_CCGC**	0.00142887	**U_8**	0.00328383
C_10	0.00376357	**R_AAAG**	0.00320846	**R_CCGG**	0.00082377	**U_9**	0.00333871
C_11	0.00306085	**R_AAAU**	0.00202569	**R_CCUA**	0.00200825	**U_rc**	0.00432667
C_12	0.00379948	**R_AAC**	0.00315752	**R_CCUC**	0.00263336	**UA_rc**	0.00317157
C_13	0.00286281	**R_AACA**	0.0026779	**R_CGA**	0.00218839	**UAA_rc**	0.00271946
C_14	0.00326994	**R_AACC**	0.00244048	**R_CGAA**	0.00195695	**UAC_rc**	0.00212171
C_15	0.00339731	**R_AACG**	0.00098237	**R_CGAC**	0.00133128	**UAG_rc**	0.00261222
C_16	0.00338596	**R_AACU**	0.00222453	**R_CGAG**	0.00131553	**UAU_rc**	0.00203265
C_17	0.00285121	**R_AAG**	0.00359444	**R_CGC**	0.00246625	**UC_rc**	0.0050512
C_18	0.00256392	**R_AAGA**	0.00296298	**R_CGCA**	0.00123002	**UCA_rc**	0.00308027
C_19	0.00419302	**R_AAGC**	0.00300301	**R_CGCC**	0.0015068	**UCC_rc**	0.0036032
C_2	0.00298127	**R_AAGG**	0.00316682	**R_CGCG**	0.00088794	**UCG_rc**	0.00174543
C_20	0.00297295	**R_AAGU**	0.00282233	**R_CGGA**	0.0015686	**UCU_rc**	0.00489616
C_21	0.00341623	**R_AAU**	0.00318448	**R_CGGC**	0.00202888	**UG_rc**	0.00358597
C_22	0.00307748	**R_AAUA**	0.00233857	**R_CGUA**	0.00072697	**UGA_rc**	0.00318941
C_23	0.00109235	**R_AAUC**	0.00217709	**R_CGUC**	0.00142901	**UGC_rc**	0.00307345
C_24	0.00045535	**R_AAUG**	0.00280841	**R_CUA**	0.0024018	**UGG_rc**	0.00303544
C_25	0.00018581	**R_AAUU**	0.00171827	**R_CUAA**	0.0019048	**UGU_rc**	0.00285147
C_26	0	**R_ACA**	0.00305537	**R_CUAC**	0.00219007	**UU_rc**	0.00376958
C_3	0.00330278	**R_ACAA**	0.0025685	**R_CUAG**	0.00167469	**UUA_rc**	0.00255778
C_4	0.0037041	**R_ACAC**	0.00226595	**R_CUC**	0.00362383	**UUC_rc**	0.00278129
C_5	0.00338976	**R_ACAG**	0.00321387	**R_CUCA**	0.00321354	**UUG_rc**	0.00191243
C_6	0.00319445	**R_ACAU**	0.00261232	**R_CUCC**	0.00317946	**UUU_rc**	0.00303845
C_7	0.00323447	**R_ACC**	0.00338953	**R_CUGA**	0.00319096	**X_1**	0
C_8	0.00363721	**R_ACCA**	0.00265064	**R_CUGC**	0.00388859	**X_10**	0

C_9	0.00330667	**R_ACCC**	0.00230767	**R_CUUA**	0.00176488	**X_11**	0
C_rc	0.00334598	**R_ACCG**	0.00163016	**R_CUUC**	0.00266856	**X_12**	0
CA_rc	0.00450543	**R_ACCU**	0.00204213	**R_GAA**	0.00309305	**X_13**	0
CAA_rc	0.00325071	**R_ACG**	0.0023245	**R_GAAA**	0.00238446	**X_14**	0
CAC_rc	0.0031366	**R_ACGA**	0.00105016	**R_GAAC**	0.00244088	**X_15**	0
CAG_rc	0.00379573	**R_ACGC**	0.00102507	**R_GAC**	0.00362093	**X_16**	0
CAU_rc	0.00293438	**R_ACGG**	0.00253455	**R_GACA**	0.00241652	**X_17**	5.43E-05
CC_rc	0.00354723	**R_ACGU**	0.00078279	**R_GACC**	0.00169044	**X_18**	0.00064101
CCA_rc	0.00304126	**R_ACU**	0.00369146	**R_GAGA**	0.00257275	**X_19**	0.00081392
CCC_rc	0.00404099	**R_ACUA**	0.00200576	**R_GAGC**	0.0028238	**X_2**	0
CCG_rc	0.00247308	**R_ACUC**	0.00240066	**R_GAUA**	0.00154106	**X_20**	0.00145605
CCU_rc	0.00373869	**R_ACUG**	0.00326929	**R_GAUC**	0.00269653	**X_21**	0.0026546
CG_rc	0.00369689	**R_AGA**	0.0038502	**R_GCA**	0.0040308	**X_22**	0.00320974
CGA_rc	0.00169924	**R_AGAA**	0.00236534	**R_GCAA**	0.00249707	**X_23**	0.00247615
CGC_rc	0.00188677	**R_AGAC**	0.00262002	**R_GCAC**	0.00352306	**X_24**	0.0015732
CGG	0.00171842	**R_AGAG**	0.00281741	**R_GCC**	0.00361277	**X_25**	0.00065952
CGU	0.0016972	**R_AGAU**	0.0028266	**R_GCCA**	0.0040212	**X_26**	0.00012069
CU_rc	0.00395602	**R_AGC**	0.00339492	**R_GCCC**	0.00275559	**X_3**	0
CUA_rc	0.0025845	**R_AGCA**	0.00323614	**R_GCGA**	0.00111164	**X_4**	0
CUC_rc	0.00395187	**R_AGCC**	0.00292959	**R_GCGC**	0.00098499	**X_5**	0
CUG_rc	0.00357377	**R_AGCG**	0.00165219	**R_GCUA**	0.00157596	**X_6**	0
CUU_rc	0.0029909	**R_AGCU**	0.00197952	**R_GGA**	0.00315631	**X_7**	0
G_1	0.00264671	**R_AGG**	0.00415888	**R_GGAA**	0.00248987	**X_8**	0
G_10	0.00347195						

5.3.2.2 DL Models

A) Sequences

Our CNN model trained on one-hot encoded sequences achieved an AUROC of 0.611 on the training set (five-fold cross-validated) and 0.621 on the independent validation set. The complete performance metrics for the CNN on sequence data are provided in Table 5.11.

B) Structure Images

Using RNA secondary structure images as input, we trained both a CNN and a ResNet50 model. The CNN and ResNet50 achieved AUROCs of 0.551 and 0.553 on the validation set, respectively. These results (also detailed in Table 5.11) show that classifying miRNAs based

on their secondary structure images is challenging, and these models did not perform as well as sequence-based approaches.

Table 5.11: Results for the CNN model developed on sequences and images, and the ResNet model developed on images

	Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
CNN Seq	0.57	80.545	80.607	80.576	0.612	0.199	56.757	58.081	57.441	0.621	0.148
CNN Images	0.48	48.119	50.792	49.444	0.503	-0.011	50.270	54.545	52.480	0.551	0.048
ResNet Images	0.48	50.584	50.660	50.621	0.504	0.012	50.811	52.020	51.436	0.554	0.028

C) LLM Models

We fine-tuned the DNABERT and BERT-base-uncased models to classify miRNAs. The DNABERT-based classifier achieved an AUROC of 0.535 on the validation set, whereas the BERT-base model achieved 0.608. The relatively low AUROC for DNABERT suggests that its specialized pre-training on genomic DNA did not directly translate into identifying exosomal miRNAs in our case. The BERT-base model performed better, possibly due to learning contextual cues during fine-tuning. The detailed results for the fine-tuned LLM model performance on the independent validation set are available in Table 5.12.

Table 5.12: Results for the fine-tuned LLM models (DNABERT and BERT-base uncased) on the independent validation set

Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DNABERT Finetuned	0.50	52.350	55.080	51.832	0.535	0.095
BERT-base-uncased Finetuned	0.47	64.865	50.000	57.180	0.608	0.150

5.3.3 Hybrid classification method

To maximize prediction performance, we developed a hybrid model that combines our best-performing ML model with the alignment-based motif and similarity-search (BLAST) results. The best-performing individual model was the Extra Trees classifier (AUC 0.707 on validation). When we incorporated the motif-search results into this model, the performance

increased, leading to a validation AUC of 0.712. We then added the BLAST similarity-search component as described in Methods, and the AUC further improved to 0.730 on the validation set. The Matthews correlation coefficient (MCC) of the hybrid approach on the validation set was 0.352, higher than any model's MCC. Table 5.8 compares the results of the best-performing combined features model alone versus the hybrid model that integrates motif and similarity-search approaches. The hybrid model improved all key metrics on both training and validation sets. The AUROC curves for the hybrid model on the training and validation sets are shown in Figure 5.3. The hybrid model outperformed the single-model approaches, demonstrating the value of incorporating biologically informed alignment-based evidence into the data-driven model.

5.3.4 Comparison with existing methods

Presently, no tool exclusively predicts exosomal miRNA. However, a few computational tools predict miRNA subcellular localization, with "exosome" being one of several possible subcellular locations, like miRNALoc and EL-RMLocNet (Asim et al., 2022; Meher, Satpathy, et al., 2021). We compared the performance of our method with these existing tools. According to their published results, miRNALoc achieves about 50% accuracy in predicting the subcellular localization of miRNAs on independent test data. EL-RMLocNet reports an AUROC of 0.629 for human miRNA localization on a benchmark dataset. We submitted our independent validation set (383 sequences) to the miRNALoc web server. It achieved an AUROC of 0.494, accuracy of 48.04%, and MCC of –0.028 on our data, significantly lower than our method (which yields AUC 0.73, accuracy 67.62%, MCC 0.352 on the same set). We could not comprehensively test EL-RMLocNet on all validation sequences because it only allows one sequence input at a time, making it impractical to get results for 383 sequences. These results indicate that our method, EmiRPred, performs substantially better in identifying exosomal miRNAs than existing general miRNA localization predictors.

5.3.5 Web Server and Standalone Software

This work aims to provide an accessible tool for the scientific community, particularly those working in diagnostics and prognostics involving exosomal miRNA biomarkers. To this end, we developed a user-friendly web server with four main modules: **Predict**, **Design**, **Motif-scan**, and **BLAST-search**.

- The **"Predict"** module accepts miRNA sequences as input and predicts whether each sequence is exosomal or non-exosomal using our best model.
- The **"Design"** module allows users to generate all possible single-nucleotide mutants of a query miRNA and then predicts which mutant sequences would be classified as exosomal. This can help in understanding the sequence determinants of exosomal localization.
- The **"Motif-scan"** module scans an input sequence for the presence of known exosomal motifs (as identified in Table 5.1) and reports any motifs found.
- The **"BLAST-search"** module performs a BLAST similarity search against our database of known exosomal and non-exosomal miRNAs, allowing users to see if their sequence has a close match in either category.

The web server designed for EmiRPred is compatible with smartphones, tablets, Macs and PCs. Along with webserver we have also developed a Python software package, a standalone command-line tool, and a GitHub repository. They are available at the following links: https://webs.iiitd.edu.in/raghava/emirpred/ and https://github.com/raghavagps/emirpred.

5.4 Discussions

Exosomes are lipid bilayer-enclosed vesicles released by most cell types in the body. They carry diverse molecular cargoes, including lipids, proteins, metabolites, and RNA species (Kalluri & LeBleu, 2020). Exosomes have emerged as a reliable and non-invasive reservoir of biomarkers for diagnosing different diseases, including cancer, neurodegenerative conditions, and cardiovascular disorders. Recently, a growing number of biomarkers derived from exosomes have been discovered. For instance, exosomal PD-L1 has been recognized as a potential biomarker for lung cancer (G. Chen et al., 2018). In addition, exosome-derived circular RNAs (circRNAs) have been associated with colorectal cancer, demonstrating promise for diagnosis and therapeutic intervention. (Vakhshiteh et al., 2021). In breast cancer, exosomal miR-21 levels are frequently elevated, and exosomal miR-1246 has been associated with chemotherapy resistance in pancreatic cancer (Ghafouri-Fard et al., 2022; Hashemi et al., 2023). In therapy, exosomes have been engineered to deliver CRISPR/Cas9 components for gene editing, demonstrating efficacy in correcting genetic mutations in preclinical models (Lu et al., 2023). Furthermore, exosomes engineered to carry small interfering RNAs (siRNAs) against oncogenes such as KRAS have demonstrated notable tumor-suppressive effects in pancreatic cancer models (Kamerkar et al., 2017). These developments highlight the

multifaceted role of exosomes in diagnostics and therapy, supporting their potential in advancing personalized and targeted treatment strategies.

However, predicting exosomal miRNA computationally poses multiple challenges. First, the short length of miRNAs leads to uninformative features, sparsity, and high sequence homology, complicating accurate classification. Second, experimentally validated datasets are limited in size and exhibit class imbalance, which hinders model training and generalization. Third, the mechanisms controlling miRNA packaging into exosomes are not fully understood, preventing the identification of consistent biological patterns computationally. Lastly, the lack of known exosomal sequence signatures requires advanced techniques to reliably identify subtle patterns and distinguishing features.

In this study, we focused on miRNAs - the most abundant RNA molecules in exosomes, which hold great potential as prognostic and diagnostic biomarkers. Our objective was to predict exosomal miRNAs using various techniques, including sequence similarity search, motif identification, and AI-based approaches. We identified several features and motifs that classify exosomal and non-exosomal miRNAs. Initially, we discovered short sequence motifs frequently found in exosomal miRNAs. The most recurrent motif identified was "GgapAAGCAC," present in 1 non-exosomal and 4 exosomal sequences in the validation set. However, only a small fraction (~5.2%) of miRNAs in the validation set contained any of these motifs, indicating that motif presence alone is insufficient for the comprehensive prediction of exosomal miRNAs. Next, we applied a similarity search technique using BLAST. As described earlier, the similarity-based approach had high precision for the miRNAs it could identify, but had limited coverage. This indicates that alignment-based techniques such as motif and BLAST search cannot capture all exosomal miRNAs alone.

Over the past few years, AI-based models have been commonly used for classification problems, particularly sequence classification. Therefore, we applied machine learning and deep learning methods to develop prediction models for exosomal miRNAs. One advantage of AI-based models over purely alignment-based models is their ability to cover the entire dataset – they predict every input sequence, regardless of whether known motifs or similar sequences are present. We systematically applied ML and DL techniques using all the feature types we could derive. We also implemented large language models specialized for sequences. Despite these extensive efforts, our best AI-based model achieved a maximum AUROC of 0.707 with an MCC of 0.269 on the validation set. To further improve the performance of individual

approaches used in the study, we developed a hybrid method that combined AI-based and alignment-based approaches. We combined these approaches using a scoring system described in the methods section. Our ensemble method - EmiRPred achieved an AUROC of 0.73 and MCC of 0.352 on the independent validation set.

It is crucial to compare new methods with the existing methods. To our awareness, there are currently no dedicated approaches designed exclusively for predicting exosomal miRNA sequences. However, there are some general miRNA localization predictors like EL-RMLocNet and miRNALoc that can predict multiple subcellular locations for miRNAs, including exosomal localization. We evaluated these tools on the independent validation dataset. Our method outperformed the existing tools in exosomal miRNA prediction. The limitation of our study is that, even though the data are derived from experimentally validated sources, our work is entirely computational. The important features and motifs identified in our study have been previously reported to be associated with exosomal miRNAs which supports our findings. However, experimental validation is required to confirm the biological significance of the identified motifs and features. In particular, lab experiments to test whether altering a specific motif in a miRNA can change its packaging into exosomes would strengthen the conclusions.

Chapter 6

Prediction of Abundant miRNAs in Normal Exosomes

6.1 Introduction

Exosomes are extracellular vesicles ranging from 30–150 nm in size and are released by almost all cell types. They carry lipids, nucleic acids, and proteins that reflect the cell's physiological state of origin. Exosomes have attracted considerable attention for their involvement in intercellular communication and their promise as diagnostic biomarkers. In particular, the miRNA content of exosomes can mirror the pathological or physiological conditions of the originating cells. For instance, distinct profiles of exosomal miRNAs have been observed in cancer, where certain exosomal miRNAs (e.g., miR-21, miR-155) contribute to tumor progression and metastasis, and their levels are elevated in patient exosomes (X. Lai et al., 2017; Melo et al., 2014). Similarly, in neurodegenerative disorders like Alzheimer's disease, specific exosomal miRNAs (e.g., miR-125b, miR-146a) are implicated in neuronal dysfunction and neuroinflammation (L. Cheng et al., 2015). The stability of miRNAs within exosomes (protected from RNases in bodily fluids) makes them attractive candidates for non-invasive diagnostics (W. Yu et al., 2021; B. Zhou et al., 2020).

Identifying which miRNAs are highly abundant in exosomes under healthy conditions is crucial for determining a reference baseline. Such a baseline helps distinguish disease-specific alterations from normal variability in exosomal miRNA content. miRNAs that are consistently abundant in exosomes of healthy individuals likely play critical physiological roles and could exhibit dysregulation in disease states. Therefore, characterizing these abundant exosomal miRNAs can improve our understanding of normal intercellular communication and facilitate more precise detection of disease-related changes in exosomal miRNA profiles.

Previous approaches to identifying exosomal miRNA biomarkers often focused on finding miRNAs with differential abundance between patients and controls. However, it is challenging to interpret the significance of observed changes without baseline knowledge of usually abundant exosomal miRNAs. For example, knowing which miRNAs are present at high levels in exosomes in healthy individuals helps researchers recognize when those miRNAs deviate in disease and by how much. The establishment of this baseline can help discover new biomarkers for a number of diseases and enable more precise non-invasive disease monitoring (Bartel, 2018; Weber et al., 2010).

In this work, we introduce AdmirePred, which is a computational tool specifically designed to predict highly abundant miRNAs in human blood exosomes under normal conditions.

AdmirePred integrates two major types of techniques: alignment based approaches and AI based (machine learning) approaches. The alignment-based component includes motif discovery and sequence similarity search methods, which identify sequence patterns or homology that might relate to preferential miRNA packaging into exosomes. The AI-based component involves training machine learning models on a comprehensive set of sequence-derived features to distinguish exosome-abundant miRNAs from others. We trained and validated our models on experimentally validated datasets of abundant and non-abundant miRNA in exosomes compiled from the EVmiRNA database and NCBI GEO (Barrett et al., 2007; T. Liu et al., 2019). The overall architecture of our approach is illustrated in Figure 6.1, which outlines data collection, feature extraction, model development, and evaluation.

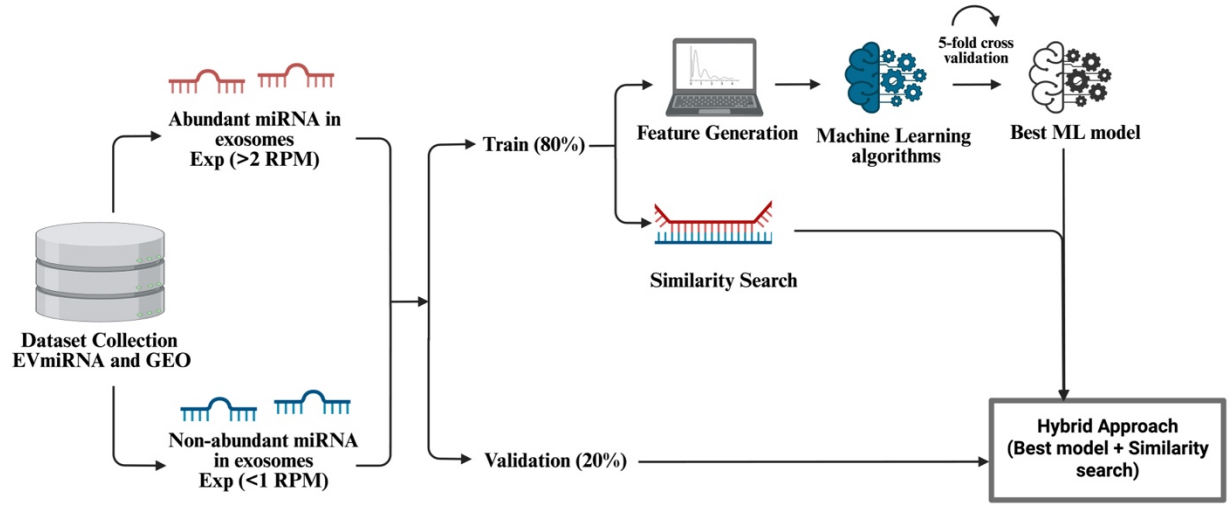


Figure 6.1: A schematic overview of the AdmirePred workflow

6.2 Methods

6.2.1 Data Collection and Preprocessing

We collected expression data for miRNAs from the EVmiRNA database for miRNA in blood exosomes (serum and plasma) of healthy human samples (n = 60). MiRNAs with an average expression value greater than 2 (in reads per kilobase million, RPKM) were classified as abundant in exosomes. In contrast, miRNAs with average expression less than 1 RPKM were classified as non-abundant. Here, we refer to abundant and non-abundant miRNAs in exosomes as exosomal and non-exosomal miRNA respectively. To validate our data and ensure robustness, we cross-verified the identified exosomal and non-exosomal miRNAs using a GEO study (accession GSE270497) that performed small RNA sequencing of plasma exosomes in

60 healthy and 120 breast cancer patients. We verified our data with the miRNA data available for the healthy individuals in this study. After verifying and removing duplicates, our curated dataset comprised 348 exosomal (abundant) miRNA sequences and 349 non-exosomal (non-abundant) miRNA sequences. The sequence lengths ranged from 16 to 25 nucleotides.

6.2.2 Motif-Search

We employed an alignment-based motif discovery approach to investigate whether certain sequence motifs are enriched in exosome-abundant miRNAs. Specifically, we used the MERCI (Motif EmeRging and with Classes Identification) tool to identify nucleotide sequence motifs that occur predominantly in exosomal miRNAs. MERCI is designed to find motifs that are exclusive to a given class of sequences (Vens et al., 2011). In our case, we wanted to identify motifs exclusively present in exosomal miRNAs but not in non-exosomal ones. We ran MERCI on the training set of miRNA sequences using our exosomal miRNAs as positive and non-exosomal miRNAs as negative classes. The tool was configured to search for motifs with and without gaps, allowing a gap length of up to 1 nucleotide.

6.2.3 Similarity Search using BLAST

In addition to motif discovery, we applied a sequence similarity search approach to identify exosomal miRNAs based on their sequence similarity to known examples. We utilized NCBI's blastn-short program optimized for short sequences to compare each query miRNA sequence against a database of our training sequences labeled as exosomal or non-exosomal. The BLAST search was conducted using a range of expectation values (e-values) from 10^{-6} to 10^{6} to tune the stringency of the matches (Boratyn et al., 2013). In the training set, self-hits were excluded, and the top hit following self-hits was considered. For the validation set, the first hit was directly used to assess performance across different e-value thresholds.

6.2.4 Machine Learning-Based Classification

To differentiate between miRNAs with high abundance in exosomes and those with low abundance, we employed a range of techniques to derive informative features from miRNA sequences. Below is an overview of these feature-generation strategies:

6.2.4.1 Feature Generation

I. **Nucleotide Composition**: We used the Nfeature tool to gather composition-based descriptors, which included the composition of nucleotides (A, C, G, U) in both the original sequence and its reverse complement for k-mers of lengths 1, 2, 3, and 4 (Mathur et al., 2021).

II. **Nucleotide-Weighted Frequency**: To evaluate how often particular k-mers appear across our sequences, we applied the TF-IDF algorithm as described in Chapter 5. We computed TF-IDF values for both the original sequences and their reverse complements, covering k-mer sizes from (1,1) up to (1,4).

III. **Distance Distribution of Nucleotides**: We also determined the distribution of distances between specific nucleotides in the miRNA sequences using Nfeature. While straightforward frequency-based measures capture the content of nucleotides, they may overlook spatial arrangement. Distance-based features capture positional relationships, which can be pivotal for miRNA structure, functionality, and regulation.

IV. **Nucleotide Repeat Index (NRI)**: The NRI quantifies how frequently each nucleotide (A, C, G, U) appears in consecutive runs within a sequence, translating these repeats into a concise numeric form. This highlights whether nucleotides tend to cluster or be scattered, a factor that can affect RNA folding properties and gene regulation.

V. **Entropy**: We calculated Shannon entropy of the sequences in this study, which reflects the degree of randomness in nucleotide composition. Sequences heavily dominated by one or two nucleotides present lower entropy values, whereas those more evenly balanced show higher entropy. We computed an overall (sequence-level) entropy and a four-value vector capturing the nucleotide-level entropy, revealing how each of A, C, G, and U contributes to overall sequence diversity.

VI. **Correlation**: We incorporated descriptors that measure positional and physicochemical relationships along the miRNA sequence, including:

   a) **Autocorrelation**: Correlation of a single property at different positions separated by a fixed lag.
   b) **Cross-Correlation**: Measures how two different properties correlate at positions that are separated by a lag.
   c) **Auto-Cross Correlation**: Combines the concepts of autocorrelation and cross-correlation into a unified descriptor, effectively capturing both same-property and different-property correlations over defined positional gaps.

VII. **Pseudo Composition**: Unlike standard composition features, pseudo-nucleotide composition converts sequences of varying lengths into uniform-length vectors while preserving long-range ordering. We used two types of pseudo-composition in this study:

   a) **Serial Correlation Pseudo Composition**: Captures how consecutive dinucleotides depend on each other along a sequence.
   b) **Parallel Correlation Pseudo Composition**: Assesses several correlations across different positions in the sequence simultaneously.

VIII. **Binary Profiles (One-Hot Encoding)**: Finally, we applied one-hot encoding to turn each nucleotide into a binary vector. Since miRNAs vary in length, we standardized sequence length to 25 nucleotides (the longest in the dataset) by padding shorter sequences with the placeholder 'X.' For example, a 22-nucleotide sequence like "AGTAUAAGTCGUGCATUCCAAU" was extended to "AGTAUAAGTCGUGCATUCCAAUXXX." After padding, each position was encoded into a binary representation suitable for downstream classification models.

6.2.4.2 Prediction Models

In this study, we applied several ML models to distinguish between abundant and non-abundant miRNAs in exosomes. Decision Tree (DT) and Random Forest (RF) rely on splitting the data into decision rules and combining multiple trees, respectively (Breiman, 2001; Bulac & Bulac, 2016). Logistic Regression uses a logistic function to estimate class probabilities, while K-Nearest Neighbors bases classification on the most common label among the k nearest neighbors (Stoltzfus, 2011; Y. Wu et al., 2002). Gaussian Naïve Bayes applies Bayes' theorem under a Gaussian distribution assumption, and Support Vector Classifier determines an optimal hyperplane to separate classes (Cristianini & Ricci, 2008; Joshi et al., 2012). Extreme Gradient Boosting refines predictions through iterative gradient-boosted trees, and Extra Trees introduces additional randomness in feature selection and splitting to enhance model robustness (T. Chen & Guestrin, 2016; Geurts et al., 2006). These models were trained on various feature sets described above to classify miRNA sequences according to their abundance in exosomes.

6.2.4.3 Cross-validation and performance metrics

The dataset consists of a total of 697 miRNA sequences that were divided into a 80:20 ratio, allocating 80% for model training and preserving 20% for validation. We applied five-fold cross-validation on the training data as described in Chapter 4. We also computed the different threshold-dependent and independent metrics for each model as described previously in Chapter 4.

6.2.5 Ensemble Method

To further enhance the performance of our best model, which was built on the optimal feature set, we introduced a strategy combining both AI-based and alignment-based methods. This approach uses a weighted scoring scheme that integrates (i) a motif-based approach, (ii) a similarity search using BLAST, and (iii) ML predictions. Each miRNA sequence receives a +0.5 score if an exosomal motif is identified and a 0 score if no motif is present, and these points are added to the probability estimates generated by the top-performing ML model, generated using the predict_proba() function from scikit-learn (Pedregosa et al., 2011). Similarly, sequences obtain a +0.5 score if they are found to be similar to the miRNA that is highly abundant in exosomes and these points are again added to the model's predicted probabilities. The combined score from ML and motif-based detection, as well as the combined score from ML and BLAST analysis, range from 0 to 1.5, as shown in Equations 8 and 9. Based on these total scores, sequences are labelled as either exosomal or non-exosomal. This hybrid or ensemble methodology has also been reported in prior studies (Arora, Patiyal, et al., 2023; Kaur et al., 2024).

$$S_b = \begin{cases} S + 0.5 & If\ hit\ is\ against\ an\ exosomal\ sequence \\ S & If\ there\ is\ no\ hit\ or\ a\ hit\ against\ non-exosomal\ sequence \end{cases} \quad \text{- Eq. 8}$$

Here, S_b is the probability score obtained from the total score calculated by adding prediction probability scores from the ML model with the best performance and BLAST-based method; the score ranges from 0 to 1.5

$$S_m = \begin{cases} S + 0.5 & If\ an\ exosomal\ motif\ is\ present \\ S & if\ no\ exosomal\ motif\ present \end{cases} \quad \text{- Eq. 9}$$

Here, S_m is the score calculated by adding the score from the motif-based approach to prediction probability scores from the best performing ML model. It ranges from 0 to 1.5.

6.3 Results

We explored a variety of methods to predict miRNAs that are abundant in exosomes, organized into three classes: (i) Alignment-based approaches, (ii) AI-based techniques, and (iii) Ensemble approaches. Alignment-based approaches encompass motif identification through MERCI and sequence similarity searches using BLAST. AI-based approaches involve the application of machine learning models to a broad range of features. To leverage the strengths of alignment-based and AI-based methods, we devised an ensemble approach that effectively merges these two approaches.

6.3.1 Motif-Search

When we used MERCI to identify motifs, the independent validation set showed low coverage coupled with a high error rate. At a gap value of 0, only 8 sequences were detected, with 3 incorrect classifications. At gap = 1, the tool captured about 15 sequences, among which 6 were misclassified.

6.3.2. Similarity-Search Using BLAST

We used blastn-short tool to compare sequences against a training dataset containing abundant and non-abundant miRNAs found in exosomes, testing e-values ranging from 10^{-6} to 10^{6} (Altschul et al., 1990). We got an ideal performance at an e-value of 10^{-2}, yielding 24 correct predictions and 0 errors out of 69 exosomal miRNA sequences in the validation set. At e-values below 10^{-2}, coverage was insufficient, whereas higher e-values led to more errors. A detailed summary of the BLAST results for e-values between 10^{-6} and 10^{6} is given in Table 6.1.

Table 6.1: The results for similarity-search using BLAST for different e-values ranging from 10^{-6} to 10^{6} [https://doi.org/10.1101/2025.03.19.644072]

	Training set				**Validation set**			
e-value	**Correct hits**	**Incorrect hits**	**No hits**	**Total**	**Correct hits**	**Incorrect hits**	**No hits**	**Total**
10^{6}	196	83	0	279	32	37	0	69
10^{5}	196	83	0	279	32	37	0	69
10^{4}	196	83	0	279	32	37	0	69

10^{3}	196	83	0	279	32	37	0	69
10^{2}	196	83	0	279	32	37	0	69
10^{1}	196	83	0	279	32	37	0	69
10^{0} (1)	167	57	55	279	43	11	15	69
10^{-1}	118	31	130	279	34	6	29	69
10^{-2}	85	6	188	279	24	0	45	69
10^{-3}	61	1	217	279	14	0	55	69
10^{-4}	51	1	227	279	13	0	56	69
10^{-5}	35	1	243	279	10	0	59	69
10^{-6}	14	1	264	279	2	0	67	69

6.3.3 Machine Learning-Based Classification

We computed a large set of features for each miRNA sequence, capturing various sequence composition and order aspects. In total, 145 features were generated for each miRNA. We then trained a variety of machine learning algorithms on optimal feature set—including LR, GNB, KNN, DT, SVC, XGB, RF, and ET to create predictive models. The AUROC curves for the best-performing ML model are given in Figure 6.3. We have discussed the performance of these ML-based models below:

I. **Nucleotide composition:** We calculated the frequencies of nucleotides and nucleotide combinations (k-mers) for k = 1 to 4. The reverse-complement 3-mer composition yielded the best performance, achieving an independent validation AUC of 0.738 on a support vector classifier. The results for ML models developed on nucleotide composition features are shown in Table 6.2.

Table 6.2: Results for ML models developed on nucleotide composition features extracted from miRNA sequences and their reverse complementary sequences

CDK-1 = (composition of nucleotides, k-mer = 1)											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	57.706	60.573	59.140	0.605	0.183	43.478	61.429	52.518	0.612	0.050
RF	0.52	57.706	58.065	57.885	0.606	0.158	59.420	57.143	58.273	0.622	0.166
LR	0.50	54.122	53.763	53.943	0.569	0.079	53.623	55.714	54.676	0.565	0.093
XGB	0.51	58.423	58.781	58.602	0.619	0.172	59.420	61.429	60.432	0.611	0.209
KNN	0.49	62.724	55.914	59.319	0.624	0.187	63.768	60.000	61.871	0.608	0.238
GNB	0.50	54.480	54.480	54.480	0.570	0.090	53.623	54.286	53.957	0.565	0.079
SVC	0.50	63.799	58.781	61.290	0.641	0.226	57.971	58.571	58.273	0.634	0.165
CDK-2 = (composition of nucleotides, k-mer = 2)											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.46	56.631	55.914	56.272	0.593	0.125	53.623	54.286	53.957	0.565	0.079
RF	0.51	66.667	64.875	65.771	0.73	0.315	68.116	58.571	63.309	0.724	0.268
LR	0.50	58.781	61.29	60.036	0.65	0.201	69.565	58.571	64.029	0.659	0.283
XGB	0.53	65.233	66.308	65.771	0.697	0.315	65.217	54.286	59.712	0.667	0.196
KNN	0.48	62.724	65.233	63.978	0.694	0.280	69.565	61.429	65.468	0.713	0.311
GNB	0.45	64.158	64.158	64.158	0.66	0.283	60.870	51.429	56.115	0.639	0.124
SVC	0.50	68.459	68.817	68.638	0.709	0.373	75.362	60.000	67.626	0.713	0.358
CDK-3 = (composition of nucleotides, k-mer = 3)											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	57.348	60.932	59.14	0.591	0.183	56.522	48.571	52.518	0.525	0.051
RF	0.52	65.591	65.233	65.412	0.708	0.308	68.116	60.000	64.029	0.677	0.282
LR	0.50	63.082	61.29	62.186	0.67	0.244	69.565	60.000	64.748	0.672	0.297
XGB	0.56	61.29	60.573	60.932	0.671	0.219	59.42	58.571	58.993	0.646	0.18
KNN	0.60	59.498	66.308	62.903	0.681	0.259	66.667	58.571	62.59	0.655	0.253
GNB	0.23	63.441	63.441	63.441	0.678	0.269	71.014	50.000	60.432	0.677	0.215
SVC	0.52	63.441	64.875	64.158	0.701	0.283	68.116	58.571	63.309	0.688	0.268
CDK-4 = (composition of nucleotides, k-mer = 4)											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	54.48	59.498	56.989	0.57	0.14	63.768	51.429	57.554	0.576	0.153
RF	0.51	64.516	62.724	63.62	0.7	0.272	66.667	64.286	65.468	0.704	0.31
LR	0.49	59.498	59.857	59.677	0.663	0.194	60.87	55.714	58.273	0.629	0.166

XGB	0.57	63.082	63.082	63.082	0.668	0.262	63.768	65.714	64.748	0.667	0.295
KNN	0.49	60.215	59.857	60.036	0.655	0.201	60.87	64.286	62.59	0.652	0.252
GNB	0.51	42.294	77.061	59.677	0.649	0.206	46.377	68.571	57.554	0.653	0.153
SVC	0.50	63.799	61.29	62.545	0.692	0.251	66.667	51.429	58.993	0.66	0.183
RDK-1 = (composition of nucleotides in reverse complementary sequence, k-mer = 1)											
		Training Set					**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.54	61.29	58.781	60.036	0.617	0.201	79.71	38.571	58.993	0.591	0.200
RF	0.55	58.423	59.140	58.781	0.610	0.176	62.319	54.286	58.273	0.595	0.167
LR	0.50	57.348	58.781	58.065	0.607	0.161	57.971	57.143	57.554	0.619	0.151
XGB	0.52	63.441	62.366	62.903	0.633	0.258	60.870	55.714	58.273	0.621	0.166
KNN	0.55	66.308	46.595	56.452	0.609	0.132	63.768	51.429	57.554	0.606	0.153
GNB	0.55	59.498	56.989	58.244	0.601	0.165	57.971	58.571	58.273	0.631	0.165
SVC	0.55	61.29	59.498	60.394	0.603	0.208	59.42	55.714	57.554	0.594	0.151
RDK-2 = (composition of nucleotides in reverse complementary sequence, k-mer = 2)											
		Training Set					**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.56	56.631	57.706	57.168	0.596	0.143	47.826	67.143	57.554	0.583	0.153
RF	0.52	63.799	65.233	64.516	0.683	0.290	65.217	65.714	65.468	0.698	0.309
LR	0.50	58.781	62.724	60.753	0.632	0.215	65.217	57.143	61.151	0.669	0.224
XGB	0.52	62.724	62.366	62.545	0.670	0.251	66.667	64.286	65.468	0.676	0.310
KNN	0.51	36.918	76.703	56.810	0.629	0.148	39.130	81.429	60.432	0.629	0.227
GNB	0.50	68.459	50.538	59.498	0.641	0.193	76.812	48.571	62.590	0.674	0.264
SVC	0.53	60.573	58.423	59.498	0.661	0.190	72.464	54.286	63.309	0.658	0.272
RDK-3 = (composition of nucleotides in reverse complementary sequence, k-mer = 3)											
		Training Set					**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	60.215	55.914	58.065	0.581	0.161	52.174	67.143	59.712	0.597	0.195
RF	0.41	69.176	54.122	61.649	0.672	0.236	71.014	57.143	64.029	0.665	0.284
LR	0.50	59.857	62.366	61.111	0.666	0.222	76.812	57.143	66.906	0.700	0.346
XGB	0.54	63.799	63.441	63.620	0.691	0.272	65.217	60.000	62.590	0.684	0.252
KNN	0.52	61.290	67.742	64.516	0.709	0.291	73.913	60.000	66.906	0.710	0.342
GNB	0.46	65.950	65.233	65.591	0.692	0.312	76.812	57.143	66.906	0.707	0.346
SVC	0.52	64.875	64.875	64.875	0.730	0.297	76.812	55.714	66.187	0.738	0.333
RDK-4 = (composition of nucleotides in reverse complementary sequence, k-mer = 4)											
		Training Set					**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.52	60.215	57.706	58.961	0.590	0.179	68.116	50.000	58.993	0.591	0.184
RF	0.51	68.100	64.158	66.129	0.715	0.323	69.565	60.000	64.748	0.709	0.297

LR	0.51	63.441	63.441	63.441	0.679	0.269	71.014	65.714	68.345	0.682	0.368
XGB	0.51	61.290	60.573	60.932	0.644	0.219	68.116	54.286	61.151	0.645	0.226
KNN	0.49	69.534	64.875	67.204	0.716	0.344	71.014	51.429	61.151	0.693	0.229
GNB	0.32	63.799	63.441	63.620	0.676	0.272	78.261	65.714	71.942	0.707	0.443
SVC	0.52	64.516	63.082	63.799	0.700	0.276	68.116	62.857	65.468	0.712	0.310

II. **Nucleotide-weighted frequency:** We computed TFIDF for miRNA sequences and their reverse complement for k-mers 1 to 4. TF–IDF features on reverse complements (for k = 1,2) were performing the best overall. An ET classifier trained on these features achieved the highest AUC of 0.744 on the independent validation set. The results for TFIDF features are shown in Table 6.3.

Table 6.3: Results for ML models developed on TFIDF features from k-mers 1 to 4, extracted from miRNA sequences and their reverse complementary sequences

TFIDF (1,1)											
		Training Set					**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
DT	0.51	49.104	57.706	53.405	0.068	0.536	57.971	51.429	54.676	0.094	0.547
RF	0.51	57.348	56.272	56.810	0.136	0.583	60.870	54.286	57.554	0.152	0.596
LR	0.50	59.140	60.932	60.036	0.201	0.621	53.623	47.143	50.360	0.008	0.553
XGB	0.47	55.197	54.122	54.659	0.093	0.580	66.667	50.000	58.273	0.169	0.614
KNN	0.41	54.480	54.480	54.480	0.090	0.548	60.870	52.857	56.835	0.138	0.549
GNB	0.50	58.065	59.857	58.961	0.179	0.625	53.623	58.571	56.115	0.122	0.574
SVC	0.51	59.498	62.007	60.753	0.215	0.628	50.725	54.286	52.518	0.050	0.626
ET	0.50	58.065	53.405	55.735	0.115	0.573	72.464	50.000	61.151	0.230	0.605
TFIDF (1,2)											
		Training Set					**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
DT	0.51	58.065	60.215	59.140	0.183	0.591	75.362	50.000	62.590	0.262	0.627
RF	0.51	65.591	63.082	64.337	0.287	0.698	72.464	52.857	62.590	0.258	0.685
LR	0.49	62.007	59.498	60.753	0.215	0.653	68.116	54.286	61.151	0.226	0.636
XGB	0.56	64.158	64.158	64.158	0.283	0.697	65.217	57.143	61.151	0.224	0.652
KNN	0.41	59.498	62.366	60.932	0.219	0.645	59.420	64.286	61.871	0.237	0.649
GNB	0.49	62.724	62.366	62.545	0.251	0.658	63.768	54.286	58.993	0.181	0.645
SVC	0.50	65.233	65.233	65.233	0.305	0.696	62.319	57.143	59.712	0.195	0.664
ET	0.51	64.875	65.591	65.233	0.305	0.704	68.116	60.000	64.029	0.282	0.737
TFIDF (1,3)											

		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
DT	0.50	62.007	62.007	62.007	0.240	0.620	65.217	57.143	61.151	0.224	0.612
RF	0.51	65.233	65.233	65.233	0.305	0.720	72.464	57.143	64.748	0.299	0.693
LR	0.50	62.007	60.932	61.470	0.229	0.674	75.362	60.000	67.626	0.358	0.674
XGB	0.49	65.591	65.233	65.412	0.308	0.705	69.565	60.000	64.748	0.297	0.671
KNN	0.41	63.082	67.384	65.233	0.305	0.694	76.812	51.429	64.029	0.292	0.683
GNB	0.27	62.007	61.649	61.828	0.237	0.677	72.464	51.429	61.871	0.244	0.672
SVC	0.51	63.799	64.158	63.978	0.280	0.700	66.667	60.000	63.309	0.267	0.693
ET	0.51	67.384	66.667	67.025	0.341	0.727	69.565	61.429	65.468	0.311	0.688
TFIDF (1,4)											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
DT	0.50	58.065	57.348	57.706	0.154	0.577	68.116	55.714	61.871	0.240	0.619
RF	0.52	64.875	63.441	64.158	0.283	0.718	73.913	57.143	65.468	0.315	0.717
LR	0.50	64.158	61.649	62.903	0.258	0.679	71.014	57.143	64.029	0.284	0.665
XGB	0.52	64.875	64.158	64.516	0.290	0.701	71.014	55.714	63.309	0.270	0.685
KNN	0.41	61.649	61.649	61.649	0.233	0.672	65.217	61.429	63.309	0.267	0.665
GNB	0.01	48.029	75.269	61.649	0.242	0.662	57.971	57.143	57.554	0.151	0.666
SVC	0.50	64.516	62.366	63.441	0.269	0.704	66.667	57.143	61.871	0.239	0.690
ET	0.52	64.516	64.875	64.695	0.294	0.705	72.464	61.429	66.906	0.341	0.723
TFIDF reverse complement sequence (RC) (1,1)											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
DT	0.51	50.896	58.781	54.839	0.097	0.548	56.522	48.571	52.518	0.051	0.536
RF	0.50	58.065	59.498	58.781	0.176	0.597	63.768	50.000	56.835	0.139	0.594
LR	0.50	59.140	60.932	60.036	0.201	0.621	53.623	47.143	50.360	0.008	0.553
XGB	0.48	56.631	55.914	56.272	0.125	0.581	65.217	54.286	59.712	0.196	0.615
KNN	0.41	54.480	54.480	54.480	0.090	0.548	60.870	52.857	56.835	0.138	0.549
GNB	0.50	58.065	59.857	58.961	0.179	0.625	53.623	58.571	56.115	0.122	0.574
SVC	0.51	59.498	62.007	60.753	0.215	0.628	50.725	54.286	52.518	0.050	0.626
ET	0.51	54.839	57.348	56.093	0.122	0.573	57.971	52.857	55.396	0.108	0.596
TFIDF reverse complement sequence (RC) (1,2)											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
DT	0.51	54.480	63.441	58.961	0.180	0.590	68.116	54.286	61.151	0.226	0.612
RF	0.51	64.875	65.591	65.233	0.305	0.701	65.217	60.000	62.590	0.252	0.695
LR	0.49	62.007	59.498	60.753	0.215	0.653	68.116	54.286	61.151	0.226	0.636
XGB	0.49	62.007	62.007	62.007	0.240	0.688	66.667	52.857	59.712	0.197	0.659

KNN	0.41	59.498	62.366	60.932	0.219	0.645	59.420	64.286	61.871	0.237	0.649
GNB	0.49	62.724	62.366	62.545	0.251	0.658	63.768	54.286	58.993	0.181	0.645
SVC	0.50	65.233	65.233	65.233	0.305	0.696	62.319	57.143	59.712	0.195	0.664
ET	0.50	66.308	64.158	65.233	0.305	0.718	73.913	61.429	67.626	0.356	0.744
TFIDF reverse complement sequence (RC) (1,3)											
		Training Set					**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
DT	0.50	60.573	63.799	62.186	0.244	0.622	65.217	65.714	65.468	0.309	0.655
RF	0.52	63.799	65.233	64.516	0.290	0.708	68.116	54.286	61.151	0.226	0.661
LR	0.50	62.007	60.932	61.470	0.229	0.674	75.362	60.000	67.626	0.358	0.674
XGB	0.51	64.875	64.875	64.875	0.297	0.704	68.116	50.000	58.993	0.184	0.686
KNN	0.41	63.082	67.384	65.233	0.305	0.694	76.812	51.429	64.029	0.292	0.683
GNB	0.27	62.007	61.649	61.828	0.237	0.677	72.464	51.429	61.871	0.244	0.672
SVC	0.51	63.799	64.158	63.978	0.280	0.700	66.667	60.000	63.309	0.267	0.693
ET	0.50	64.875	67.384	66.129	0.323	0.714	71.014	55.714	63.309	0.270	0.683
TFIDF reverse complement sequence (RC) (1,4)											
		Training Set					**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
DT	0.50	55.914	55.556	55.735	0.115	0.557	62.319	52.857	57.554	0.152	0.576
RF	0.52	64.516	66.667	65.591	0.312	0.715	73.913	58.571	66.187	0.329	0.710
LR	0.50	64.158	61.649	62.903	0.258	0.679	71.014	57.143	64.029	0.284	0.665
XGB	0.52	64.875	65.233	65.054	0.301	0.692	69.565	61.429	65.468	0.311	0.714
KNN	0.41	61.649	61.649	61.649	0.233	0.672	65.217	61.429	63.309	0.267	0.665
GNB	0.01	48.029	75.269	61.649	0.242	0.662	57.971	57.143	57.554	0.151	0.666
SVC	0.50	64.516	62.366	63.441	0.269	0.704	66.667	57.143	61.871	0.239	0.690
ET	0.51	66.308	66.308	66.308	0.326	0.717	79.710	58.571	69.065	0.391	0.725

III. **Distance distribution of nucleotides:** We computed the distance distribution among nucleotides using Nfeature (Mathur et al., 2021). The highest AUC performance for these features was 0.609 for training and 0.621 for the independent validation set. The results for ML models developed on these features are given in Table 6.4.

Table 6.4: Results for ML models developed on features extracted by calculating the distance distribution of nucleotides in miRNA sequences

		Training Set					**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	54.839	53.763	54.301	0.540	0.086	23.188	80.000	51.799	0.525	0.039

RF	0.51	56.989	58.781	57.885	0.609	0.158	47.826	65.714	56.835	0.621	0.138
LR	0.5	62.366	37.634	50.000	0.536	0.050	44.928	60.000	52.518	0.539	0.050
XGB	0.49	58.065	55.556	56.810	0.589	0.136	55.072	57.143	56.115	0.567	0.122
KNN	0.49	53.763	54.122	53.943	0.573	0.079	52.174	55.714	53.957	0.569	0.079
GNB	0.52	58.423	52.688	55.556	0.578	0.111	53.623	54.286	53.957	0.536	0.079
SVC	0.52	56.631	54.839	55.735	0.602	0.115	53.623	60.000	56.835	0.601	0.137

IV. **Nucleotide repeat index:** We calculated a **repeat index** that measures the extent of consecutive nucleotide repeats (e.g., "AAA" or "GGGG" runs) in the sequence. These features did not perform well, yielding the highest AUC of 0.593 for training and 0.594 for independent validation sets. The results for ML models developed on this feature are shown in Table 6.5.

Table 6.5: Results for ML models developed on features extracted by calculating nucleotide repeat index in miRNA sequences

		Training Set					**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.56	55.556	52.688	54.122	0.569	0.082	60.870	50.000	55.396	0.573	0.109
RF	0.53	57.706	54.839	56.272	0.602	0.125	55.072	58.571	56.835	0.574	0.137
LR	0.49	62.724	49.104	55.914	0.594	0.119	62.319	47.143	54.676	0.576	0.096
XGB	0.51	58.065	55.197	56.631	0.591	0.133	69.565	45.714	57.554	0.573	0.157
KNN	0.53	55.197	54.480	54.839	0.570	0.097	66.667	51.429	58.993	0.571	0.183
GNB	0.61	55.556	58.065	56.810	0.593	0.136	66.667	48.571	57.554	0.594	0.155
SVC	0.53	51.254	54.839	53.047	0.576	0.061	53.623	52.857	53.237	0.557	0.065

V. **Entropy-based features:** We calculated Shannon entropy at two different levels: (i) sequence-level entropy, treating the entire miRNA as a sequence of nucleotides, and (ii) nucleotide-level entropy, focusing on the distribution of each nucleotide across the sequence. These entropy measures capture the randomness or complexity of the sequence. The highest performance achieved by entropy features was an AUC of 0.56 for sequence-level entropy and 0.59 for nucleotide-level entropy. The results for both nucleotide and sequence level entropy are given in Table 6.6.

Table 6.6: Results for ML models developed on features extracted by calculating Shannon entropy on nucleotide and sequence level of miRNA sequences

Entropy – nucleotide level											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.56	57.348	58.065	57.706	0.566	0.154	50.725	50.000	50.360	0.552	0.007
RF	0.5	55.914	58.065	56.989	0.598	0.140	52.174	55.714	53.957	0.584	0.079
LR	0.5	54.480	59.140	56.810	0.567	0.136	36.232	68.571	52.518	0.547	0.051
XGB	0.5	57.706	59.140	58.423	0.591	0.168	57.971	52.857	55.396	0.587	0.108
KNN	0.49	55.914	56.631	56.272	0.580	0.125	50.725	57.143	53.957	0.583	0.079
GNB	0.5	30.108	79.211	54.659	0.557	0.107	17.391	87.143	52.518	0.551	0.063
SVC	0.49	55.556	58.781	57.168	0.593	0.143	44.928	61.429	53.237	0.594	0.064
Entropy – sequence level											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	46.237	59.498	52.867	0.540	0.058	52.174	52.857	52.518	0.522	0.050
RF	0.51	50.538	51.613	51.075	0.535	0.022	56.522	48.571	52.518	0.522	0.051
LR	0.5	37.634	52.688	45.161	0.429	-0.098	31.884	61.429	46.763	0.439	-0.070
XGB	0.52	53.405	55.914	54.659	0.558	0.093	57.971	47.143	52.518	0.545	0.051
KNN	0.51	72.043	30.108	51.075	0.527	0.024	76.812	28.571	52.518	0.529	0.061
GNB	0.46	55.914	49.104	52.509	0.524	0.050	49.275	61.429	55.396	0.548	0.108
SVC	0.5	63.082	40.860	51.971	0.544	0.040	85.507	12.857	48.921	0.561	-0.024

VI. **Correlation:** We calculated the autocorrelation of identical nucleotides and the auto-cross correlation between different nucleotides of nucleotide pairs. The best performing among these was the autocorrelation of dinucleotide sequences, which reached an AUC of ~0.62 on the independent validation set. The results are given in Table 6.7.

Table 6.7: Results for ML models developed on correlation features of miRNA sequences

Autocorrelation											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	55.197	56.631	55.914	0.558	0.118	42.029	65.714	53.957	0.559	0.080
RF	0.53	53.047	59.498	56.272	0.598	0.126	47.826	72.857	60.432	0.614	0.214
LR	0.52	53.405	57.706	55.556	0.581	0.111	57.971	54.286	56.115	0.607	0.123
XGB	0.52	59.140	57.706	58.423	0.613	0.168	47.826	67.143	57.554	0.620	0.153
KNN	0.51	40.143	69.176	54.659	0.563	0.097	31.884	71.429	51.799	0.557	0.036
GNB	0.62	54.480	57.348	55.914	0.583	0.118	49.275	61.429	55.396	0.594	0.108
SVC	0.53	53.405	54.839	54.122	0.581	0.082	59.420	58.571	58.993	0.600	0.180

Auto-cross Correlation											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	56.989	52.330	54.659	0.568	0.093	76.812	37.143	56.835	0.557	0.152
RF	0.52	58.423	55.197	56.810	0.598	0.136	52.174	61.429	56.835	0.604	0.137
LR	0.52	53.405	57.706	55.556	0.581	0.111	57.971	54.286	56.115	0.607	0.123
XGB	0.52	59.140	57.706	58.423	0.613	0.168	47.826	67.143	57.554	0.620	0.153
KNN	0.51	40.143	69.176	54.659	0.563	0.097	31.884	71.429	51.799	0.557	0.036
GNB	0.62	54.480	57.348	55.914	0.583	0.118	49.275	61.429	55.396	0.594	0.108
SVC	0.53	51.971	54.839	53.405	0.583	0.068	56.522	58.571	57.554	0.588	0.151

VII. **Pseudo Composition:** We implemented pseudo-composition features, which included serial and parallel correlation pseudo-composition for dinucleotides. The best performer here was the serial correlation pseudo-composition, achieving ~0.708 AUC on the training set and ~0.707 on the validation set, indicating a stable predictive contribution. The results for pseudo composition are given in Table 6.8.

Table 6.8: Results for ML models developed on pseudo composition of nucleotides in miRNA sequences

Pseudo Composition of nucleotides											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.52	59.857	58.781	59.319	0.598	0.186	43.478	75.714	59.712	0.630	0.203
RF	0.51	66.308	65.233	65.771	0.707	0.315	69.565	55.714	62.590	0.715	0.255
LR	0.5	59.498	59.498	59.498	0.649	0.190	66.667	57.143	61.871	0.652	0.239
XGB	0.52	62.724	62.724	62.724	0.666	0.254	68.116	61.429	64.748	0.659	0.296
KNN	0.49	65.233	64.875	65.054	0.704	0.301	69.565	64.286	66.906	0.697	0.339
GNB	0.45	62.007	63.082	62.545	0.653	0.251	60.87	54.286	57.554	0.633	0.152
SVC	0.5	61.290	60.215	60.753	0.663	0.215	72.464	55.714	64.029	0.666	0.286
Parallel correlation pseudo composition											
		Training Set					Independent Validation Set				
Model	Thr	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
DT	0.52	59.498	59.498	59.498	0.631	0.190	52.174	61.429	56.835	0.582	0.137
RF	0.51	66.667	67.025	66.846	0.720	0.337	69.565	61.429	65.468	0.692	0.311
LR	0.5	60.215	60.215	60.215	0.650	0.204	66.667	57.143	61.871	0.655	0.239
XGB	0.53	63.441	63.441	63.441	0.689	0.269	65.217	61.429	63.309	0.664	0.267
KNN	0.49	63.441	64.516	63.978	0.700	0.280	68.116	57.143	62.59	0.696	0.254

GNB	0.44	63.082	63.799	63.441	0.658	0.269	57.971	54.286	56.115	0.626	0.123
SVC	0.5	62.724	60.573	61.649	0.670	0.233	71.014	54.286	62.59	0.669	0.257
Serial correlation pseudo composition											
		Training Set					**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	59.140	59.857	59.498	0.630	0.190	75.362	41.429	58.273	0.647	0.178
RF	0.51	67.742	66.667	67.204	0.717	0.344	65.217	60.000	62.590	0.701	0.252
LR	0.5	58.781	60.932	59.857	0.650	0.197	69.565	58.571	64.029	0.656	0.283
XGB	0.53	60.932	61.649	61.290	0.665	0.226	57.971	55.714	56.835	0.642	0.137
KNN	0.49	65.950	65.591	65.771	0.708	0.315	71.014	61.429	66.187	0.707	0.326
GNB	0.49	62.007	62.007	62.007	0.662	0.240	59.420	55.714	57.554	0.637	0.151
SVC	0.5	61.649	61.290	61.470	0.669	0.229	69.565	57.143	63.309	0.669	0.269

VIII. **Binary profiles:** We used One-hot encoding to represent each miRNA as a binary vector. In this encoding, each position in the miRNA sequence is represented by four binary variables. The ML model developed using binary profiles achieved up to 0.75 AUC on the independent validation set. The results for ML models developed on the binary profile of miRNA sequences are given in Table 6.9.

Table 6.9: Results for ML models developed on binary profiles extracted for miRNA sequences by performing one-hot encoding on the sequences

		Training Set					**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.49	55.914	56.631	56.272	0.581	0.125	52.174	57.143	54.676	0.563	0.093
RF	0.51	67.384	65.591	66.487	0.747	0.330	68.116	70.000	69.065	0.754	0.381
LR	0.51	63.082	63.441	63.262	0.685	0.265	66.667	54.286	60.432	0.665	0.211
XGB	0.51	62.366	61.649	62.007	0.685	0.240	68.116	52.857	60.432	0.683	0.212
KNN	0.43	67.025	63.082	65.054	0.709	0.301	76.812	57.143	66.906	0.713	0.346
GNB	0.51	62.366	63.082	62.724	0.673	0.254	71.014	57.143	64.029	0.670	0.284
SVC	0.51	64.516	64.516	64.516	0.702	0.290	79.710	57.143	68.345	0.706	0.378
ET	0.51	64.158	68.459	66.308	0.740	0.326	68.116	60.000	64.029	0.745	0.282

IX. **Best features:** We combined the best-performing features – composition-based features (TFIDF of reverse complementary sequences for k-mers (1,2)) and binary features. These features, when combined, achieved the highest AUC of 0.763 for training and 0.769 for independent validation sets on an ET classifier. The results for the combined best features are given in Table 6.10. We also conducted a Mann-Whitney

test on these features, which showed that 31 features were statistically significant ($p<0.05$) among 145 total features. The results of the Mann–Whitney analysis are shown in Table 6.11.

Table 6.10: The performance metrics for ML models developed on combining binary profile and reverse complement TFIDF features [https://doi.org/10.1101/2025.03.19.644072]

Best features combined – Binary-profile based features + Reverse complement TFIDF (1,2)											
		Training Set					Independent Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	54.122	60.573	57.348	0.573	0.147	65.217	44.286	54.676	0.548	0.097
RF	0.52	67.384	69.892	68.638	0.745	0.373	68.116	67.143	67.626	0.735	0.353
LR	0.51	65.233	65.950	65.591	0.682	0.312	73.913	57.143	65.468	0.688	0.315
XGB	0.57	67.384	67.025	67.204	0.734	0.344	73.913	65.714	69.784	0.721	0.398
KNN	0.50	66.308	62.724	64.516	0.712	0.291	85.507	50.000	67.626	0.723	0.379
GNB	0.51	63.082	62.724	62.903	0.681	0.258	69.565	58.571	64.029	0.672	0.283
SVC	0.51	66.667	64.875	65.771	0.735	0.315	78.261	64.286	71.223	0.760	0.429
ET	0.51	65.950	68.459	67.204	0.763	0.344	71.014	70.000	70.504	0.769	0.410

Table 6.11: Results for the Mann-Whitney U test for the best features (here, Features like A_1, A_2,... stand for OHE features; and features like A_rc, T_rc,... stand for TFIDF features for reverse complementary strand for k-mer range (1,2))

Feature	Mean_E	Mean_NE	Mean_Diff	p-Value	Feature	Mean_E	Mean_NE	Mean_Diff	p-Value
A_1	0.256	0.324	-0.068	0.048	U_15	0.167	0.318	-0.151	0.000
C_1	0.244	0.181	0.064	0.040	X_15	0.000	0.000	0.000	1.000
G_1	0.092	0.086	0.006	0.781	A_16	0.267	0.178	0.090	0.004
U_1	0.408	0.410	-0.002	0.964	C_16	0.250	0.266	-0.016	0.620
X_1	0.000	0.000	0.000	1.000	G_16	0.218	0.241	-0.022	0.484
A_2	0.319	0.355	-0.036	0.311	U_16	0.264	0.315	-0.051	0.140
C_2	0.210	0.172	0.038	0.204	X_16	0.000	0.000	0.000	1.000
G_2	0.247	0.261	-0.014	0.680	A_17	0.152	0.232	-0.080	0.008
U_2	0.224	0.212	0.012	0.699	C_17	0.279	0.221	0.058	0.077
X_2	0.000	0.000	0.000	1.000	G_17	0.276	0.226	0.050	0.132
A_3	0.284	0.332	-0.048	0.171	U_17	0.293	0.321	-0.028	0.427
C_3	0.207	0.192	0.015	0.623	X_17	0.000	0.000	0.000	1.000
G_3	0.250	0.232	0.018	0.581	A_18	0.181	0.223	-0.042	0.163
U_3	0.259	0.244	0.015	0.647	C_18	0.204	0.252	-0.048	0.130
X_3	0.000	0.000	0.000	1.000	G_18	0.287	0.226	0.061	0.066
A_4	0.250	0.289	-0.039	0.242	U_18	0.316	0.287	0.030	0.396
C_4	0.259	0.206	0.052	0.102	X_18	0.011	0.011	0.000	0.998

G_4	0.293	0.304	-0.011	0.760	A_19	0.187	0.209	-0.022	0.459
U_4	0.198	0.201	-0.002	0.940	C_19	0.195	0.246	-0.051	0.105
X_4	0.000	0.000	0.000	1.000	G_19	0.302	0.192	0.110	0.001
A_5	0.276	0.255	0.021	0.534	U_19	0.293	0.324	-0.031	0.381
C_5	0.181	0.198	-0.017	0.575	X_19	0.023	0.029	-0.006	0.638
G_5	0.305	0.327	-0.022	0.532	A_20	0.144	0.249	-0.106	0.000
U_5	0.239	0.221	0.018	0.575	C_20	0.224	0.172	0.052	0.084
X_5	0.000	0.000	0.000	1.000	G_20	0.310	0.235	0.075	0.026
A_6	0.190	0.287	-0.097	0.003	U_20	0.282	0.292	-0.011	0.756
C_6	0.290	0.244	0.047	0.164	X_20	0.040	0.052	-0.011	0.475
G_6	0.256	0.229	0.027	0.414	A_21	0.158	0.226	-0.068	0.022
U_6	0.264	0.241	0.024	0.472	C_21	0.250	0.186	0.064	0.042
X_6	0.000	0.000	0.000	1.000	G_21	0.210	0.255	-0.045	0.158
A_7	0.287	0.312	-0.025	0.472	U_21	0.310	0.244	0.067	0.049
C_7	0.264	0.241	0.024	0.472	X_21	0.072	0.089	-0.017	0.410
G_7	0.207	0.206	0.001	0.985	A_22	0.126	0.163	-0.037	0.167
U_7	0.241	0.241	0.001	0.983	C_22	0.138	0.155	-0.017	0.531
X_7	0.000	0.000	0.000	1.000	G_22	0.227	0.155	0.072	0.015
A_8	0.221	0.272	-0.051	0.119	U_22	0.264	0.181	0.084	0.008
C_8	0.261	0.206	0.055	0.085	X_22	0.244	0.347	-0.102	0.003
G_8	0.230	0.226	0.004	0.912	A_23	0.034	0.026	0.009	0.503
U_8	0.287	0.295	-0.008	0.822	C_23	0.034	0.026	0.009	0.503
X_8	0.000	0.000	0.000	1.000	G_23	0.049	0.060	-0.011	0.511
A_9	0.190	0.261	-0.071	0.025	U_23	0.063	0.049	0.015	0.405
C_9	0.244	0.235	0.009	0.774	X_23	0.819	0.840	-0.021	0.471
G_9	0.201	0.215	-0.014	0.655	A_24	0.003	0.006	-0.003	0.566
U_9	0.365	0.289	0.076	0.034	C_24	0.000	0.003	-0.003	0.319
X_9	0.000	0.000	0.000	1.000	G_24	0.003	0.009	-0.006	0.318
A_10	0.207	0.281	-0.074	0.023	U_24	0.011	0.000	0.011	0.045
C_10	0.264	0.238	0.027	0.420	X_24	0.983	0.983	0.000	0.997
G_10	0.310	0.209	0.101	0.002	A_25	0.000	0.000	0.000	1.000
U_10	0.218	0.272	-0.054	0.099	C_25	0.000	0.000	0.000	1.000
X_10	0.000	0.000	0.000	1.000	G_25	0.003	0.000	0.003	0.318
A_11	0.178	0.272	-0.094	0.003	U_25	0.000	0.003	-0.003	0.319
C_11	0.227	0.209	0.018	0.569	X_25	0.997	0.997	0.000	1.000
G_11	0.322	0.275	0.047	0.178	A_rc	0.404	0.407	-0.003	0.670
U_11	0.273	0.244	0.029	0.375	AA_rc	0.126	0.150	-0.024	0.067
X_11	0.000	0.000	0.000	1.000	AC_rc	0.127	0.099	0.028	0.000
A_12	0.239	0.246	-0.008	0.808	AG_rc	0.129	0.122	0.008	0.514
C_12	0.233	0.235	-0.002	0.946	AU_rc	0.101	0.120	-0.019	0.044
G_12	0.250	0.223	0.027	0.411	C_rc	0.376	0.341	0.034	0.002
U_12	0.279	0.295	-0.016	0.633	CA_rc	0.145	0.137	0.009	0.214

X_12	0.000	0.000	0.000	1.000	CC_rc	0.140	0.118	0.022	0.057
A_13	0.218	0.244	-0.025	0.431	CG_rc	0.085	0.045	0.040	0.000
C_13	0.216	0.186	0.029	0.335	CU_rc	0.114	0.125	-0.011	0.098
G_13	0.270	0.255	0.015	0.651	G_rc	0.342	0.316	0.026	0.008
U_13	0.296	0.315	-0.019	0.582	GA_rc	0.119	0.094	0.025	0.001
X_13	0.000	0.000	0.000	1.000	GC_rc	0.101	0.107	-0.006	0.249
A_14	0.267	0.241	0.027	0.421	GG_rc	0.116	0.111	0.005	0.360
C_14	0.204	0.221	-0.017	0.592	GU_rc	0.110	0.103	0.007	0.520
G_14	0.259	0.246	0.012	0.711	U_rc	0.330	0.386	-0.056	6.13E-07
U_14	0.270	0.292	-0.022	0.516	UA_rc	0.096	0.120	-0.023	4.1E-03
X_14	0.000	0.000	0.000	1.000	UC_rc	0.110	0.108	0.001	8.6E-01
A_15	0.273	0.266	0.007	0.847	UG_rc	0.119	0.122	-0.003	4.5E-01
C_15	0.239	0.206	0.032	0.307	UU_rc	0.103	0.152	-0.049	6.8E-06
G_15	0.322	0.209	0.113	0.001					

X. **Feature Importance:** Finally, we examined the top features in the Extra Trees classifier developed on the best-performing features to interpret the model. We selected the 20 most important features according to the model's feature importance. Out of these 20, about 15 features were significantly different ($p<0.05$) between the two classes. In one-hot encoding (binary) features, A at 16^{th}, G at 19^{th}, G at 15^{th}, G at 10^{th} positions were seen to be significantly higher in exosomal miRNA sequences, whereas, U at 15^{th}, A at 11^{th}, and A at 20^{th} positions were seen to be considerably lower in exosomal miRNA sequences. In TF-IDF features calculated for the reverse complements of miRNA sequences for the (1,2) range, we observed that C, CG, GA, and AC incidence were significantly higher in reverse complementary exosomal sequences. In contrast, U, UU, UA, and AU were found at much lower levels in exosomal miRNAs compared to non-exosomal miRNAs. A comparison of the top 20 important features that can effectively classify exosomal and non-exosomal miRNA is shown in Figure 6.2, and the top 20 features with their importance are given in Table 6.12.

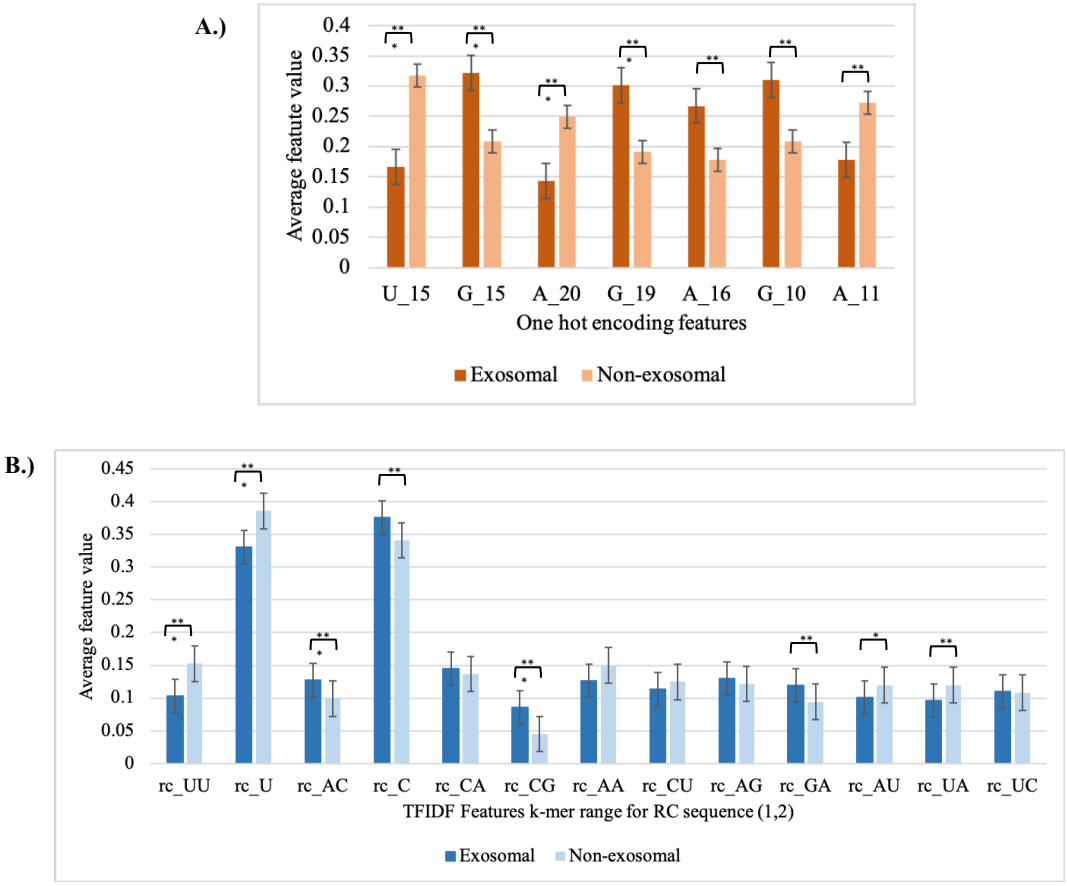


Figure 6.2: The top 20 most important features for differentiating abundant and non-abundant miRNA in exosomes: A) One-hot encoding (binary) features, B) TFIDF features

Table 6.12: Results for the top 20 most important features in the best features set extracted using Extra Tree Model (here, Features like A_1, A_2,.. Stand for OHE features; and features like A_rc, T_rc,.. Stand for TFIDF features for reverse complementary strand for k-mer range (1,2))

Feature	Importance	Feature	Importance
UU_rc	0.0174	**CA_rc**	0.0126
U_rc	0.0158	**CG_rc**	0.0122
U_15	0.0148	**G_10**	0.0117
AC_rc	0.0146	**AA_rc**	0.0115
X_22	0.0140	**CU_rc**	0.0112

G_15	0.0137	**AG_rc**	0.0109
A_20	0.0136	**GA_rc**	0.0108
G_19	0.0130	**AU_rc**	0.0107
A_16	0.0129	**A_11**	0.0107
C_rc	0.0127	**UA_rc**	0.0107

6.3.4 Ensemble Method

To improve the predictive performance of our best-performing ML model, we combined Alignment-based approaches with AI-based approaches to develop an ensemble method. As motif-search was not performing well on the dataset, yielding a very poor coverage and high error rate, we combined similarity-search with AI-based methods, as it had relatively higher coverage with zero error. The hybrid model achieved an AUC of 0.854 on the independent validation set, notably higher than the 0.769 of the ML model alone. It also improved the validation MCC to 0.559. Table 6.13 shows the performance of the hybrid model, and Figure 6.3 shows the AUC curve of the ensemble approach.

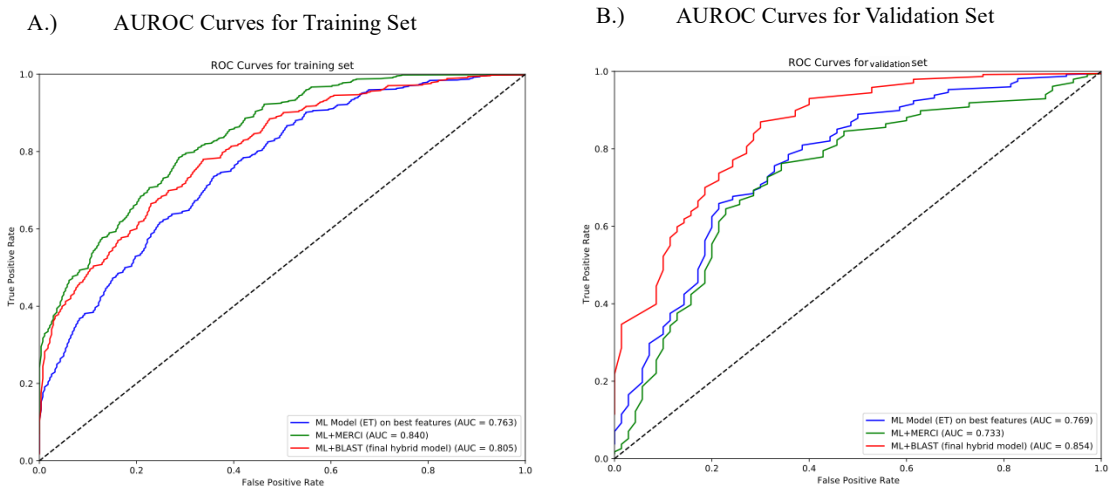


Figure 6.3: The AUROCs plots depicted for the performance of the best ML model, and ensemble models (ML + motif-search using MERCI, ML + similarity-search using BLAST), for A) training set and B) validation set [https://doi.org/10.1101/2025.03.19.644072]

Table 6.13: The results for the hybrid/ensemble model: AI-based methods developed on the best-performing features combined with the similarity-search (Alignment-based) method [https://doi.org/10.1101/2025.03.19.644072]

		Training dataset (cross-validation)					Validation/Independent dataset				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	0.51	58.78	60.57	59.68	0.676	0.194	65.22	52.86	58.99	0.671	0.182
RF	0.52	73.12	71.68	72.40	0.801	0.448	75.36	71.43	73.38	0.832	0.468
LR	0.55	69.89	70.25	70.07	0.781	0.401	81.16	68.57	74.82	0.824	0.501
XGB	0.61	70.25	73.48	71.86	0.787	0.437	84.06	72.86	78.42	0.821	0.572
KN	0.50	71.68	65.23	68.46	0.778	0.370	88.41	57.14	72.66	0.825	0.479
GNB	0.54	69.18	69.53	69.35	0.780	0.387	79.71	70.00	74.82	0.825	0.499
SVC	0.52	70.61	69.53	70.07	0.797	0.401	79.70	71.43	77.70	0.846	0.513
ET	0.50	76.34	67.38	71.86	0.805	0.439	84.06	71.43	77.70	0.854	0.559

6.3.5 Comparison with existing methods

We compared AdmirePred's predictions with those of existing computational tools. In particular, we evaluated EmiRPred, a method developed to predict whether a given miRNA is exosomal or non-exosomal based on its subcellular location (Arora & Raghava, 2024). In addition, two general miRNA subcellular localization predictors were used - miRNALoc and EL-RMLocNet. Both of these include "exosome" as one of several possible localization compartments (Asim et al., 2022; Meher, Satpathy, et al., 2021). We used our independent validation set of 139 sequences as input to these tools. We found that miRNALoc's web server could process our sequences and produce predictions labeling each miRNA to one of multiple compartments (nucleus, cytoplasm, exosome, etc.). However, miRNALoc's performance on our dataset was limited – it achieved an AUC of only ~0.422 on our validation set, compared to AdmirePred, which showed a significantly higher AUC of 0.854. We also attempted to use EL-RMLocNet, an LSTM-based model for multi-compartment RNA localization. Unfortunately, EL-RMLocNet only processes only one sequence at a time, which makes it unfeasible to use it for 139 sequences. Since the training data for EmiRPred shared many miRNAs with AdmirePred, we selected only those miRNA sequences that were uniquely found in the AdmirePred dataset and submitted them to the EmiRPred server. EmiRPred could classify these miRNAs as exosomal or non-exosomal, achieving an AUC of 0.623. These results collectively suggest that AdmirePred performs better than existing computational tools in identifying miRNAs enriched in exosomes.

6.3.6 Web Server and Standalone Software

To promote the accessibility of our approach to the broader research community, we developed a web server for AdmirePred. The web server provides three main modules: Predict, Design, and BLAST-Search.

I. **Predict module:** Allows users to input one or multiple miRNA sequences (in FASTA format or plain text). Upon submission, the server computes all the necessary features and applies our trained hybrid model to predict whether each miRNA will likely be abundant in exosomes. The output provides a probability score for each miRNA in the exosomal (abundant) class, and a classification (Exosomal vs. Non-exosomal) based on the default threshold. This enables experimental biologists to quickly screen miRNAs of interest for their propensity to be found in exosomes.

II. **Design module:** This tool helps users explore sequence variants of a miRNA. A user can submit a miRNA sequence of interest, and the module will generate all possible single-nucleotide mutants of that miRNA. For each mutant sequence, AdmirePred provides a prediction score for being abundant in exosomes. This allows researchers to identify which nucleotide positions are most critical for a miRNA's exosomal localization.

III. **BLAST-search module:** This feature provides an interface to perform a similarity search against our curated database of exosomal and non-exosomal miRNAs. Users can input an miRNA, and the server will run a BLAST search against the dataset used in AdmirePred. The results show the closest matches among known exosomal and non-exosomal miRNAs for different e-values. This can give users a quick sense of whether their query has known exosomal miRNA homologs.

The web server is designed to be accessible on various devices and platforms. The interface is simple and responsive, allowing easy use from desktop computers and tablets or smartphones. In addition to the web server, we provide a Python package, a standalone command-line tool, and a GitHub repository for AdmirePred. We envision that AdmirePred will help researchers discover the most promising miRNA candidates for exosome-based biomarkers or therapeutic targets.

6.4 Discussions

Exosomal miRNAs play an essential role in mediating cell-to-cell communication and contributing to disease development, positioning them as promising candidates for diagnostic and therapeutic applications. (Mosquera-Heredia et al., 2021; Tastan et al., 2022). These small non-coding RNAs are selectively packaged within exosomes, which shield them from enzymatic degradation and enable their delivery to recipient cells (J. Zhang et al., 2015). Studies have shown the utility of exosomal miRNAs across various diseases and their influence on gene expression. In cancer, for example, exosomal miR-155 and miR-21 are known regulators of metastasis and tumor progression; their levels are elevated in exosomes derived from patients with pancreatic and breast cancers (X. Lai et al., 2017; Melo et al., 2014). In neurodegenerative diseases like Alzheimer's, exosomal miR-146a and miR-125b have been implicated in promoting neuroinflammation and synaptic dysfunction (L. Cheng et al., 2015). Similarly, in cardiovascular conditions, exosomal miR-133a and miR-1 levels rise in the serum of patients after myocardial injury, reflecting cardiac damage (Kuwabara et al., 2011). These examples underscore that changes in exosomal miRNA content can serve as a window into pathological processes occurring in tissues.

To facilitate exosome-based diagnostic and therapeutic applications, we carried out a study to predict which miRNAs are most abundant in exosomes under normal physiological conditions. Our goal is to provide researchers with a reference point for comparing miRNA levels in exosomes under various disease states. As a first step, we employed alignment-based methods, including both motif identification and similarity searches. For motif detection, we explored Gap values of 0, 1, and 2; however, this approach showed limited coverage (about 7.6%) and a high error rate. We then implemented a similarity search via the blastn-short tool, which demonstrated a 0% error rate but covered only ~ 35% of independent validation set at an e-value of 10^{-2}.

To address the shortcomings observed with alignment-based techniques, we developed machine learning models to predict which miRNAs appear most abundantly in exosomes. We extracted many features from miRNA sequences, including binary-profile representations and composition-based features such as nucleotide composition, entropy, correlation, and TF-IDF. Among these, binary-profile features and TF-IDF features for the reverse complement sequences with k-mers (1,2) emerged as the top performing features. By combining these features, we achieved a maximum AUC of 0.769 and an MCC of 0.410 on the independent

validation dataset. We then introduced an ensemble method that incorporates both alignment-based and AI-based classification, which improved the performance, leading to a prediction AUC of 0.854 and an MCC of 0.559.

We compared our approach against existing computational methods for predicting miRNA subcellular localization. These tools include EmiRPred, EL-RMLocnet, and miRNALoc, which determine where miRNAs localize within cells (Arora & Raghava, 2024; Asim et al., 2022; Meher, Satpathy, et al., 2021). Our method outperformed these existing methods, particularly in identifying abundant miRNAs in exosomes. A key factor in this improvement could be the incorporation of miRNA expression profiles, as elevated expression levels typically increase the likelihood of specific miRNAs being packaged into exosomes. The sorting mechanisms of exosomes often prioritize miRNAs essential for intercellular communication, which are more prevalent when their expression is elevated in specific cell types or disease states. By integrating these expression patterns, our framework captured important nuances and yielded more accurate predictions.

Despite the strengths of AdmirePred, there are areas for future improvement. One limitation is that we focused on miRNAs abundant in the blood exosomes of healthy individuals. It remains to be explored whether the same model and features apply to exosomes from other fluids or disease states.

Chapter 7

Identification of Salivary Biomarkers for Gastric Cancer

7.1 Introduction

Gastric cancer (GC) is a highly prevalent and deadly malignancy worldwide, and it remains among the top causes of cancer-related mortality (Smyth et al., 2020). Even with improvements in treatment, outcomes for GC patients are often poor because the disease is typically diagnosed at an advanced stage (Thrift & El-Serag, 2020). Early detection is critical for improving survival, yet existing diagnostic methods such as upper gastrointestinal endoscopy coupled with histopathological biopsy are invasive, expensive, and unfeasible for routine screening (Zhu et al., 2016). This creates a need for reliable, non-invasive biomarkers that can enable earlier diagnosis and better patient outcomes.

In recent years, liquid biopsy approaches have gained traction for cancer detection, leveraging circulating biomarkers found in accessible biofluids such as urine, blood, and saliva (Arora, Kaur, et al., 2023; Bao et al., 2024; Swarup et al., 2023). Saliva is an attractive medium because it can be collected easily, without invasive procedures, and it is known to reflect systemic disease states (Surdu et al., 2025). It carries various biomolecules like extracellular RNA (exRNA) species that serve as potential biomarkers for a number of diseases, including GC (Ghosh et al., 2024; Kaczor-Urbanowicz et al., 2022; F. Li, Yoshizawa, et al., 2018). These salivary exRNAs encompass mRNA, piRNA, miRNA, miscellaneous small RNAs, circRNA, and other RNA types (Figure 7.1).

A number of biomarker-driven approaches have been investigated in the past for the early detection of gastric cancer (GC) through biofluids such as plasma, serum, and saliva. Traditional blood-based markers like carbohydrate antigen 19-9 (CA19-9), carcinoembryonic antigen (CEA), and pepsinogen levels, are routinely used for GC screening. However, they have lower sensitivity and specificity, especially in early stages (Shibata et al., 2022). Circulating microRNA signatures in plasma have shown some promise; certain studies identified miR-20a, miR-185, miR-92b, and miR-210 as potential GC biomarkers, resulting in area-under-ROC values of ~0.65–0.75. However, such findings require validation in larger patient cohorts (X. Zhou et al., 2015).

Previously, Li et al. (2018) have explored the salivary extracellular RNA (exRNA) biomarkers for the non-invasive detection of gastric cancer (GC) by analyzing saliva samples. Their study identified a set of five biomarkers, including SPINK7, PPL, SEMA4B, miR-140-5p, and miR-301a-3p. These biomarkers were validated through reverse transcription quantitative real-time

PCR (RT-qPCR). The salivary biomarker panel identified in this study achieved an area under the ROC curve (AUC) of 0.81, which further increased to 0.87 when combined with patient demographic information. This study highlights the promise of salivary exRNAs for GC screening (Kaczor-Urbanowicz et al., 2022; F. Li, Yoshizawa, et al., 2018).

In the present study, we aimed to classify 98 GC patients versus 100 non-GC controls based on salivary exRNA profiles with high accuracy. We employed both a simple threshold-based classification and a machine learning-based approach. As a result, we identified a panel of eight mRNA biomarkers that can distinguish GC from normal samples, achieving an AUROC of 0.905 on an independent validation set. To ensure a comprehensive search for biomarkers, we also explored additional "secondary" biomarker sets beyond this primary eight-mRNA panel. This was accomplished by iteratively removing the already identified primary biomarkers from the feature pool and reapplying feature selection techniques. This strategy enabled us to uncover distinct but relevant secondary biomarker panels that exhibited varying degrees of overlap and correlation with the primary set, thereby contributing to a broader understanding of salivary exRNA profiles associated with GC. The complete workflow followed in this study is illustrated in Figure 7.2. In summary, the findings highlight the promise of salivary exRNA as a non-invasive diagnostic tool for gastric cancer, providing a potential alternative to conventional invasive techniques.

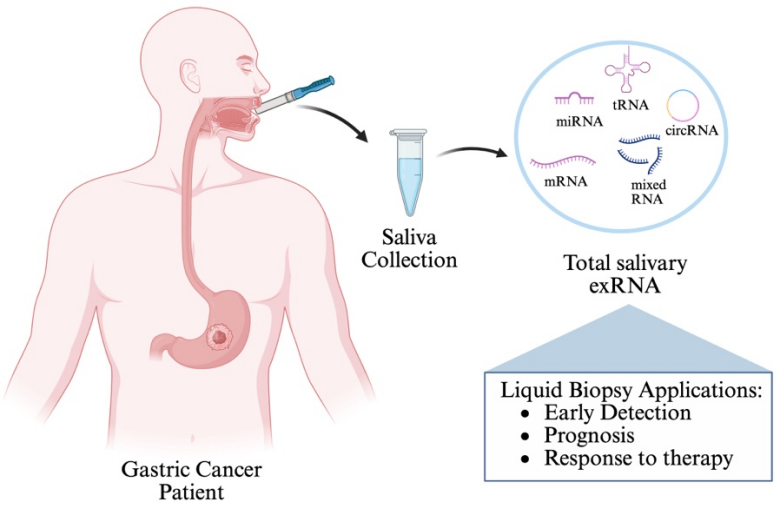


Figure 7.1: Extracellular RNA in saliva from Gastric Cancer patients

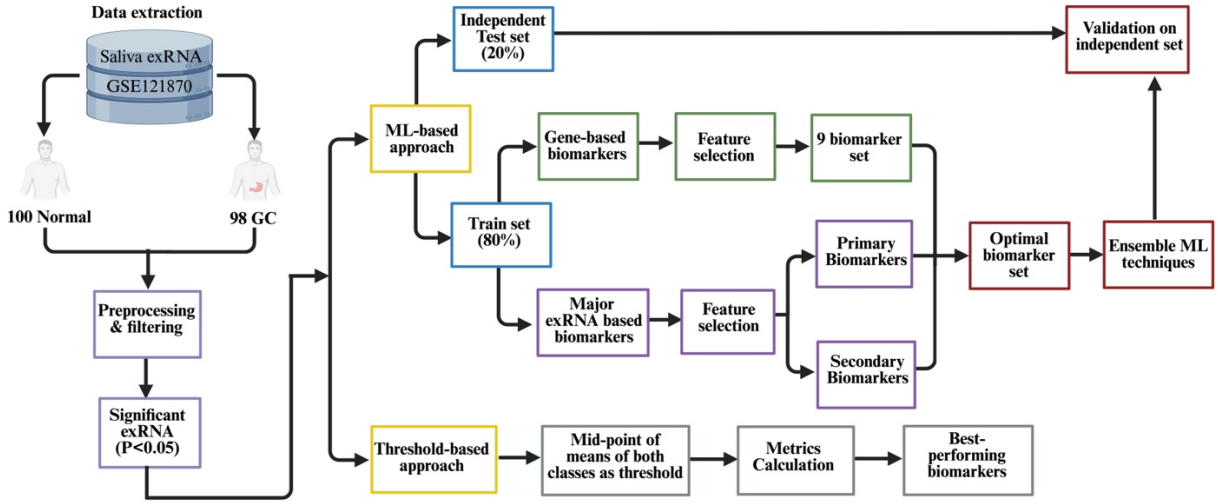


Figure 7.2: Methodology followed in the study

7.2 Methods

7.2.1 Data Collection

The dataset used in this study was obtained from the Gene Expression Omnibus (GEO) under accession number GSE121870 (F. Li, Kaczor-Urbanowicz, et al., 2018). It consists of salivary extracellular RNA (exRNA) sequencing data derived from cell-free saliva of 198 participants, comprising 100 non-gastric cancer (Normal) controls and 98 gastric cancer (GC) patients. The sequencing data is in reads per million (RPM) format for 19,704 mRNAs, 55,104 mixed RNAs (e.g., lincRNA, snRNA, etc.), 1,583 miRNAs, 2,281 piRNAs, 26 tRNAs, and 245 circRNAs.

7.2.2 Data Preprocessing

To refine the dataset, exRNA features missing in more than 20% of the samples were excluded. Subsequently, a constant value 1.0 was added to each RPM expression value, and then they were log2-transformed (Bhalla, Chaudhary, et al., 2017). We applied this log transformation to stabilize variance (Equation 1). Furthermore, we applied a Mann-Whitney U test to identify the statistically significant exRNA differences between groups. It is also known as the Wilcoxon rank-sum test, a non-parametric statistical method used to compare two independent groups when the data is not normally distributed (Arora & Raghava, 2024). After performing this test, we extracted the exRNAs that had a p-value < 0.05. Out of these significant features,

the most common RNA classes used as biomarkers (mRNA, circRNA, piRNA, miRNA, and tRNA) were selected for further analysis.

$$X\,(\text{transformed value}) = \log2\,(\text{RPM} + 1) \qquad \text{- Eq. 1}$$

7.2.3 Threshold-based Approach

In this method, we classified the samples into normal and GC using a threshold-based technique. For each salivary exRNA, we calculated the mean expression in the GC and normal samples, then used the midpoint of these two means as a threshold to classify samples as GC or normal. In doing so, we computed metrics such as true positives, true negatives, false positives, false negatives, and accuracy for each exRNA. Here, accuracy is the proportion of Normal and GC subjects that were correctly classified (Equation 2). We then identified the salivary exRNAs with the highest classification accuracies using this threshold method.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \times 100 \qquad \text{- Eq. 2}$$

7.2.4 Machine Learning-based Approach

We next applied several machine learning (ML) algorithms using the salivary exRNAs as features. The classifiers evaluated included ensemble tree-based methods (Random Forest (RF) and Extra Trees (ET)), Support Vector Classifier (SVC), Logistic Regression (LR), k-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGB), AdaBoost, and Gradient Boosting (GB) (Breiman, 2001; Cristianini & Ricci, 2008; Geurts et al., 2006; Stoltzfus, 2011; Y. Wu et al., 2002). In addition, we employed ensemble learning strategies: a Stacking Classifier (SC), which combines several base models using a meta-classifier to enhance predictive performance by learning the optimal combination of their outputs; and a Voting Classifier (VC) aggregates predictions from multiple base models and determines the final output based on majority voting.

7.2.4.1 Feature Selection

To identify the most informative biomarkers, we applied state-of-the-art feature selection techniques including Iterative Feature Selection (IFS), Recursive Feature Elimination (RFE), and an L1-regularized Support Vector Classifier (SVC-L1). Recursive Feature Elimination (RFE) progressively eliminates the least important features based on model performance,

aiming to optimize predictive accuracy. (Kaur et al., 2024). IFS starts from an empty feature set and progressively adds features, assessing model performance at each step to determine the optimal subset. SVC-L1 utilizes a linear Support Vector Classifier with L1 regularization, fostering sparsity by removing irrelevant features while retaining critical biomarkers essential for classification.

7.2.4.2 Gene-based Biomarkers

Firstly, we focused on identifying protein-coding genes (mRNAs) biomarkers, since mRNAs were the most abundant class of exRNAs in the dataset. Gene-based biomarkers are among the most commonly studied biomarkers for disease diagnosis and prognosis. We applied the above ML and feature selection pipeline to the mRNA features to determine whether a gene-only biomarker panel could distinguish GC from normal samples.

7.2.4.3 Salivary exRNA Biomarkers

Next, we expanded our analysis to include other classes of RNA commonly used as biomarkers in addition to mRNAs, such as miRNAs, piRNAs, circRNAs, and tRNAs. These RNA categories together yielded a total of 362 significantly different salivary exRNA features. We further applied our ML approach on this comprehensive feature set to identify the most relevant biomarkers for classifying GC vs. normal samples.

A. Primary Biomarkers

We identified an initial set of top-performing biomarkers using the feature selection approach that gave the best classification performance - IFS. This yielded a primary set (P) of biomarkers. Feature selection methods are designed to capture the most informative features while minimizing redundancy. In cases where features are highly correlated, many selection algorithms prefer to retain just one representative feature to prevent multicollinearity and overlap in information. This challenge is particularly evident in biological datasets, where biomarkers often show co-expression or functional relationships. As a result, different feature selection strategies may prioritize slightly different biomarker sets based on their specific ranking approaches. Hence, multiple biomarkers may be relevant in the classification of a disease, but only a subset is selected using feature selection methods, possibly due to underlying correlations.

B. Secondary Biomarkers

To obtain additional relevant biomarker sets, we performed an iterative feature selection process. We subtracted the set of identified primary biomarkers (P) from the total set of significant salivary exRNAs (A = 362), leaving a reduced feature pool (A–P). We then applied the feature selection methods on the set (A–P) and obtained a first secondary set of biomarkers, denoted as S1. This S1 set consisted of 11 salivary exRNAs. We repeated this process again: removing the P and S1 features from A, and then performing feature selection on the remaining features (A–P–S1). This led to the identification of a second secondary set of biomarkers, denoted as S2.

C. Correlation Among Biomarkers

To evaluate the relationships among the identified biomarkers, we computed Pearson's correlation coefficients for all pairwise combinations within and across the biomarker sets P, S1, and S2. This analysis aimed to determine whether the expression patterns of biomarkers in these sets were significantly correlated with each other, thereby providing insights into potential co-regulation or shared biological relevance among the identified candidates.

7.2.4.4 Cross Validation

To minimize bias and prevent overfitting, a five-fold cross-validation strategy was employed. Initially, the dataset was split into training and independent validation sets in an 80:20 ratio. The training set was further partitioned into five subsets for cross-validation. This method has been previously described in Chapter 4.

7.2.4.5 Performance Metrics

The evaluation of the prediction models was based on both threshold-dependent and threshold-independent parameters. Threshold-dependent metrics included sensitivity (which measures the true positive rate, i.e., how well the model identifies GC patients), specificity (the true negative rate for correctly predicting normal subjects), accuracy (the proportion of GC and normal subjects that were correctly classified), and MCC, which reflects the overall agreement between predicted and observed values. As a threshold-independent metric, we used AUROC, which indicates the model's capacity to differentiate between the two classes across all possible thresholds.

7.3 Results

7.3.1 Data Pre-processing

The dataset initially included 19,704 mRNAs, 55,104 miscellaneous RNAs (e.g., lincRNAs, snRNAs, etc.), 1,583 miRNAs, 2,281 piRNAs, 26 tRNAs, and 245 circRNAs. After applying the filtering criteria, a total of 33,212 exRNA features remained, comprising 11,532 mRNAs, 516 miRNAs, 110 piRNAs, 26 tRNAs, 10 circRNAs, and 21,018 miscellaneous RNAs. After applying the Mann-Whitney U test on the data, we were left with 1,027 exRNAs that were significantly different between the two groups (p-value < 0.05). The distribution of these 1,027 significant exRNAs across RNA categories is shown in Figure 7.3, and the total number of upregulated and downregulated exRNAs in each category is given in Figure 7.4. Out of these significant features, the most common RNA classes used for biomarkers (mRNAs, miRNAs, circRNAs, piRNAs, and tRNAs) were selected for further analysis, which comprised a total of 362 exRNAs.

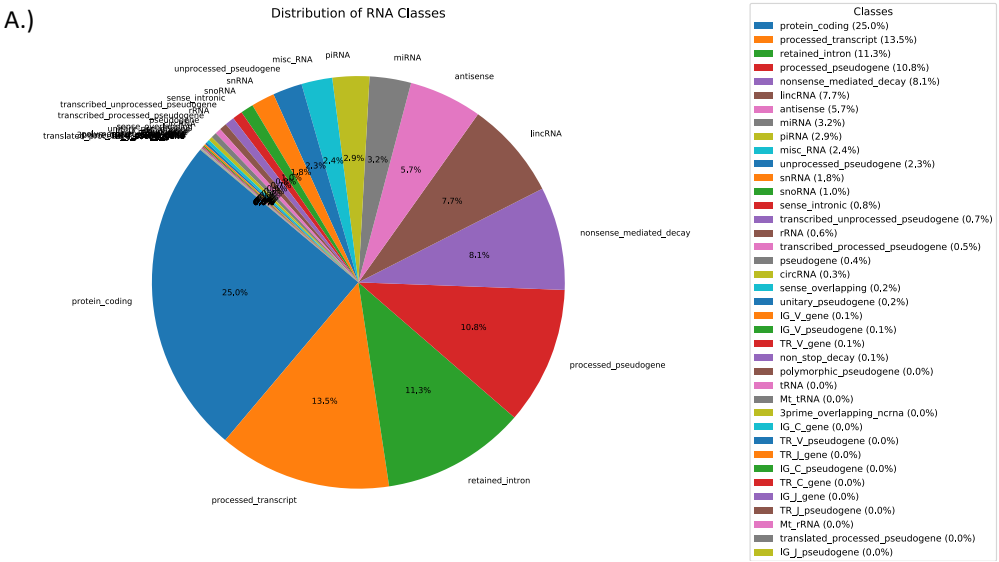
7.3.2 Threshold-based Approach

The top individual biomarker using the threshold-based classification approach was the protein-coding gene GSDMA, which achieved about 64.14% accuracy in classifying GC vs. normal samples (this gene was downregulated in GC samples). The next best performer was SMR3B (~62.63% accuracy, upregulated in GC). Several other exRNAs showed accuracies of approximately 61.6%—notably NPTXR and TBCD (both upregulated in GC), and KCNC4 and KRT26 (both downregulated in GC). The results for the top 15 biomarkers identified by this threshold method are shown in Figure 7.5.

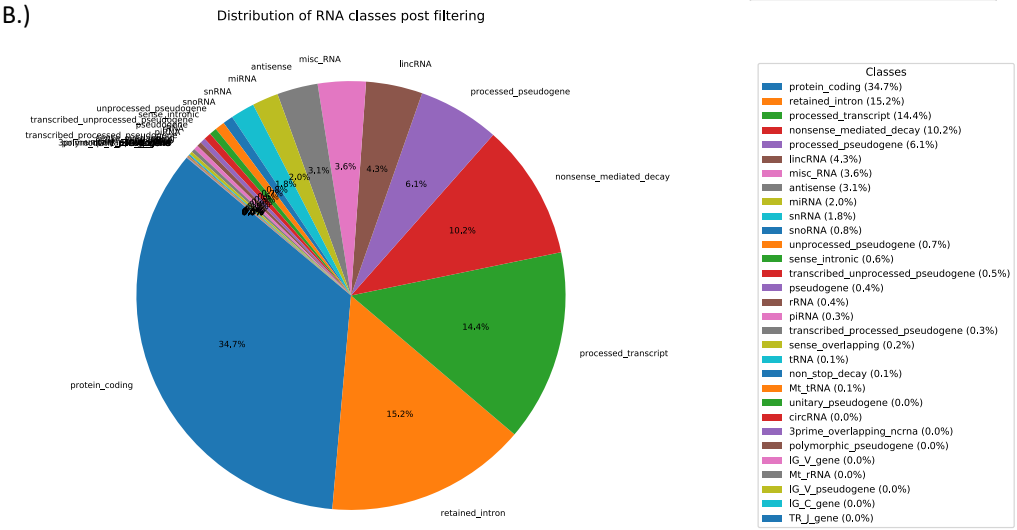
7.3.3 Machine Learning-based Approach

We applied a suite of ML classifiers (as described in Methods) to build predictive models for GC classification. Two main analyses were performed: one using only gene (mRNA) features, and another incorporating multiple exRNA categories.

A.)


Distribution of RNA Classes
protein_coding
processed_transcript
retained_intron
processed_pseudogene
nonsense_mediated_decay
lincRNA
antisense
miRNA
piRNA
misc_RNA
25.0%
13.5%
11.3%
10.8%
8.1%
7.7%
5.7%
3.2%
2.9%
2.4%
Classes
protein_coding (25.0%)
processed_transcript (13.5%)
retained_intron (11.3%)
processed_pseudogene (10.8%)
nonsense_mediated_decay (8.1%)
lincRNA (7.7%)
antisense (5.7%)
miRNA (3.2%)
piRNA (2.9%)
misc_RNA (2.4%)
unprocessed_pseudogene (2.3%)
snRNA (1.8%)
snoRNA (1.0%)
sense_intronic (0.8%)
transcribed_unprocessed_pseudogene (0.7%)
rRNA (0.6%)
transcribed_processed_pseudogene (0.5%)
pseudogene (0.4%)
circRNA (0.3%)
sense_overlapping (0.2%)
unitary_pseudogene (0.2%)
IG_V_gene (0.1%)
IG_V_pseudogene (0.1%)
TR_V_gene (0.1%)
non_stop_decay (0.1%)
polymorphic_pseudogene (0.0%)
tRNA (0.0%)
Mt_tRNA (0.0%)
3prime_overlapping_ncrna (0.0%)
IG_C_gene (0.0%)
TR_V_pseudogene (0.0%)
TR_J_gene (0.0%)
IG_C_pseudogene (0.0%)
TR_C_gene (0.0%)
IG_J_gene (0.0%)
TR_J_pseudogene (0.0%)
Mt_rRNA (0.0%)
translated_processed_pseudogene (0.0%)
IG_J_pseudogene (0.0%)

B.)


Distribution of RNA classes post filtering
protein_coding
processed_transcript
retained_intron
nonsense_mediated_decay
processed_pseudogene
lincRNA
misc_RNA
antisense
34.7%
15.2%
14.4%
10.2%
6.1%
4.3%
3.6%
3.1%
2.0%
Classes
protein_coding (34.7%)
retained_intron (15.2%)
processed_transcript (14.4%)
nonsense_mediated_decay (10.2%)
processed_pseudogene (6.1%)
lincRNA (4.3%)
misc_RNA (3.6%)
antisense (3.1%)
miRNA (2.0%)
snRNA (1.8%)
snoRNA (0.8%)
unprocessed_pseudogene (0.7%)
sense_intronic (0.6%)
transcribed_unprocessed_pseudogene (0.5%)
pseudogene (0.4%)
rRNA (0.4%)
piRNA (0.3%)
transcribed_processed_pseudogene (0.3%)
sense_overlapping (0.2%)
tRNA (0.1%)
non_stop_decay (0.1%)
Mt_tRNA (0.1%)
unitary_pseudogene (0.0%)
circRNA (0.0%)
3prime_overlapping_ncrna (0.0%)
polymorphic_pseudogene (0.0%)
IG_V_gene (0.0%)
Mt_rRNA (0.0%)
IG_V_pseudogene (0.0%)
IG_C_gene (0.0%)
TR_J_gene (0.0%)

C.)

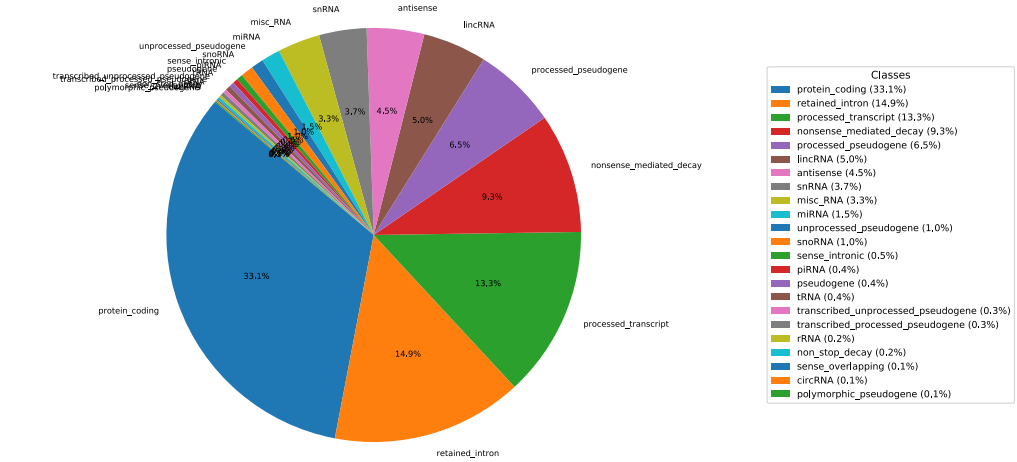

Distribution of differentially expressed RNA (p<0.05)
protein_coding
processed_transcript
retained_intron
nonsense_mediated_decay
processed_pseudogene
lincRNA
antisense
snRNA
misc_RNA
33.1%
14.9%
13.3%
9.3%
6.5%
5.0%
4.5%
3.7%
3.3%
Classes
protein_coding (33.1%)
retained_intron (14.9%)
processed_transcript (13.3%)
nonsense_mediated_decay (9.3%)
processed_pseudogene (6.5%)
lincRNA (5.0%)
antisense (4.5%)
snRNA (3.7%)
misc_RNA (3.3%)
miRNA (1.5%)
unprocessed_pseudogene (1.0%)
snoRNA (1.0%)
sense_intronic (0.5%)
piRNA (0.4%)
pseudogene (0.4%)
tRNA (0.4%)
transcribed_unprocessed_pseudogene (0.3%)
transcribed_processed_pseudogene (0.3%)
rRNA (0.2%)
non_stop_decay (0.2%)
sense_overlapping (0.1%)
circRNA (0.1%)
polymorphic_pseudogene (0.1%)

Figure 7.3: Distribution of different exRNA classes in A.) the whole dataset, B.) After filtering step: removing exRNA with >80% data as zeroes, C.) Significantly different exRNA in Normal vs GC found using Mann-Whitney U test (p-value<0.05)

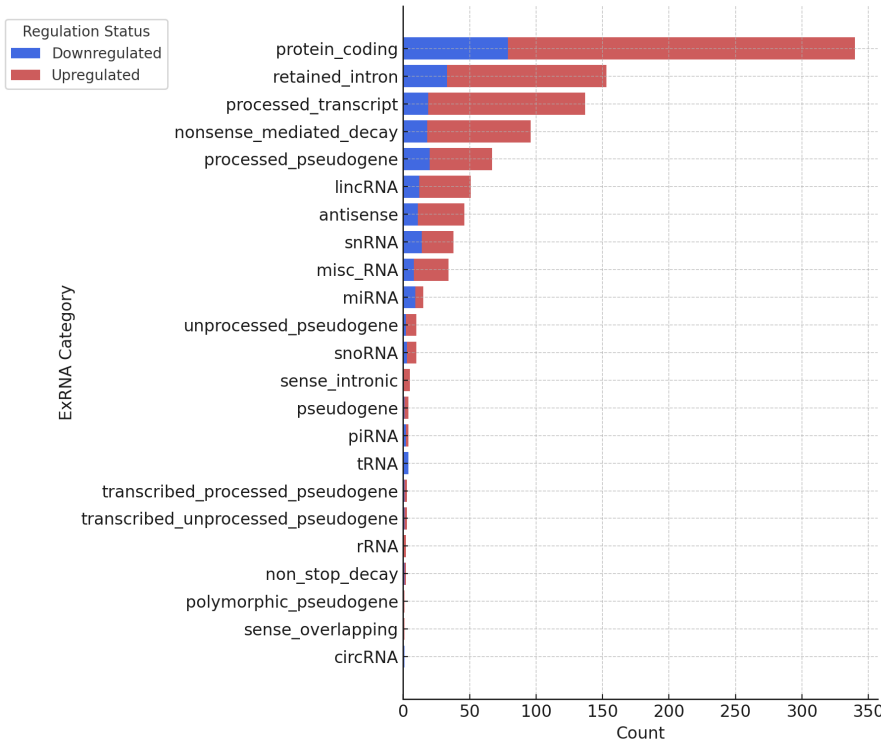


Figure 7.4: Number of upregulated and downregulated exRNA in each category

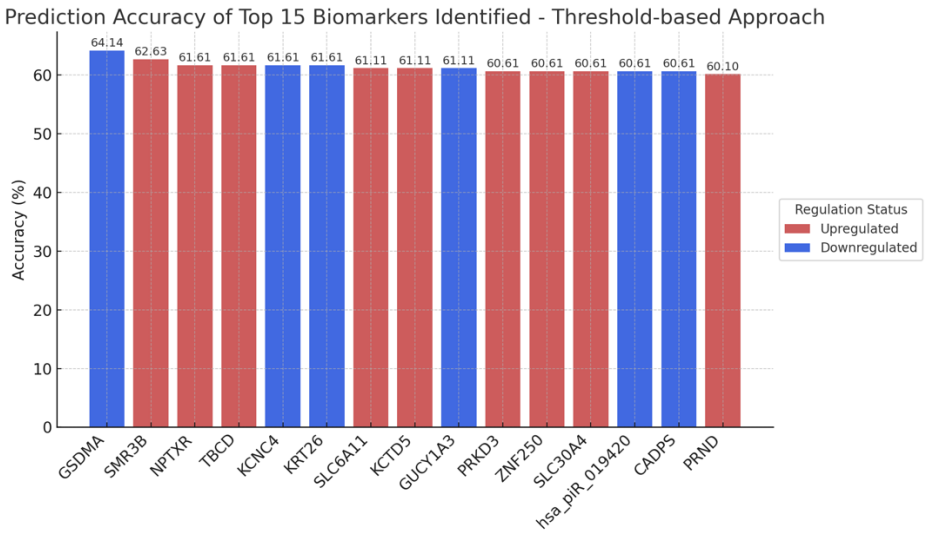


Figure 7.5: Prediction Accuracies of Top 15 biomarkers identified using threshold-based approach

7.3.3.1 Gene-based Biomarkers

To identify gene-based biomarkers, we selected 340 significantly different genes (mRNAs) from the initial pool of salivary exRNAs. Machine learning models trained on this 340-gene set achieved a maximum independent validation AUC of 0.975 (obtained with a Logistic Regression model). Next, we applied feature selection to this gene set to identify a smaller signature of genes with high discriminative power. This process yielded nine gene biomarkers (GSDMA, CCDC141, HSP90B1, SLC30A4, ATP8B3, ARHGAP23, NPTXR, WRB, and SMR3B), which resulted in an AUC of 0.893 on the independent validation set using a voting classifier. The performance of models using all 340 gene features and the selected 9-gene subset is summarized in Table 7.1.

Table 7.1: Results for gene-based biomarker sets

All significant genes (n=340)											
	Training Set						**Independent Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**
ET	0.49	0.73	0.70	71.52	0.431	0.825	0.80	0.65	72.50	0.455	0.823
RF	0.51	0.77	0.79	77.85	0.557	0.846	0.80	0.70	75.00	0.503	0.845
LR	0.51	0.91	0.91	91.14	0.823	0.977	0.95	0.85	90.00	0.804	0.975
XGB	0.48	0.72	0.75	73.42	0.468	0.784	0.80	0.75	77.50	0.551	0.798
KNN	0.43	0.71	0.75	72.78	0.456	0.787	0.80	0.70	75.00	0.503	0.794
SVC	0.50	0.87	0.89	87.98	0.759	0.962	0.95	0.80	87.50	0.759	0.965
Feature Selection (9 biomarkers)											
		Training Set					**Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**
ET	0.49	0.83	0.80	81.65	0.633	0.861	0.75	0.80	77.50	0.551	0.860
RF	0.50	0.73	0.73	72.78	0.456	0.829	0.75	0.65	70.00	0.402	0.820
LR	0.50	0.79	0.73	75.95	0.521	0.848	0.80	0.85	82.50	0.651	0.870
XGB	0.48	0.72	0.71	71.52	0.43	0.814	0.75	0.70	72.50	0.451	0.810
KNN	0.49	0.80	0.80	80.22	0.636	0.880	0.80	0.85	82.50	0.651	0.885
SVC	0.51	0.81	0.80	80.38	0.608	0.860	0.70	0.80	75.00	0.503	0.853
Ensemble Classifiers for Feature Selection Set (n=9)											
		Training Set					**Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**

SC (Base: SVC, ET, LR, KNN, GNB; Meta: KNN)	0.48	0.81	0.80	80.38	0.608	0.883	0.90	0.80	85.00	0.704	0.893
VC (SVC, ET, KNN, GNB)	0.53	0.81	0.81	81.01	0.620	0.871	0.85	0.80	82.50	0.651	0.870

7.3.3.2 Salivary exRNA Biomarkers

Next, we built models using the combined set of 362 significant exRNAs spanning multiple RNA classes (mRNAs, miRNAs, piRNAs, circRNAs, and tRNAs). Using all 362 features, the Logistic Regression classifier achieved the best performance, with an AUROC of approximately 0.97 on the training set and 0.965 on the independent validation set (Table 7.3). We applied feature selection methods to identify a smaller biomarker panel from these 362 features.

A) Primary Biomarkers

It was observed that model performance gradually decreased as the number of biomarkers was reduced, but it remained high even with a substantially trimmed feature set. Using an iterative feature selection (IFS) method, we identified a subset of 24 features (23 mRNAs and 1 tRNA; denoted as set B) that achieved an AUROC of 0.937 on the independent validation set, comparable to the performance with all 362 features. We further applied feature selection techniques to identify the top-performing biomarkers from this set, which helped us pinpoint the top 8 biomarkers (CCDC141, CDHR3, GSDMA, STARD13, TBCD, WRB, ARHGAP23, BX842679.1). This 8-mRNA panel achieved an AUC of 0.873 on the independent validation set with a single SVC model, which increased to 0.905 when an ensemble method (stacking classifier) was applied. We named this top-performing set the primary biomarker set (P). Notably, although this primary set consists only of mRNAs, it exceeded the performance of the larger 24-biomarker set. The detailed results for the complete feature set (Set A, n = 362), the 24-feature set (Set B, n = 24), and the 8-feature primary set (Set P, n = 8) are given in Table 7.2, and the corresponding AUROC curves are shown in Figure 7.5.

Table 7.2: The results on various feature sets after performing feature selection – a) n = 362, b) n = 24, and c) n = 8; here, Thr: threshold, Spec: specificity, Sens: sensitivity

All selected exRNA (Set A, n=362)											
		Training Set					Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**
ET	0.51	0.74	0.73	73.42	0.469	0.843	0.80	0.70	75.00	0.503	0.849
RF	0.51	0.77	0.76	76.58	0.532	0.850	0.80	0.75	77.50	0.551	0.850
LR	0.48	0.91	0.91	91.14	0.823	0.970	0.95	0.90	92.50	0.851	0.965
XGB	0.54	0.83	0.84	83.54	0.671	0.890	0.85	0.80	82.50	0.651	0.860
KNN	0.41	0.82	0.59	70.25	0.419	0.784	0.90	0.60	75.00	0.524	0.779
SVC	0.53	0.88	0.89	88.61	0.772	0.942	0.95	0.80	87.50	0.759	0.955
Feature Selection (Set B, n=24)											
		Training Set					Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**
ET	0.51	0.81	0.84	82.28	0.646	0.883	0.80	0.85	82.50	0.651	0.895
RF	0.50	0.76	0.75	75.32	0.506	0.864	0.80	0.75	77.50	0.551	0.842
LR	0.51	0.93	0.90	90.38	0.770	0.930	0.90	0.95	92.50	0.831	0.937
XGB	0.54	0.77	0.78	77.22	0.544	0.845	0.70	0.70	70.00	0.400	0.835
KNN	0.56	0.77	0.80	78.48	0.570	0.854	0.75	0.80	77.50	0.551	0.864
SVC	0.52	0.90	0.91	91.00	0.801	0.935	0.90	0.90	90.00	0.800	0.930
Feature Selection (Set P, n=8)											
		Training Set					Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**
ET	0.52	0.77	0.78	77.22	0.544	0.834	0.65	0.85	75.00	0.510	0.830
RF	0.50	0.78	0.78	77.85	0.557	0.823	0.75	0.80	77.50	0.551	0.833
LR	0.50	0.76	0.78	76.58	0.532	0.838	0.70	0.85	77.50	0.556	0.825
XGB	0.47	0.72	0.70	70.89	0.418	0.770	0.70	0.70	70.00	0.400	0.788
KNN	0.34	0.83	0.79	81.01	0.621	0.812	0.85	0.75	80.00	0.603	0.825
SVC	0.51	0.81	0.81	80.85	0.657	0.864	0.70	0.95	82.50	0.671	0.873
Ensemble Classifiers for Feature Selection Set P (n=8)											
		Training Set					Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc (%)**	**MCC**	**AUC**
SC (Base: SVC, RF, LR, KNN, MLP, GB, GNB; Meta: SVC)	0.54	0.87	0.86	84.48	0.706	0.891	0.87	0.90	89.00	0.77	0.905

VC (SVC, LR, AB, ET, KNN, GNB)	0.50	0.79	0.80	79.75	0.595	0.863	0.85	0.90	87.50	0.751	0.880

B) Secondary Biomarkers

After identifying the primary set, we also evaluated additional secondary biomarker sets derived as described earlier in the methods section. The best model based on the first secondary set (S1) achieved an independent validation AUROC of ~0.788, and for the second secondary set (S2), the highest AUROC was ~0.798. A combination of the primary and secondary biomarker sets (P + S1 + S2) yielded a maximum AUROC of 0.868 on the independent validation set (using a Logistic Regression classifier). However, these secondary panels, either alone or in combination with the primary panel, did not surpass the performance of the primary set alone (Table 7.3; see also Figure 7.6 for AUROC curves).

Table 7.3: The results for sets S1 and S2 (Secondary feature sets) and sets S1 and S2 combined with set P (Primary features)

Secondary Biomarkers (Set S1, n=11)											
		Training Set					**Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
ET	0.52	0.68	0.71	69.62	0.392	0.771	0.60	0.80	70.00	0.408	0.768
RF	0.53	0.70	0.71	70.35	0.397	0.777	0.75	0.70	72.50	0.451	0.788
LR	0.51	0.69	0.69	68.99	0.380	0.782	0.70	0.66	67.50	0.351	0.752
XGB	0.53	0.65	0.66	65.82	0.316	0.746	0.75	0.55	65.00	0.306	0.742
KNN	0.43	0.79	0.63	70.89	0.426	0.742	0.65	0.60	62.50	0.250	0.743
SVC	0.48	0.69	0.69	68.99	0.380	0.759	0.70	0.65	67.50	0.350	0.750
Secondary Biomarkers (Set S2, n=11)											
		Training Set					**Validation Set**				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
ET	0.5	0.68	0.68	67.72	0.354	0.797	0.75	0.70	72.50	0.451	0.798
RF	0.49	0.68	0.68	67.72	0.354	0.785	0.70	0.65	67.50	0.350	0.788
LR	0.48	0.69	0.69	68.99	0.380	0.759	0.65	0.65	65.00	0.300	0.735
XGB	0.47	0.69	0.69	68.99	0.380	0.748	0.80	0.70	75.00	0.503	0.744
KNN	0.51	0.68	0.73	70.25	0.405	0.741	0.75	0.55	65.00	0.306	0.734
SVC	0.49	0.72	0.70	70.89	0.418	0.781	0.75	0.60	67.50	0.354	0.771
Primary + Secondary Biomarkers (Sets P+S1+S2, n=30)											

		Training Set					Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
ET	0.5	0.81	0.83	81.65	0.633	0.893	0.85	0.75	80.00	0.603	0.857
RF	0.53	0.76	0.76	75.95	0.519	0.850	0.75	0.80	77.50	0.551	0.863
LR	0.51	0.85	0.85	84.81	0.696	0.917	0.85	0.70	77.50	0.556	0.868
XGB	0.48	0.79	0.80	79.75	0.595	0.851	0.70	0.80	75.00	0.503	0.830
KNN	0.61	0.71	0.81	75.95	0.521	0.821	0.60	0.95	77.50	0.587	0.826
SVC	0.48	0.78	0.79	78.48	0.570	0.869	0.75	0.80	77.50	0.551	0.859

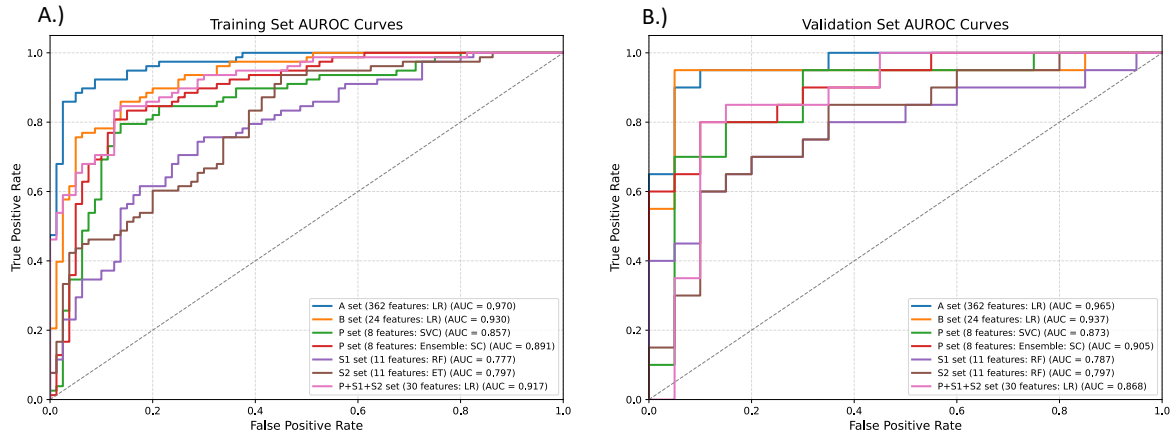


Figure 7.6: The AUROC plots for Training and Independent Validation sets for different feature sets. Here, Set A: All significant biomarkers, Set B: 24 biomarkers after feature selection, Set P: 8 biomarkers after feature selection, also named as the primary set of biomarkers in this study, Set S1: First set of secondary biomarkers, Set S2: Second set of secondary biomarkers.

C) Correlation Among Biomarkers

We calculated Pearson's correlation coefficients among the biomarkers in sets P, S1, and S2. The top 3 highest correlated biomarker pairs for each comparison are given in Table 7.4. The most strongly correlated pair for the primary and first secondary sets was GSDMA (from set P) with GUCY1A3 (from S1), with a Pearson correlation of 0.529. For sets P and S2, the highest correlation was between ARHGAP23 and LPPR5 (r = 0.411). For the two secondary sets S1 and S2, the most correlated pair was KRT26 and ZNF277 (r = 0.365). Overall, while some biomarkers in the primary and secondary panels were correlated with each other, each set largely contained unique biomarkers with relatively low inter-set correlations.

Table 7.4: Pearson correlation for biomarkers present in sets P (Primary Set), S1 (Secondary Set 1), and S2 (Secondary Set 2)

Top 3 correlated biomarkers in sets P and S1		
Biomarker (P)	**Biomarker (S1)**	**correlation**
GSDMA	GUCY1A3	0.529
GSDMA	KRT26	0.404
GSDMA	hsa_circ_000353	0.327
Top 3 correlated biomarkers in sets P and S2		
Biomarker (P)	**Biomarker (S2)**	**correlation**
ARHGAP23	LPPR5	0.411
TBCD	EXOSC6	0.403
TBCD	ZMAT5	0.341
Top 3 correlated biomarkers in sets S1 and S2		
Biomarker (S1)	**Biomarker (S2)**	**correlation**
KRT26	ZNF277	0.365
GUCY1A3	ZNF277	0.287
KCTD5	EXOSC6	0.286

7.3.4 Comparison with other studies

It is essential to compare our findings against those of previously published studies. Earlier research focused on identifying salivary biomarkers for gastric cancer (GC) (Kaczor-Urbanowicz et al., 2022; F. Li, Yoshizawa, et al., 2018). They proposed a five-biomarker panel (PPL, SPINK7, SEMA4B, miR-301a-3p, miR-140-5p) to distinguish GC vs Normal. These biomarkers and demographic factors reported the highest AUC of 0.87. However, when we applied these biomarkers to our dataset for GC vs Normal classification, the highest AUROC achieved on the independent validation set was only 0.563. Furthermore, when these biomarkers were integrated into our primary biomarker set (P), the AUROC on the independent validation dataset declined from 0.905 to 0.830. These findings indicate that the salivary

exRNA signature identified in our study outperformed the previously reported biomarker panel on our cohort (Tables 7.5 and 7.6).

Table 7.5: Comparison of the performance of biomarkers found in our study with previous studies

Study	Technique	Biomarkers	AUROC
Arora et al.	RNASeq (100 Normal, 98 GC)	8 mRNA (GSDMA, CCDC141, TBCD, STARD13, WRB, ARHGAP23, CDHR3, BX842679.1)	0.905
Kaczor-Urbanowicz et al.	Microarray (50 Normal, 50 GC)	Demographics	0.68
		Demographics + 2 miRNA (miR-140-5p and miR-301a-3p)	0.75
		Demographics + 3 mRNA (SPINK7, PPL, SEMA4B) + 2 miRNA (miR-140-5p and miR-301a-3p)	0.78
Li et al.	Microarray (31 Normal, 63 GC for mRNA; 10 Normal, 10 GC for miRNA)	Demographics	0.69
		3 mRNA (SPINK7, PPL, SEMA4B) + 2 miRNA (miR-140-5p and miR-301a-3p)	0.81
		Demographics 3 mRNA (SPINK7, PPL, SEMA4B) + 2 miRNA (miR-140-5p and miR-301a-3p)	0.87

Table 7.6: The results for ML models developed on a) a 5 biomarker panel discovered in Li. et al. study, b) 13 biomarkers - 5 biomarkers identified in Li et al. study and 8 primary biomarkers identified in this study

Li et al. biomarkers (n=5)											
		Training Set					Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
ET	0.48	0.49	0.45	46.84	-0.063	0.455	0.35	0.55	45.00	-0.102	0.453
RF	0.5	0.51	0.51	51.27	0.025	0.499	0.45	0.75	60.00	0.210	0.510
LR	0.48	0.55	0.54	53.27	0.032	0.542	0.60	0.50	55.00	0.101	0.563
XGB	0.48	0.47	0.50	48.73	-0.026	0.438	0.35	0.45	40.00	-0.201	0.464
KNN	0.01	0.45	0.51	48.10	-0.039	0.481	0.50	0.50	50.00	0.000	0.500
SVC	0.49	0.49	0.51	50.00	0.000	0.479	0.75	0.20	47.50	-0.060	0.467
Primary biomarkers + Li et al. biomarkers (n=13)											
		Training Set					Validation Set				
Model	**Thr**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**	**Sens**	**Spec**	**Acc**	**MCC**	**AUC**
ET	0.51	0.78	0.78	77.85	0.557	0.809	0.65	0.85	75.00	0.510	0.808
RF	0.49	0.72	0.71	71.52	0.430	0.789	0.70	0.80	75.00	0.503	0.779
LR	0.49	0.74	0.70	72.15	0.444	0.808	0.75	0.70	72.50	0.451	0.803
XGB	0.49	0.65	0.66	65.82	0.316	0.736	0.65	0.75	70.00	0.402	0.742
KNN	0.41	0.81	0.66	73.42	0.475	0.788	0.70	0.75	72.50	0.451	0.794
SVC	0.5	0.77	0.76	76.58	0.532	0.828	0.70	0.85	77.50	0.556	0.830

7.4 Discussions

Gastric cancer (GC) continues to be a major global health challenge because of its high death rate and late detection. Traditional diagnostic methods like endoscopy and tissue biopsy are invasive, expensive, and not ideal for regular screening. In recent years, saliva has gained attention as a promising option for non-invasive biomarker discovery. It carries a wide variety of extracellular molecules, including RNA, that can capture changes happening throughout the body (Arora, Kaur, et al., 2023).

In this study, we focused on finding salivary extracellular RNA (exRNA) biomarkers that can distinguish gastric cancer (GC) patients from healthy individuals. We started by using the Mann-Whitney U test to identify salivary exRNAs with p-values<0.05, which lead to ~99,000 salivary exRNA from 98 GC patients and 100 normal controls. Next, we carried out a

threshold-based classification, where we used the midpoint between the mean expression levels of the two groups as the threshold. Through this method, we found that GSDMA (downregulated in GC) could distinguish GC from normal samples with an accuracy of 64.14%. This was followed by SMR3B (upregulated) with 62.63% accuracy, and then NPTXR (upregulated), KCNC4 (downregulated), TBCD (upregulated), and KRT26 (downregulated), each achieving an accuracy of 61.61%.

To further improve the prediction of GC vs. Normal samples, we applied machine learning (ML) models to the data. We split the dataset into 80% for training and 20% for independent validation. First, we built a classification model using only the gene features (n = 340) from the significant salivary exRNAs. The Logistic Regression (LR) model trained on these genes achieved an AUC of 0.975 on the independent validation set. After applying feature selection techniques, we narrowed down the features to 9 key gene biomarkers (ARHGAP23, ATP8B3, CCDC141, GSDMA, HSP90B1, NPTXR, SLC30A4, SMR3B, WRB), which reached the highest AUC of 0.893 using a voting classifier.

Furthermore, we wanted to find additional relevant exRNAs beyond mRNAs. We expanded our analysis to include other common RNA types used as biomarkers such as circRNA, miRNA, piRNA, and tRNA, along with mRNA. This combined set of 362 salivary exRNAs achieved an AUC of 0.965 and an MCC of 0.851 using a Logistic Regression model. However, we aimed to identify a smaller set of top biomarkers that could differentiate GC from normal samples with high accuracy. By applying feature selection methods, we found a panel of 24 features (23 genes and 1 tRNA) that performed comparably, with an AUROC of 0.937 and an MCC of 0.851 on the independent validation set. We further reduced the feature set down to an optimal primary set of 8 mRNAs (CCDC141, CDHR3, GSDMA, TBCD, STARD13, ARHGAP23, BX842679.1, WRB) using feature selection techniques. This primary set of biomarkers achieved the highest AUROC of 0.905 and MCC of 0.770 on the independent validation set using an ensemble stacking classifier, surpassing the performance of the larger mRNA-only set.

Feature selection methods are designed to improve model performance by keeping the most important features while removing redundant ones. When features are highly correlated, many methods select just one to avoid multicollinearity and overlapping information. However, in biomarker discovery, removing one feature from a group of highly correlated biomarkers can be risky as each may have unique biological importance. To address this, we worked on

identifying multiple meaningful sets of biomarkers. We called the first set of 8 mRNAs our primary biomarker set (P). To find more relevant biomarkers, we first removed this primary set from the original 362 biomarkers (set A) and then performed feature selection again, which led to the first secondary set (S1) with 11 salivary exRNAs (ALX1, ATR, CADPS, KRT26, KPTN, KCTD5, GUCY1A3, hsa_circ_000353, HIST1H3A, TMEM98, ZNF688). We repeated this process to identify a second secondary set (S2), which also contained 11 salivary exRNAs (CALB2, ATP8B3, KIAA2022, MRPL51, MDH1B, PABPC3, PRKD3, ZNF277, EXOSC6, LPPR5, ZMAT5). We also checked how correlated the biomarkers were within each set. In sets P and S1, the strongest correlation was between GSDMA and GUCY1A3 (Pearson's r = 0.529). For sets P and S2, the highest correlation was between ARHGAP23 and LPPR5 (r = 0.411), and between sets S1 and S2, the most correlated pair was KRT26 and ZNF277 (r = 0.365).

We compared the performance of our machine learning (ML) models with previous studies and found that our set of biomarkers performed better than earlier reported panels. Previous studies had achieved a highest AUC of 0.87 by combining three mRNAs and two miRNAs with demographic information. However, when we tested those biomarkers on our dataset for GC vs. Normal classification, the highest AUROC achieved was only 0.563 on the independent validation set. Moreover, when these biomarkers were added to our primary biomarker set (P), the AUROC dropped from 0.905 to 0.830 on the independent validation set.

The primary set of biomarkers identified includes CCDC141 (downregulated), GSDMA (downregulated), CDHR3 (upregulated), TBCD (upregulated), ARHGAP23 (upregulated), STARD13 (downregulated), WRB (downregulated), and BX842679.1 (downregulated). These biomarkers have shown links to Gastric Cancer in previous studies. CCDC141 has been shown to be associated with GC and involved in metastasis and other cellular processes related to cancer progression (Tanikawa et al., 2018). GSDMA is shown to regulate apoptosis in gastric epithelial cells and is often silenced in GC (Saeki et al., 2000, 2009). CDHR3 is involved in cell adhesion and has been reported to be dysregulated in GC (Hao et al., 2019; Qiu et al., 2020). STARD13 functions as a tumor suppressor and is a known direct target of miR-125b in GC (Chang et al., 2016). ARHGAP23 is often seen to be upregulated in GC. It also regulates the actin cytoskeleton and cell motility (X. Liu et al., 2023). However, for BX842679.1, there is limited information regarding its role in GC, but it has been reported to be linked with myeloid leukemia (Mulindwa et al., 2020). TBCD is associated with microtubule dynamics

and cell proliferation, but no direct association with GC has been reported yet (Fanarraga et al., 2010).

A major limitation of this study is that it is based mainly on computational and bioinformatics analysis of the existing datasets. Although the results offer valuable insights into potential biomarkers, they should be considered preliminary. Further clinical studies and wet-lab experiments are necessary to confirm their diagnostic and prognostic value.

Chapter 8

Summary

Non-invasive diagnostics and therapeutics have significantly transformed medical practices by providing patient-friendly, easily accessible, and pain-free alternatives to conventional invasive methods. Saliva as a readily obtainable biofluid, has emerged as an excellent medium for such applications. It enables rapid and stress-free sample collection and contains biomarkers reflective of systemic health. One of the pathways through which biomarkers reach biofluids including saliva is via exosomes. Exosomes are tiny extracellular vesicles secreted by almost all cell types into saliva and other biofluids. They carry specific biological molecules such as different types of RNAs, proteins, and lipids. Due to their unique cargo reflective of their cells of origin, exosomes hold promise in disease detection and targeted therapeutic delivery.

Chapter 1 highlights the increasing demand for accessible and non-invasive methods for diagnostics and therapeutics. It emphasizes saliva's promising role as a convenient and non-invasive sample source, stressing the importance of developing saliva-based diagnostic tools. Additionally, it explores the biological mechanisms by which biomarkers are delivered into saliva, particularly focusing on exosome-mediated pathways. The chapter examines exosome biogenesis and their molecular constituents, pointing out the existing gaps in understanding how biomolecules are selectively incorporated into exosomes. Lastly, it provides an overview of the thesis structure and outlines the main objectives of this research.

Chapter 2 presents a thorough literature review on current methodologies and tools for discovering non-invasive biomarkers, emphasizing saliva-based and exosome-based approaches. This chapter discusses the clinical significance of biomarkers found in saliva and highlights known salivary markers approved by the FDA and their existing databases. The chapter subsequently addresses the diagnostic and therapeutic potential of exosomes. It also reviews available databases and computational tools designed for predicting exosomal biomarkers.

Chapter 3 discusses the development of a database for salivary biomarkers found in humans. Saliva has enormous potential for early diagnosis and patient prognosis because it is a non-invasive diagnostic fluid. Salivary biomarker data is widely dispersed throughout a number of sources and academic papers. It's crucial to compile all data on salivary biomarkers onto a single platform to expedite the development of non-invasive diagnosis and prognosis techniques. We gathered extensive data on five categories of salivary biomarkers, including proteins, metabolites, microorganisms, miRNA, and genes from a range of sources, including SalivaTecDB, the Human Metabolome Database, and PubMed. SalivaDB contains 15821

entries for 48 disease categories and 201 diseases. These entries are divided into five classes based on the type of biomolecule: 6067 entries for proteins, 3987 for metabolites, 2909 for microbes, 2272 for miRNAs, and 586 for genes. We created a web-based interface that offers a variety of options, such as browsing, basic keyword search, advanced search, and similarity search. A web server for this database has been built to provide services to the scientific community, which is available at https://webs.iiitd.edu.in/raghava/salivadb/.

Chapter 4 underlines the development of a computational tool to predict exosomal proteins. It is vital to develop non-invasive diagnostic methods and treatments to avoid subjecting patients to painful procedures. Exosomal proteins represent promising biomarkers for advancing such approaches, since they can be obtained from body fluids without invasive surgery. In this study, we compiled a non-redundant dataset of human proteins consisting of 2,831 exosomal and 2,831 non-exosomal proteins, with no two sequences sharing >40% identity. A conventional similarity-based approach (BLAST) proved ineffective due to the low sequence homology in this dataset. We therefore developed machine-learning models using protein sequence features (composition and evolutionary profiles), which achieved AUROC of about 0.73. Further analysis revealed that exosomal proteins contain distinctive sequence motifs that can aid in their identification. Using this insight, we devised a hybrid predictor that combines motif-based scores with the machine-learning model. This hybrid approach achieved an AUROC of 0.85 and MCC of 0.56 on an independent test set. The hybrid model outperformed existing prediction methods when evaluated on an independent validation set. Finally, we implemented our hybrid method in a standalone software tool and webserver, ExoProPred (https://webs.iiitd.edu.in/raghava/exopropred/), to enable researchers to predict exosomal proteins and examine the functional motifs they contain.

Chapter 5 underscores the development of an *in silico* tool for predicting exosomal miRNA. In this study, we examined the features of exosomal microRNAs (miRNAs) to identify biomarkers for liquid biopsy. We gathered 956 exosomal and 956 non-exosomal miRNA sequences from miRBase and RNALocate databases to construct predictive models. Our preliminary analyses indicated some nucleotide preferences at specific positions within exosomal miRNAs. We applied multiple prediction approaches, including alignment-based methods, artificial intelligence (AI)-based approaches, and ensemble methods. The alignment-based methods comprised a motif-based approach using MERCI and a similarity-based BLAST technique, achieving high precision but relatively low coverage (~29%). The AI-based

approaches incorporated ML, DL, and LLM, which attained an AUC of up to 0.707 and an MCC of 0.268 on an independent validation set. The ensemble method, combining alignment-based and AI-based approaches, further improved performance, achieving an MCC of 0.352 and AUC of 0.73 on an independent validation dataset. To facilitate broader usage, we developed EmiRPred, a user-friendly web server designed to support the scientific community in predicting exosomal miRNAs and identifying relevant motifs (https://webs.iiitd.edu.in/raghava/emirpred/).

Chapter 6 discusses the development of a bioinformatics tool for predicting abundant miRNA in exosomes. Non-invasive disease diagnosis through liquid biopsy commonly utilizes blood-derived exosomes, which carry a variety of biomolecules, notably microRNAs (miRNAs) originating from their parental cells. Identifying miRNA-based biomarkers for diseases necessitates predicting miRNAs that are highly abundant in exosomes under healthy conditions. This baseline benchmarking is critical for exploring their physiological functions and disease-related changes. In our study, we developed predictive models to identify highly abundant exosomal miRNAs based on their nucleotide sequences. The dataset used in this study consisted of 348 abundant and 349 non-abundant miRNAs collected from EVmiRNA and GEO databases. Initially, we employed alignment-based methods such as motif identification and sequence similarity searches, but they resulted in limited coverage. Therefore, we shifted our focus to alignment-free methods, specifically machine learning techniques utilizing a diverse set of features. An Extra Trees classifier incorporating TF-IDF and binary profile-based features achieved the best performance, with an AUC of 0.77. To further enhance the performance, we developed a hybrid approach combining machine learning with alignment-based strategies, which improved performance significantly, achieving an AUC of 0.854 on an independent dataset. To facilitate ongoing research in non-invasive diagnostics and therapeutic development, we established AdmirePred, a comprehensive resource available as a web server, standalone software, and Python package accessible at https://webs.iiitd.edu.in/raghava/admirepred/.

Chapter 7 focuses on the discovery of saliva-based biomarkers for the prediction of Gastric Cancer (GC) patients. Saliva-based biomarkers provide a non-invasive, convenient, and patient-friendly strategy for gastric cancer (GC) detection, eliminating the requirement for uncomfortable procedures. This study aimed to discover extracellular RNA (exRNA) biomarkers in saliva that distinguish GC patients from normal controls. We analyzed a public

dataset of salivary exRNA profiles from 98 GC cases and 100 healthy individuals. After data filtering and applying a Mann-Whitney U test, we obtained a set of exRNAs significantly different between GC and normal saliva samples ($p < 0.05$). Using a threshold-based classification approach, we identified several candidate biomarkers, including GSDMA, SMR3B, NPTXR, TBCD, KCNC4, and KRT26, which individually achieved classification accuracies of approximately 61–64%. We then employed machine learning techniques. The data were split into training (80%) and independent validation (20%) sets. First, using only gene features (n = 340 significant genes), a logistic regression model attained an area under the curve (AUC) of 0.975 on the independent validation set. Applying feature selection on these genes yielded nine key gene biomarkers (GSDMA, CCDC141, HSP90B1, SLC30A4, ATP8B3, ARHGAP23, NPTXR, WRB, and SMR3B) that achieved an AUC of 0.893 on the independent set. Next, we expanded the feature set to five major RNA types in saliva (mRNA, miRNA, tRNA, circRNA, and piRNA), totaling 362 significant exRNAs. Using all 362 features, the best model (logistic regression) reached an AUC of 0.965 on the independent validation set. Feature selection on this multi-class set identified 24 optimal biomarkers (23 mRNAs and 1 tRNA) with an AUC of 0.937. Further refinement produced a primary panel of eight mRNA biomarkers (GSDMA, CCDC141, TBCD, STARD13, WRB, ARHGAP23, CDHR3, BX842679.1) that yielded an AUC of 0.905 and MCC of 0.770 with an ensemble stacking classifier. In addition to this primary panel, we identified secondary biomarker sets by iteratively removing the primary markers and reapplying feature selection. These secondary sets did not exceed an AUC of 0.80 on validation. Finally, a comparative analysis against existing GC biomarker studies demonstrated that our primary biomarker panel outperforms previously reported panels. Overall, our results underscore the potential of salivary exRNA biomarkers as an effective non-invasive tool for GC diagnosis.

The work presented in this thesis highlights various computational tools, databases, and methodologies developed to advance the identification and prediction of non-invasive biomarkers, particularly emphasizing saliva and exosome-based diagnostics and therapeutics. Resources such as SalivaDB, ExoProPred, EmiRPred, AdmirePred, and identified gastric cancer (GC) salivary biomarkers collectively offer comprehensive platforms for researchers and clinicians worldwide. These advancements significantly contribute toward transitioning from invasive medical procedures to more accessible, patient-friendly, and non-invasive approaches. We anticipate that our findings and tools will meaningfully assist clinicians,

researchers, and the broader scientific community in developing innovative non-invasive diagnostic and therapeutic strategies for improved patient care.

Bibliography

Abels, E. R., & Breakefield, X. O. (2016). Introduction to Extracellular Vesicles: Biogenesis, RNA Cargo Selection, Content, Release, and Uptake. *Cellular and Molecular Neurobiology*, *36*(3), 301–312. https://doi.org/10.1007/s10571-016-0366-z

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. https://doi.org/10.3389/fninf.2014.00014

Aegle Therapeutics. (2024). *Pipeline*. Https://Aegletherapeutics.Com/Pipeline/.

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. https://doi.org/10.3389/fmicb.2023.1148579

Ahmed, F., Ansari, H. R., & Raghava, G. P. (2009). Prediction of guide strand of microRNAs from its sequence and secondary structure. *BMC Bioinformatics*, *10*(1), 105. https://doi.org/10.1186/1471-2105-10-105

Ahmed, F., Kaundal, R., & Raghava, G. P. (2013). PHDcleav: a SVM based method for predicting human Dicer cleavage sites using sequence and secondary structure of miRNA precursors. *BMC Bioinformatics*, *14*(S14), S9. https://doi.org/10.1186/1471-2105-14-S14-S9

Ahmed, S. V., Jayawarna, C., & Jude, E. (2006). Post lumbar puncture headache: Diagnosis and management. In *Postgraduate Medical Journal*. https://doi.org/10.1136/pgmj.2006.044792

Alix-Panabières, C., & Pantel, K. (2016). Clinical Applications of Circulating Tumor Cells and Circulating Tumor DNA as Liquid Biopsy. *Cancer Discovery*, *6*(5), 479–491. https://doi.org/10.1158/2159-8290.CD-15-1483

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Alum, E. U. (2025). AI-driven biomarker discovery: enhancing precision in cancer diagnosis and prognosis. *Discover Oncology*, *16*(1), 313. https://doi.org/10.1007/s12672-025-02064-7

Alvarez-Erviti, L., Seow, Y., Yin, H., Betts, C., Lakhal, S., & Wood, M. J. A. (2011). Delivery of siRNA to the mouse brain by systemic injection of targeted exosomes. *Nature Biotechnology*, *29*(4), 341–345. https://doi.org/10.1038/nbt.1807

Armakolas, A., Kotsari, M., & Koskinas, J. (2023). Liquid Biopsies, Novel Approaches and Future Directions. *Cancers*, *15*(5), 1579. https://doi.org/10.3390/cancers15051579

Arora, A., Kaur, D., Patiyal, S., Kaur, D., Tomer, R., & Raghava, G. P. S. (2023). SalivaDB-a comprehensive database for salivary biomarkers in humans. *Database : The Journal of Biological Databases and Curation*, *2023*. https://doi.org/10.1093/database/baad002

Arora, A., Patiyal, S., Sharma, N., Devi, N. L., Kaur, D., & Raghava, G. P. S. (2023). A random forest model for predicting exosomal proteins using evolutionary information and motifs. *Proteomics*, e2300231. https://doi.org/10.1002/pmic.202300231

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*(6), e2300231. https://doi.org/10.1002/pmic.202300231

Arora, A., & Raghava, G. P. S. (2024). *Prediction of exosomal miRNA-based biomarkers for liquid biopsy*. https://doi.org/10.1101/2024.06.20.599824

Arrais, J. P., Rosa, N., Melo, J., Coelho, E. D., Amaral, D., Correia, M. J., Barros, M., & Oliveira, J. L. (2013a). OralCard: A bioinformatic tool for the study of oral proteome. *Archives of Oral Biology*. https://doi.org/10.1016/j.archoralbio.2012.12.012

Arrais, J. P., Rosa, N., Melo, J., Coelho, E. D., Amaral, D., Correia, M. J., Barros, M., & Oliveira, J. L. (2013b). OralCard: a bioinformatic tool for the study of oral proteome. *Archives of Oral Biology*, *58*(7), 762–772. https://doi.org/10.1016/j.archoralbio.2012.12.012

Arslan, F., Lai, R. C., Smeets, M. B., Akeroyd, L., Choo, A., Aguor, E. N. E., Timmers, L., van Rijen, H. V., Doevendans, P. A., Pasterkamp, G., Lim, S. K., & de Kleijn, D. P. (2013). Mesenchymal stem cell-derived exosomes increase ATP levels, decrease oxidative stress and activate PI3K/Akt pathway to enhance myocardial viability and prevent adverse remodeling after myocardial ischemia/reperfusion injury. *Stem Cell Research*, *10*(3), 301–312. https://doi.org/10.1016/j.scr.2013.01.002

Aruna Bio. (2024). *Aruna Bio's AB126 therapeutic exosome has CNS specificity*. Https://Www.Arunabio.Com/Ab126work.

Asim, M. N., Ibrahim, M. A., Malik, M. I., Zehe, C., Cloarec, O., Trygg, J., Dengel, A., & Ahmed, S. (2022). EL-RMLocNet: An explainable LSTM network for RNA-associated multi-compartment localization prediction. *Computational and Structural Biotechnology Journal*, *20*, 3986–4002. https://doi.org/10.1016/j.csbj.2022.07.031

Asim, M. N., Malik, M. I., Zehe, C., Trygg, J., Dengel, A., & Ahmed, S. (2020). MirLocPredictor: A ConvNet-Based Multi-Label MicroRNA Subcellular Localization Predictor by Incorporating k-Mer Positional Information. *Genes*, *11*(12), 1475. https://doi.org/10.3390/genes11121475

Bader, K. (2022, November 8). *FDA Accepts Phase 1 Clinical Trials of Exosome Therapy for AD*. Https://Www.Dermatologytimes.Com/View/Fda-Accepts-Phase-1-Clinical-Trials-of-Exosome-Therapy-for-Ad/1000.

Bai, T., Yan, K., & Liu, B. (2023). DAmiRLocGNet: miRNA subcellular localization prediction by combining miRNA–disease associations and graph convolutional networks. *Briefings in Bioinformatics*, *24*(4). https://doi.org/10.1093/bib/bbad212

Bakhsh, T., Alhazmi, S., Farsi, A., Yusuf, A. S., Alharthi, A., Qahl, S. H., Alghamdi, M. A., Alzahrani, F. A., Elgaddar, O. H., Ibrahim, M. A., & Bahieldin, A. (2024). Molecular detection of exosomal miRNAs of blood serum for prognosis of colorectal cancer. *Scientific Reports*, *14*(1), 8902. https://doi.org/10.1038/s41598-024-58536-3

Banavar, G., Ogundijo, O., Julian, C., Toma, R., Camacho, F., Torres, P. J., Hu, L., Chandra, T., Piscitello, A., Kenny, L., Vasani, S., Batstone, M., Dimitrova, N., Vuyisich, M., Amar, S., & Punyadeera, C. (2023). Detecting salivary host and microbiome RNA signature for aiding diagnosis of oral and throat cancer. *Oral Oncology*, *145*, 106480. https://doi.org/10.1016/j.oraloncology.2023.106480

Bao, Y., Zhang, D., Guo, H., & Ma, W. (2024). Beyond blood: Advancing the frontiers of liquid biopsy in oncology and personalized medicine. *Cancer Science*, *115*(4), 1060–1072. https://doi.org/10.1111/cas.16097

Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, *12*(1), 56–68. https://doi.org/10.1038/nrg2918

Barile, L., Moccetti, T., Marbán, E., & Vassalli, G. (2016). Roles of exosomes in cardioprotection. *European Heart Journal*, ehw304. https://doi.org/10.1093/eurheartj/ehw304

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., & Edgar, R. (2007). NCBI GEO: Mining tens of

millions of expression profiles - Database and tools update. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkl887
Bartel, D. P. (2018). Metazoan MicroRNAs. *Cell*, *173*(1), 20–51. https://doi.org/10.1016/j.cell.2018.03.006
Becker, A., Thakur, B. K., Weiss, J. M., Kim, H. S., Peinado, H., & Lyden, D. (2016). Extracellular Vesicles in Cancer: Cell-to-Cell Mediators of Metastasis. *Cancer Cell*, *30*(6), 836–848. https://doi.org/10.1016/j.ccell.2016.10.009
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*(21), 10937–10949. https://doi.org/10.1093/nar/gks832
Belstrøm, D., Sembler-Møller, M. L., Grande, M. A., Kirkby, N., Cotton, S. L., Paster, B. J., & Holmstrup, P. (2017). Microbial profile comparisons of saliva, pooled and site-specific subgingival samples in periodontitis patients. *PLOS ONE*, *12*(8), e0182992. https://doi.org/10.1371/journal.pone.0182992
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*(4), 349–356. https://doi.org/10.1093/protein/gzh037
Besse, B., Charrier, M., Lapierre, V., Dansin, E., Lantz, O., Planchard, D., Le Chevalier, T., Livartoski, A., Barlesi, F., Laplanche, A., Ploix, S., Vimond, N., Peguillet, I., Théry, C., Lacroix, L., Zoernig, I., Dhodapkar, K., Dhodapkar, M., Viaud, S., … Chaput, N. (2016). Dendritic cell-derived exosomes as maintenance immunotherapy after first line chemotherapy in NSCLC. *OncoImmunology*, *5*(4). https://doi.org/10.1080/2162402X.2015.1071008
Bhalla, S., Chaudhary, K., Kumar, R., Sehgal, M., Kaur, H., Sharma, S., & Raghava, G. P. S. (2017). Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *Scientific Reports*, *7*(1), 44997. https://doi.org/10.1038/srep44997
Bhalla, S., Verma, R., Kaur, H., Kumar, R., Usmani, S. S., Sharma, S., & Raghava, G. P. S. (2017a). CancerPDF: A repository of cancer-associated peptidome found in human biofluids. *Scientific Reports*. https://doi.org/10.1038/s41598-017-01633-3
Bhalla, S., Verma, R., Kaur, H., Kumar, R., Usmani, S. S., Sharma, S., & Raghava, G. P. S. (2017b). CancerPDF: A repository of cancer-associated peptidome found in human biofluids. *Scientific Reports*, *7*(1), 1511. https://doi.org/10.1038/s41598-017-01633-3
Bi, Y., Li, F., Guo, X., Wang, Z., Pan, T., Guo, Y., Webb, G. I., Yao, J., Jia, C., & Song, J. (2022). Clarion is a multi-label problem transformation method for identifying mRNA subcellular localizations. *Briefings in Bioinformatics*, *23*(6). https://doi.org/10.1093/bib/bbac467
Blanco-González, A., Cabezón, A., Seco-González, A., Conde-Torres, D., Antelo-Riveiro, P., Piñeiro, Á., & Garcia-Fandino, R. (2023). The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies. *Pharmaceuticals (Basel, Switzerland)*, *16*(6). https://doi.org/10.3390/ph16060891
Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J., & Zaretskaya, I. (2013). BLAST: a more efficient report with usability improvements. *Nucleic Acids Research*, *41*(Web Server issue), W29-33. https://doi.org/10.1093/nar/gkt282
Boukouris, S., & Mathivanan, S. (2015). Exosomes in bodily fluids are a highly stable resource of disease biomarkers. *Proteomics. Clinical Applications*, *9*(3–4), 358–367. https://doi.org/10.1002/prca.201400114
Breiman, L. (2001). Random forests. *Machine Learning*. https://doi.org/10.1023/A:1010933404324

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K. D., Maglott, D. R., & Murphy, T. D. (2015). Gene: a gene-centered information resource at NCBI. *Nucleic Acids Research*, *43*(Database issue), D36-42. https://doi.org/10.1093/nar/gku1055

Budnik, V., Ruiz-Cañada, C., & Wendler, F. (2016). Extracellular vesicles round off communication in the nervous system. *Nature Reviews Neuroscience*, *17*(3), 160–172. https://doi.org/10.1038/nrn.2015.29

Bulac, C., & Bulac, A. (2016). Decision Trees. In *Advanced Solutions in Power Systems: HVDC, FACTS, and AI Techniques*. https://doi.org/10.1002/9781119175391.ch18

Buscail, E., Alix-Panabières, C., Quincy, P., Cauvin, T., Chauvet, A., Degrandi, O., Caumont, C., Verdon, S., Lamrissi, I., Moranvillier, I., Buscail, C., Marty, M., Laurent, C., Vendrely, V., Moreau-Gaudry, F., Bedel, A., Dabernat, S., & Chiche, L. (2019). High Clinical Value of Liquid Biopsy to Detect Circulating Tumor Cells and Tumor Exosomes in Pancreatic Ductal Adenocarcinoma Patients Eligible for Up-Front Surgery. *Cancers*, *11*(11), 1656. https://doi.org/10.3390/cancers11111656

Cai, J., Wang, T., Deng, X., Tang, L., & Liu, L. (2023). GM-lncLoc: LncRNAs subcellular localization prediction based on graph neural network with meta-learning. *BMC Genomics*, *24*(1), 52. https://doi.org/10.1186/s12864-022-09034-1

Campo, J., Perea, M. A., del Romero, J., Cano, J., Hernando, V., & Bascones, A. (2006). Oral transmission of HIV, reality or fiction? An update. *Oral Diseases*, *12*(3), 219–228. https://doi.org/10.1111/j.1601-0825.2005.01187.x

Chang, S., He, S., Qiu, G., Lu, J., Wang, J., Liu, J., Fan, L., Zhao, W., & Che, X. (2016). MicroRNA-125b promotes invasion and metastasis of gastric cancer by targeting STARD13 and NEU1. *Tumour Biology : The Journal of the International Society for Oncodevelopmental Biology and Medicine*, *37*(9), 12141–12151. https://doi.org/10.1007/s13277-016-5094-y

Chauhan, R., Ghanshala, K. K., & Joshi, R. C. (2018). Convolutional Neural Network (CNN) for Image Detection and Recognition. *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 278–282. https://doi.org/10.1109/ICSCCC.2018.8703316

Chen, G., Huang, A. C., Zhang, W., Zhang, G., Wu, M., Xu, W., Yu, Z., Yang, J., Wang, B., Sun, H., Xia, H., Man, Q., Zhong, W., Antelo, L. F., Wu, B., Xiong, X., Liu, X., Guan, L., Li, T., … Guo, W. (2018). Exosomal PD-L1 contributes to immunosuppression and is associated with anti-PD-1 response. *Nature*, *560*(7718), 382–386. https://doi.org/10.1038/s41586-018-0392-8

Chen, I.-H., Xue, L., Hsu, C.-C., Paez, J. S. P., Pan, L., Andaluz, H., Wendt, M. K., Iliuk, A. B., Zhu, J.-K., & Tao, W. A. (2017). Phosphoproteins in extracellular vesicles as candidate markers for breast cancer. *Proceedings of the National Academy of Sciences*, *114*(12), 3175–3180. https://doi.org/10.1073/pnas.1618088114

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Cheng, F., Wang, Z., Huang, Y., Duan, Y., & Wang, X. (2015). Investigation of salivary free amino acid profile for early diagnosis of breast cancer with ultra performance liquid chromatography-mass spectrometry. *Clinica Chimica Acta; International Journal of Clinical Chemistry*, *447*, 23–31. https://doi.org/10.1016/j.cca.2015.05.008

Cheng, L., Doecke, J. D., Sharples, R. A., Villemagne, V. L., Fowler, C. J., Rembach, A., Martins, R. N., Rowe, C. C., Macaulay, S. L., Masters, C. L., Hill, A. F., & Australian Imaging, B. and L. (AIBL) R. G. (2015). Prognostic serum miRNA biomarkers associated with Alzheimer's disease shows concordance with neuropsychological and

neuroimaging assessment. *Molecular Psychiatry*, *20*(10), 1188–1196. https://doi.org/10.1038/mp.2014.127

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., … Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, *15*(141). https://doi.org/10.1098/rsif.2017.0387

Choudhury, S., Bajiya, N., Patiyal, S., & Raghava, G. P. S. (2024). MRSLpred—a hybrid approach for predicting multi-label subcellular localization of mRNA at the genome scale. *Frontiers in Bioinformatics*, *4*. https://doi.org/10.3389/fbinf.2024.1341479

Clayton, A., Turkes, A., Navabi, H., Mason, M. D., & Tabi, Z. (2005). Induction of heat shock proteins in B-cell exosomes. *Journal of Cell Science*, *118*(16), 3631–3638. https://doi.org/10.1242/jcs.02494

Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A. A., Wong, F., Mattox, A., Hruban, R. H., Wolfgang, C. L., Goggins, M. G., Dal Molin, M., Wang, T.-L., Roden, R., Klein, A. P., Ptak, J., Dobbyn, L., … Papadopoulos, N. (2018a). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, *359*(6378), 926–930. https://doi.org/10.1126/science.aar3247

Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A. A., Wong, F., Mattox, A., Hruban, R. H., Wolfgang, C. L., Goggins, M. G., Dal Molin, M., Wang, T.-L., Roden, R., Klein, A. P., Ptak, J., Dobbyn, L., … Papadopoulos, N. (2018b). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, *359*(6378), 926–930. https://doi.org/10.1126/science.aar3247

Colombo, M., Raposo, G., & Théry, C. (2014). Biogenesis, Secretion, and Intercellular Interactions of Exosomes and Other Extracellular Vesicles. *Annual Review of Cell and Developmental Biology*, *30*(1), 255–289. https://doi.org/10.1146/annurev-cellbio-101512-122326

Condrat, C. E., Thompson, D. C., Barbu, M. G., Bugnar, O. L., Boboc, A., Cretoiu, D., Suciu, N., Cretoiu, S. M., & Voinea, S. C. (2020). miRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis. *Cells*, *9*(2). https://doi.org/10.3390/cells9020276

Cristianini, N., & Ricci, E. (2008). Support Vector Machines. In *Encyclopedia of Algorithms* (pp. 928–932). Springer US. https://doi.org/10.1007/978-0-387-30162-4_415

Cui, T., Dou, Y., Tan, P., Ni, Z., Liu, T., Wang, D., Huang, Y., Cai, K., Zhao, X., Xu, D., Lin, H., & Wang, D. (2022). RNALocate v2.0: an updated resource for RNA subcellular localization with increased coverage and annotation. *Nucleic Acids Research*, *50*(D1), D333–D339. https://doi.org/10.1093/nar/gkab825

Dai, S., Wei, D., Wu, Z., Zhou, X., Wei, X., Huang, H., & Li, G. (2008). Phase I Clinical Trial of Autologous Ascites-derived Exosomes Combined With GM-CSF for Colorectal Cancer. *Molecular Therapy*, *16*(4), 782–790. https://doi.org/10.1038/mt.2008.1

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

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*. https://doi.org/10.1093/bib/bbaa259

Dhanda, S. K., Vir, P., & Raghava, G. P. S. (2013). Designing of interferon-gamma inducing MHC class-II binders. *Biology Direct*, *8*, 30. https://doi.org/10.1186/1745-6150-8-30

Doyle, L. M., & Wang, M. Z. (2019). Overview of Extracellular Vesicles, Their Origin, Composition, Purpose, and Methods for Exosome Isolation and Analysis. *Cells*, *8*(7). https://doi.org/10.3390/cells8070727

Doyle, L., & Wang, M. (2019). Overview of Extracellular Vesicles, Their Origin, Composition, Purpose, and Methods for Exosome Isolation and Analysis. *Cells*, *8*(7), 727. https://doi.org/10.3390/cells8070727

El Fekih, R., Hurley, J., Tadigotla, V., Alghamdi, A., Srivastava, A., Coticchia, C., Choi, J., Allos, H., Yatim, K., Alhaddad, J., Eskandari, S., Chu, P., Mihali, A. B., Lape, I. T., Lima Filho, M. P., Aoyama, B. T., Chandraker, A., Safa, K., Markmann, J. F., … Azzi, J. R. (2021). Discovery and Validation of a Urinary Exosome mRNA Signature for the Diagnosis of Human Kidney Transplant Rejection. *Journal of the American Society of Nephrology*, *32*(4), 994–1004. https://doi.org/10.1681/ASN.2020060850

Fan, J., Lee, C.-S., Kim, S., Chen, C., Aghaloo, T., & Lee, M. (2020). Generation of Small RNA-Modulated Exosome Mimetics for Bone Regeneration. *ACS Nano*, *14*(9), 11973–11984. https://doi.org/10.1021/acsnano.0c05122

Fanarraga, M. L., Bellido, J., Jaén, C., Villegas, J. C., & Zabala, J. C. (2010). TBCD links centriologenesis, spindle microtubule dynamics, and midbody abscission in human cells. *PloS One*, *5*(1), e8846. https://doi.org/10.1371/journal.pone.0008846

Feng, S., Liang, Y., Du, W., Lv, W., & Li, Y. (2020). LncLocation: Efficient Subcellular Location Prediction of Long Non-Coding RNA-Based Multi-Source Heterogeneous Feature Fusion. *International Journal of Molecular Sciences*, *21*(19), 7271. https://doi.org/10.3390/ijms21197271

Foj, L., Ferrer, F., Serra, M., Arévalo, A., Gavagnach, M., Giménez, N., & Filella, X. (2017). Exosomal and Non-Exosomal Urinary miRNAs in Prostate Cancer Detection and Prognosis. *The Prostate*, *77*(6), 573–583. https://doi.org/10.1002/pros.23295

Frampton, A. E., Prado, M. M., López-Jiménez, E., Fajardo-Puerta, A. B., Jawad, Z. A. R., Lawton, P., Giovannetti, E., Habib, N. A., Castellano, L., Stebbing, J., Krell, J., & Jiao, L. R. (2018). Glypican-1 is enriched in circulating-exosomes in pancreatic cancer and correlates with tumor burden. *Oncotarget*, *9*(27), 19006–19013. https://doi.org/10.18632/oncotarget.24873

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*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gai, C., Camussi, F., Broccoletti, R., Gambino, A., Cabras, M., Molinaro, L., Carossa, S., Camussi, G., & Arduino, P. G. (2018). Salivary extracellular vesicle-associated miRNAs as potential biomarkers in oral squamous cell carcinoma. *BMC Cancer*, *18*(1), 439. https://doi.org/10.1186/s12885-018-4364-z

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*(2), 129–140. http://www.ncbi.nlm.nih.gov/pubmed/18928201

Garg, A., Singhal, N., Kumar, R., & Kumar, M. (2020). mRNALoc: a novel machine-learning based in-silico tool to predict mRNA subcellular localization. *Nucleic Acids Research*, *48*(W1), W239–W243. https://doi.org/10.1093/nar/gkaa385

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*. https://doi.org/10.1007/s10994-006-6226-1

Ghafouri-Fard, S., Khoshbakht, T., Hussen, B. M., Taheri, M., & Samadian, M. (2022). A Review on the Role of miR-1246 in the Pathoetiology of Different Cancers. *Frontiers in Molecular Biosciences*, *8*. https://doi.org/10.3389/fmolb.2021.771835

Ghosh, A. K., Nath, A., Elangovan, E., Banerjee, A., Ramalingam, K., & Sethuraman, S. (2024). Exploring Salivary Biomarkers for Tumor Diagnosis: A Narrative Review. *Cureus*, *16*(7), e65725. https://doi.org/10.7759/cureus.65725

Gomez-Navarro, N., & Miller, E. (2016). Protein sorting at the ER-Golgi interface. *The Journal of Cell Biology*, *215*(6), 769–778. https://doi.org/10.1083/jcb.201610031

Han, Y., Jia, L., Zheng, Y., & Li, W. (2018). Salivary Exosomes: Emerging Roles in Systemic Disease. *International Journal of Biological Sciences*, *14*(6), 633–643. https://doi.org/10.7150/ijbs.25018

Hao, S., Lv, J., Yang, Q., Wang, A., Li, Z., Guo, Y., & Zhang, G. (2019). Identification of Key Genes and Circular RNAs in Human Gastric Cancer. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research*, *25*, 2488–2504. https://doi.org/10.12659/MSM.915382

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*(5), 593–600. https://doi.org/10.1016/j.gpb.2019.04.004

Hashemi, M., Mirdamadi, M. S. A., Talebi, Y., Khaniabad, N., Banaei, G., Daneii, P., Gholami, S., Ghorbani, A., Tavakolpournegari, A., Farsani, Z. M., Zarrabi, A., Nabavi, N., Zandieh, M. A., Rashidi, M., Taheriazam, A., Entezari, M., & Khan, H. (2023). Pre-clinical and clinical importance of miR-21 in human cancers: Tumorigenesis, therapy response, delivery approaches and targeting agents. *Pharmacological Research*, *187*, 106568. https://doi.org/10.1016/j.phrs.2022.106568

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Heitzer, E., Haque, I. S., Roberts, C. E. S., & Speicher, M. R. (2019). Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nature Reviews Genetics*, *20*(2), 71–88. https://doi.org/10.1038/s41576-018-0071-5

Hellhammer, D. H., Wüst, S., & Kudielka, B. M. (2009). Salivary cortisol as a biomarker in stress research. *Psychoneuroendocrinology*, *34*(2), 163–171. https://doi.org/10.1016/j.psyneuen.2008.10.026

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. https://doi.org/10.3389/fimmu.2022.792046

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*(6), 2106–2149. https://doi.org/10.1021/acsbiomaterials.1c00217

Hung, M. E., & Leonard, J. N. (2015). Stabilization of Exosome-targeting Peptides via Engineered Glycosylation. *Journal of Biological Chemistry*, *290*(13), 8166–8172. https://doi.org/10.1074/jbc.M114.621383

Hyun, S., Choi, H., Sub, Y., Hong, D., Ahn, S.-H., Choi, K., Ryu, S., Park, C., Gee, H. Y., & Choi, C. (2025). *Safety and anti-inflammatory effects of ILB-202, an engineered extracellular vesicles for NF-κB inhibition: A double-blind, randomized, placebo-controlled phase 1 trial*. https://doi.org/10.1101/2025.03.25.25324589

Iorgulescu, G. (2009). Saliva between normal and pathological. Important factors in determining systemic and oral health. In *Journal of medicine and life*.

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. https://doi.org/10.1016/j.compbiomed.2022.105297

Javaid, M. A., Ahmed, A. S., Durand, R., & Tran, S. D. (2016). Saliva as a diagnostic tool for oral and systemic diseases. *Journal of Oral Biology and Craniofacial Research*, *6*(1), 67–76. https://doi.org/10.1016/j.jobcr.2015.08.006

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*(6), 699–712. https://doi.org/10.1002/pmic.201300452

Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, *37*(15), 2112–2120. https://doi.org/10.1093/bioinformatics/btab083

Jiang, H., Zhao, H., Zhang, M., He, Y., Li, X., Xu, Y., & Liu, X. (2022). Hypoxia Induced Changes of Exosome Cargo and Subsequent Biological Effects. *Frontiers in Immunology*, *13*, 824188. https://doi.org/10.3389/fimmu.2022.824188

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*(1), 449. https://doi.org/10.1186/s12967-021-03084-x

Joshi, D., Mishra, A., & Anand, S. (2012). A naïve Gaussian Bayes classifier for detection of mental activity in gait signature. *Computer Methods in Biomechanics and Biomedical Engineering*, *15*(4), 411–416. https://doi.org/10.1080/10255842.2010.539562

Kaczor-Urbanowicz, K. E., Martin Carreras-Presas, C., Aro, K., Tu, M., Garcia-Godoy, F., & Wong, D. T. (2017). Saliva diagnostics – Current views and directions. *Experimental Biology and Medicine*, *242*(5), 459–472. https://doi.org/10.1177/1535370216681550

Kaczor-Urbanowicz, K. E., Saad, M., Grogan, T. R., Li, F., Heo, Y. J., Elashoff, D., Bresalier, R. S., Wong, D. T. W., & Kim, Y. (2022). Performance of Salivary Extracellular RNA Biomarker Panels for Gastric Cancer Differs between Distinct Populations. *Cancers*, *14*(15). https://doi.org/10.3390/cancers14153632

Kahlert, C., Melo, S. A., Protopopov, A., Tang, J., Seth, S., Koch, M., Zhang, J., Weitz, J., Chin, L., Futreal, A., & Kalluri, R. (2014). Identification of Double-stranded Genomic DNA Spanning All Chromosomes with Mutated KRAS and p53 DNA in the Serum Exosomes of Patients with Pancreatic Cancer. *Journal of Biological Chemistry*, *289*(7), 3869–3875. https://doi.org/10.1074/jbc.C113.532267

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A Convolutional Neural Network for Modelling Sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 655–665. https://doi.org/10.3115/v1/P14-1062

Kalluri, R., & LeBleu, V. S. (2020). The biology, function, and biomedical applications of exosomes. *Science (New York, N.Y.)*, *367*(6478). https://doi.org/10.1126/science.aau6977

Kamerkar, S., LeBleu, V. S., Sugimoto, H., Yang, S., Ruivo, C. F., Melo, S. A., Lee, J. J., & Kalluri, R. (2017). Exosomes facilitate therapeutic targeting of oncogenic KRAS in pancreatic cancer. *Nature*, *546*(7659), 498–503. https://doi.org/10.1038/nature22341

Kamerkar, S., Leng, C., Burenkova, O., Jang, S. C., McCoy, C., Zhang, K., Dooley, K., Kasera, S., Zi, T., Sisó, S., Dahlberg, W., Sia, C. L., Patel, S., Schmidt, K., Economides, K., Soos, T., Burzyn, D., & Sathyanarayanan, S. (2022). Exosome-mediated genetic reprogramming of tumor-associated macrophages by exoASO-STAT6 leads to potent

monotherapy antitumor activity. *Science Advances*, *8*(7). https://doi.org/10.1126/sciadv.abj7002

Kandaswamy, K. K., Pugalenthi, G., Hartmann, E., Kalies, K.-U., Möller, S., Suganthan, P. N., & Martinetz, T. (2010a). SPRED: A machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes. *Biochemical and Biophysical Research Communications*, *391*(3), 1306–1311. https://doi.org/10.1016/j.bbrc.2009.12.019

Kandaswamy, K. K., Pugalenthi, G., Hartmann, E., Kalies, K.-U., Möller, S., Suganthan, P. N., & Martinetz, T. (2010b). SPRED: A machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes. *Biochemical and Biophysical Research Communications*, *391*(3), 1306–1311. https://doi.org/10.1016/j.bbrc.2009.12.019

Katakowski, M., Buller, B., Zheng, X., Lu, Y., Rogers, T., Osobamiro, O., Shu, W., Jiang, F., & Chopp, M. (2013). Exosomes from marrow stromal cells expressing miR-146b inhibit glioma growth. *Cancer Letters*, *335*(1), 201–204. https://doi.org/10.1016/j.canlet.2013.02.019

Kaur, D., Arora, A., Vigneshwar, P., & Raghava, G. P. S. (2024). Prediction of peptide hormones using an ensemble of machine learning and similarity-based methods. *PROTEOMICS*. https://doi.org/10.1002/pmic.202400004

Kim, D., Kang, B., Kim, O. Y., Choi, D., Lee, J., Kim, S. R., Go, G., Yoon, Y. J., Kim, J. H., Jang, S. C., Park, K., Choi, E., Kim, K. P., Desiderio, D. M., Kim, Y., Lötvall, J., Hwang, D., & Gho, Y. S. (2013). EVpedia: an integrated database of high-throughput data for systemic analyses of extracellular vesicles. *Journal of Extracellular Vesicles*, *2*(1). https://doi.org/10.3402/jev.v2i0.20384

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., & Bryant, S. H. (2016). PubChem Substance and Compound databases. *Nucleic Acids Research*, *44*(D1), D1202-13. https://doi.org/10.1093/nar/gkv951

Kooijmans, S. A. A., Aleza, C. G., Roffler, S. R., van Solinge, W. W., Vader, P., & Schiffelers, R. M. (2016). Display of GPI-anchored anti-EGFR nanobodies on extracellular vesicles promotes tumour cell targeting. *Journal of Extracellular Vesicles*, *5*, 31053. https://doi.org/10.3402/jev.v5.31053

Kordelas, L., Rebmann, V., Ludwig, A.-K., Radtke, S., Ruesing, J., Doeppner, T. R., Epple, M., Horn, P. A., Beelen, D. W., & Giebel, B. (2014). MSC-derived exosomes: a novel tool to treat therapy-refractory graft-versus-host disease. *Leukemia*, *28*(4), 970–973. https://doi.org/10.1038/leu.2014.41

Kourembanas, S. (2015). Exosomes: Vehicles of Intercellular Signaling, Biomarkers, and Vectors of Cell Therapy. *Annual Review of Physiology*, *77*(1), 13–27. https://doi.org/10.1146/annurev-physiol-021014-071641

Kowal, J., Arras, G., Colombo, M., Jouve, M., Morath, J. P., Primdal-Bengtson, B., Dingli, F., Loew, D., Tkach, M., & Théry, C. (2016). Proteomic comparison defines novel markers to characterize heterogeneous populations of extracellular vesicle subtypes. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(8), E968-77. https://doi.org/10.1073/pnas.1521230113

Kowal, J., Tkach, M., & Théry, C. (2014). Biogenesis and secretion of exosomes. *Current Opinion in Cell Biology*, *29*, 116–125. https://doi.org/10.1016/j.ceb.2014.05.004

Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Research*, *47*(D1), D155–D162. https://doi.org/10.1093/nar/gky1141

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*(20), 5520–5525. https://doi.org/10.1073/pnas.1602171113

Kubala, E., Strzelecka, P., Grzegocka, M., Lietz-Kijak, D., Gronwald, H., Skomro, P., & Kijak, E. (2018). A Review of Selected Studies That Determine the Physical and Chemical Properties of Saliva in the Field of Dental Treatment. In *BioMed Research International*. https://doi.org/10.1155/2018/6572381

Kumar, M., Gromiha, M. M., & Raghava, G. P. S. (2007). Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics*, *8*, 463. https://doi.org/10.1186/1471-2105-8-463

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*(1), 56. https://doi.org/10.1186/s12929-022-00837-8

Kuwabara, Y., Ono, K., Horie, T., Nishi, H., Nagao, K., Kinoshita, M., Watanabe, S., Baba, O., Kojima, Y., Shizuta, S., Imai, M., Tamura, T., Kita, T., & Kimura, T. (2011). Increased microRNA-1 and microRNA-133a levels in serum of patients with cardiovascular disease indicate myocardial damage. *Circulation. Cardiovascular Genetics*, *4*(4), 446–454. https://doi.org/10.1161/CIRCGENETICS.110.958975

Lai, H., Li, Y., Zhang, H., Hu, J., Liao, J., Su, Y., Li, Q., Chen, B., Li, C., Wang, Z., Li, Y., Wang, J., Meng, Z., Huang, Z., & Huang, S. (2022). exoRBase 2.0: an atlas of mRNA, lncRNA and circRNA in extracellular vesicles from human biofluids. *Nucleic Acids Research*, *50*(D1), D118–D128. https://doi.org/10.1093/nar/gkab1085

Lai, X., Wang, M., McElyea, S. D., Sherman, S., House, M., & Korc, M. (2017). A microRNA signature in circulating exosomes is superior to exosomal glypican-1 levels for diagnosing pancreatic cancer. *Cancer Letters*, *393*, 86–93. https://doi.org/10.1016/j.canlet.2017.02.019

Lässer, C., Seyed Alikhani, V., Ekström, K., Eldh, M., Torregrosa Paredes, P., Bossios, A., Sjöstrand, M., Gabrielsson, S., Lötvall, J., & Valadi, H. (2011). Human saliva, plasma and breast milk exosomes contain RNA: uptake by macrophages. *Journal of Translational Medicine*, *9*(1), 9. https://doi.org/10.1186/1479-5876-9-9

Lau, C., Kim, Y., Chia, D., Spielmann, N., Eibl, G., Elashoff, D., Wei, F., Lin, Y.-L., Moro, A., Grogan, T., Chiang, S., Feinstein, E., Schafer, C., Farrell, J., & Wong, D. T. W. (2013). Role of pancreatic cancer-derived exosomes in salivary biomarker development. *The Journal of Biological Chemistry*, *288*(37), 26888–26897. https://doi.org/10.1074/jbc.M113.452458

Lau, W. W., Hardt, M., Zhang, Y. H., Freire, M., & Ruhl, S. (2021a). The Human Salivary Proteome Wiki: A Community-Driven Research Platform. *Journal of Dental Research*. https://doi.org/10.1177/00220345211014432

Lau, W. W., Hardt, M., Zhang, Y. H., Freire, M., & Ruhl, S. (2021b). The Human Salivary Proteome Wiki: A Community-Driven Research Platform. *Journal of Dental Research*, *100*(13), 1510–1519. https://doi.org/10.1177/00220345211014432

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. In *Nature* (Vol. 521, Issue 7553, pp. 436–444). Nature Publishing Group. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Lee, M. G., Ohana, E., Park, H. W., Yang, D., & Muallem, S. (2012). Molecular mechanism of pancreatic and salivary gland fluid and HCO3 secretion. *Physiological Reviews*, *92*(1), 39–74. https://doi.org/10.1152/physrev.00011.2011

Lee, Y.-H., & Wong, D. T. (2009). Saliva: an emerging biofluid for early detection of diseases. *American Journal of Dentistry*, *22*(4), 241–248.

Li, F., Kaczor-Urbanowicz, K. E., Sun, J., Majem, B., Lo, H.-C., Kim, Y., Koyano, K., Rao, S. L., Kang, S. Y., Kim, S. M., Kim, K.-M., Kim, S., Chia, D., Elashoff, D., Grogan, T. R., Xiao, X., & Wong, D. T. W. (2018). Characterization of Human Salivary Extracellular RNA by Next-generation Sequencing. *Clinical Chemistry*, *64*(7), 1085–1095. https://doi.org/10.1373/clinchem.2017.285072

Li, F., Yoshizawa, J. M., Kim, K.-M., Kanjanapangka, J., Grogan, T. R., Wang, X., Elashoff, D. E., Ishikawa, S., Chia, D., Liao, W., Akin, D., Yan, X., Lee, M.-S., Choi, R., Kim, S.-M., Kang, S.-Y., Bae, J.-M., Sohn, T.-S., Lee, J.-H., … Wong, D. T. W. (2018). Discovery and Validation of Salivary Extracellular RNA Biomarkers for Noninvasive Detection of Gastric Cancer. *Clinical Chemistry*, *64*(10), 1513–1521. https://doi.org/10.1373/clinchem.2018.290569

Li, M., Zhao, B., Li, Y., Ding, P., Yin, R., Kan, S., & Zeng, M. (2024). SGCL-LncLoc: An Interpretable Deep Learning Model for Improving lncRNA Subcellular Localization Prediction with Supervised Graph Contrastive Learning. *Big Data Mining and Analytics*, *7*(3), 765–780. https://doi.org/10.26599/BDMA.2024.9020002

Li, W. (Jess), Wang, Y., Liu, R., Kasinski, A. L., Shen, H., Slack, F. J., & Tang, D. G. (2021). MicroRNA-34a: Potent Tumor Suppressor, Cancer Stem Cell Inhibitor, and Potential Anticancer Therapeutic. *Frontiers in Cell and Developmental Biology*, *9*. https://doi.org/10.3389/fcell.2021.640587

Li, W., Li, C., Zhou, T., Liu, X., Liu, X., Li, X., & Chen, D. (2017). Role of exosomal proteins in cancer diagnosis. *Molecular Cancer*, *16*(1), 145. https://doi.org/10.1186/s12943-017-0706-8

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*(1), 011503. https://doi.org/10.1063/1.5087122

Li, Y., Zhou, X., St John, M. A. R., & Wong, D. T. W. (2004). RNA profiling of cell-free saliva using microarray technology. *Journal of Dental Research*, *83*(3), 199–203. https://doi.org/10.1177/154405910408300303

Liang, X., Wu, Q., Wang, Y., & Li, S. (2022). MicroRNAs as early diagnostic biomarkers for non-small cell lung cancer (Review). *Oncology Reports*, *49*(1), 8. https://doi.org/10.3892/or.2022.8445

Liang, Y., You, X., Zhang, Z., Qiu, S., Li, S., & Fu, L. (2024). MGFmiRNAloc: Predicting miRNA Subcellular Localization Using Molecular Graph Feature and Convolutional Block Attention Module. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *21*(5), 1348–1357. https://doi.org/10.1109/TCBB.2024.3383438

Lightner, A. L., Sengupta, V., Qian, S., Ransom, J. T., Suzuki, S., Park, D. J., Melson, T. I., Williams, B. P., Walsh, J. J., & Awili, M. (2023). Bone Marrow Mesenchymal Stem Cell-Derived Extracellular Vesicle Infusion for the Treatment of Respiratory Failure From COVID-19. *CHEST*, *164*(6), 1444–1453. https://doi.org/10.1016/j.chest.2023.06.024

Lin, Y., Pan, X., & Shen, H.-B. (2021). lncLocator 2.0: a cell-line-specific subcellular localization predictor for long non-coding RNAs with interpretable deep learning. *Bioinformatics*, *37*(16), 2308–2316. https://doi.org/10.1093/bioinformatics/btab127

Liu, T., Zhang, Q., Zhang, J., Li, C., Miao, Y.-R., Lei, Q., Li, Q., & Guo, A.-Y. (2019). EVmiRNA: a database of miRNA profiling in extracellular vesicles. *Nucleic Acids Research*, *47*(D1), D89–D93. https://doi.org/10.1093/nar/gky985

Liu, X., Li, X., Wang, L., Yu, K., Wu, D., Tao, P., & Li, Y. (2023). Pan-cancer analysis identified ARHGAP23 as a potential biomarker for pancreatic adenocarcinoma. *Molecular and Clinical Oncology*, *19*(6), 100. https://doi.org/10.3892/mco.2023.2696

Liu, Z., Bai, T., Liu, B., & Yu, L. (2024). MulStack: An ensemble learning prediction model of multilabel mRNA subcellular localization. *Computers in Biology and Medicine*, *175*, 108289. https://doi.org/10.1016/j.compbiomed.2024.108289

Llena-Puy, C. (2006). The rôle of saliva in maintaining oral health and as an aid to diagnosis. In *Medicina oral, patología oral y cirugía bucal*.

Lopez-Verrilli, M. A., & Court, F. A. (2013). Exosomes: mediators of communication in eukaryotes. *Biological Research*, *46*(1), 5–11. https://doi.org/10.4067/S0716-97602013000100001

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26. https://doi.org/10.1186/1748-7188-6-26

Lu, Y., Godbout, K., Lamothe, G., & Tremblay, J. P. (2023). CRISPR-Cas9 delivery strategies with engineered extracellular vesicles. *Molecular Therapy - Nucleic Acids*, *34*, 102040. https://doi.org/10.1016/j.omtn.2023.102040

Luan, X., Sansanaphongpricha, K., Myers, I., Chen, H., Yuan, H., & Sun, D. (2017). Engineering exosomes as refined biological nanoplatforms for drug delivery. *Acta Pharmacologica Sinica*, *38*(6), 754–763. https://doi.org/10.1038/aps.2017.12

Ludwig, J. A., & Weinstein, J. N. (2005). Biomarkers in Cancer Staging, Prognosis and Treatment Selection. *Nature Reviews Cancer*, *5*(11), 845–856. https://doi.org/10.1038/nrc1739

M, T., Koti, M. S., B A, N., V, G., K P, S., Mathivanan, S. K., & Dalu, G. T. (2024). Lung cancer diagnosis based on weighted convolutional neural network using gene data expression. *Scientific Reports*, *14*(1), 3656. https://doi.org/10.1038/s41598-024-54124-7

Malamud, D. (2011). Saliva as a diagnostic fluid. *Dental Clinics of North America*, *55*(1), 159–178. https://doi.org/10.1016/j.cden.2010.08.004

Malathi, N., Mythili, S., & Vasanthi, H. R. (2014). Salivary diagnostics: a brief review. *ISRN Dentistry*, *2014*, 158786. https://doi.org/10.1155/2014/158786

Malon, R. S. P., Sadir, S., Balakrishnan, M., & Córcoles, E. P. (2014). Saliva-Based Biosensors: Noninvasive Monitoring Tool for Clinical Diagnostics. *BioMed Research International*, *2014*, 1–20. https://doi.org/10.1155/2014/962903

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., & Ferrari, R. (2018). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*, *19*(2), 286–302. https://doi.org/10.1093/bib/bbw114

Mathieu, M., Martin-Jaular, L., Lavieu, G., & Théry, C. (2019). Specificities of secretion and uptake of exosomes and other extracellular vesicles for cell-to-cell communication. *Nature Cell Biology*, *21*(1), 9–17. https://doi.org/10.1038/s41556-018-0250-9

Mathivanan, S., Fahner, C. J., Reid, G. E., & Simpson, R. J. (2012). ExoCarta 2012: database of exosomal proteins, RNA and lipids. *Nucleic Acids Research*, *40*(D1), D1241–D1244. https://doi.org/10.1093/nar/gkr828

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*, 2021.12.14.472723. https://doi.org/10.1101/2021.12.14.472723

McDonald, C. M., Marbán, E., Hendrix, S., Hogan, N., Ruckdeschel Smith, R., Eagle, M., Finkel, R. S., Tian, C., Janas, J., Harmelink, M. M., Varadhachary, A. S., Taylor, M. D., Hor, K. N., Mayer, O. H., Henricson, E. K., Furlong, P., Ascheim, D. D., Rogy, S., Williams, P., … Rogers, R. G. (2022). Repeated intravenous cardiosphere-derived cell therapy in late-stage Duchenne muscular dystrophy (HOPE-2): a multicentre, randomised, double-blind, placebo-controlled, phase 2 trial. *The Lancet*, *399*(10329), 1049–1058. https://doi.org/10.1016/S0140-6736(22)00012-5

Meher, P. K., Rai, A., & Rao, A. R. (2021). mLoc-mRNA: predicting multiple sub-cellular localization of mRNAs using random forest algorithm coupled with feature selection via elastic net. *BMC Bioinformatics*, *22*(1), 342. https://doi.org/10.1186/s12859-021-04264-8

Meher, P. K., Satpathy, S., & Rao, A. R. (2021). Publisher Correction: miRNALoc: predicting miRNA subcellular localizations based on principal component scores of physico-chemical properties and pseudo compositions of di-nucleotides. *Scientific Reports*, *11*(1), 3287. https://doi.org/10.1038/s41598-021-81881-6

Meldolesi, J. (2022). Unconventional Protein Secretion Dependent on Two Extracellular Vesicles: Exosomes and Ectosomes. *Frontiers in Cell and Developmental Biology*, *10*, 877344. https://doi.org/10.3389/fcell.2022.877344

Melo, S. A., Luecke, L. B., Kahlert, C., Fernandez, A. F., Gammon, S. T., Kaye, J., LeBleu, V. S., Mittendorf, E. A., Weitz, J., Rahbari, N., Reissfelder, C., Pilarsky, C., Fraga, M. F., Piwnica-Worms, D., & Kalluri, R. (2015). Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. *Nature*, *523*(7559), 177–182. https://doi.org/10.1038/nature14581

Melo, S. A., Sugimoto, H., O'Connell, J. T., Kato, N., Villanueva, A., Vidal, A., Qiu, L., Vitkin, E., Perelman, L. T., Melo, C. A., Lucci, A., Ivan, C., Calin, G. A., & Kalluri, R. (2014). Cancer Exosomes Perform Cell-Independent MicroRNA Biogenesis and Promote Tumorigenesis. *Cancer Cell*, *26*(5), 707–721. https://doi.org/10.1016/j.ccell.2014.09.005

Menini, M., De Giovanni, E., Bagnasco, F., Delucchi, F., Pera, F., Baldi, D., & Pesce, P. (2021). Salivary Micro-RNA and Oral Squamous Cell Carcinoma: A Systematic Review. *Journal of Personalized Medicine*, *11*(2), 101. https://doi.org/10.3390/jpm11020101

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., Peterson, A., Noteboom, J., O'Briant, K. C., Allen, A., Lin, D. W., Urban, N., Drescher, C. W., Knudsen, B. S., Stirewalt, D. L., Gentleman, R., Vessella, R. L., Nelson, P. S., Martin, D. B., & Tewari, M. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences*, *105*(30), 10513–10518. https://doi.org/10.1073/pnas.0804549105

Moreno-Gonzalo, O., Villarroya-Beltri, C., & Sánchez-Madrid, F. (2014). Post-translational modifications of exosomal proteins. *Frontiers in Immunology*, *5*, 383. https://doi.org/10.3389/fimmu.2014.00383

Mosquera-Heredia, M. I., Morales, L. C., Vidal, O. M., Barceló, E., Silvera-Redondo, C., Vélez, J. I., & Garavito-Galofre, P. (2021). Exosomes: Potential Disease Biomarkers and New Therapeutic Targets. *Biomedicines*, *9*(8). https://doi.org/10.3390/biomedicines9081061

Mulcahy, L. A., Pink, R. C., & Carter, D. R. F. (2014). Routes and mechanisms of extracellular vesicle uptake. *Journal of Extracellular Vesicles*, *3*(1). https://doi.org/10.3402/jev.v3.24641

Mulindwa, J., Noyes, H., Ilboudo, H., Pagani, L., Nyangiri, O., Kimuda, M. P., Ahouty, B., Asina, O. F., Ofon, E., Kamoto, K., Kabore, J. W., Koffi, M., Ngoyi, D. M., Simo, G.,

Chisi, J., Sidibe, I., Enyaru, J., Simuunza, M., Alibu, P., … TrypanoGEN Research Group of the H3Africa Consortium. (2020). High Levels of Genetic Diversity within Nilo-Saharan Populations: Implications for Human Adaptation. *American Journal of Human Genetics*, *107*(3), 473–486. https://doi.org/10.1016/j.ajhg.2020.07.007

Nieman, L. K. (2018). Diagnosis of Cushing's Syndrome in the Modern Era. *Endocrinology and Metabolism Clinics of North America*, *47*(2), 259–273. https://doi.org/10.1016/j.ecl.2018.02.001

Nonaka, T., & Wong, D. T. W. (2023). Saliva diagnostics: Salivaomics, saliva exosomics, and saliva liquid biopsy. *Journal of the American Dental Association (1939)*, *154*(8), 696–704. https://doi.org/10.1016/j.adaj.2023.05.006

O'Brien, K., Breyne, K., Ughetto, S., Laurent, L. C., & Breakefield, X. O. (2020a). RNA delivery by extracellular vesicles in mammalian cells and its applications. *Nature Reviews Molecular Cell Biology*, *21*(10), 585–606. https://doi.org/10.1038/s41580-020-0251-y

O'Brien, K., Breyne, K., Ughetto, S., Laurent, L. C., & Breakefield, X. O. (2020b). RNA delivery by extracellular vesicles in mammalian cells and its applications. *Nature Reviews Molecular Cell Biology*, *21*(10), 585–606. https://doi.org/10.1038/s41580-020-0251-y

Ohno, S., Takanashi, M., Sudo, K., Ueda, S., Ishikawa, A., Matsuyama, N., Fujita, K., Mizutani, T., Ohgi, T., Ochiya, T., Gotoh, N., & Kuroda, M. (2013). Systemically Injected Exosomes Targeted to EGFR Deliver Antitumor MicroRNA to Breast Cancer Cells. *Molecular Therapy*, *21*(1), 185–191. https://doi.org/10.1038/mt.2012.180

Ostrowski, M., Carmo, N. B., Krumeich, S., Fanget, I., Raposo, G., Savina, A., Moita, C. F., Schauer, K., Hume, A. N., Freitas, R. P., Goud, B., Benaroch, P., Hacohen, N., Fukuda, M., Desnos, C., Seabra, M. C., Darchen, F., Amigorena, S., Moita, L. F., & Thery, C. (2010). Rab27a and Rab27b control different steps of the exosome secretion pathway. *Nature Cell Biology*, *12*(1), 19–30. https://doi.org/10.1038/ncb2000

Otmani, K., Rouas, R., Berehab, M., & Lewalle, P. (2024). The regulatory mechanisms of oncomiRs in cancer. *Biomedicine & Pharmacotherapy*, *171*, 116165. https://doi.org/10.1016/j.biopha.2024.116165

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*. https://doi.org/10.1089/cmb.2022.0241

Pantel, K., & Alix-Panabières, C. (2013). Real-time Liquid Biopsy in Cancer Patients: Fact or Fiction? *Cancer Research*, *73*(21), 6384–6388. https://doi.org/10.1158/0008-5472.CAN-13-2030

Panwar, B., Arora, A., & Raghava, G. P. S. (2014). Prediction and classification of ncRNAs using structural information. *BMC Genomics*, *15*, 127. https://doi.org/10.1186/1471-2164-15-127

Pathan, M., Fonseka, P., Chitti, S. V, Kang, T., Sanwlani, R., Van Deun, J., Hendrix, A., & Mathivanan, S. (2019). Vesiclepedia 2019: a compendium of RNA, proteins, lipids and metabolites in extracellular vesicles. *Nucleic Acids Research*, *47*(D1), D516–D519. https://doi.org/10.1093/nar/gky1029

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*.

Pitt, J. M., Charrier, M., Viaud, S., André, F., Besse, B., Chaput, N., & Zitvogel, L. (2014). Dendritic Cell–Derived Exosomes as Immunotherapies in the Fight against Cancer. *The Journal of Immunology*, *193*(3), 1006–1011. https://doi.org/10.4049/jimmunol.1400703

Poersch, A., Grassi, M. L., Carvalho, V. P. de, Lanfredi, G. P., Palma, C. de 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. https://doi.org/10.1016/j.jprot.2016.05.005

Poggio, E. D., McClelland, R. L., Blank, K. N., Hansen, S., Bansal, S., Bomback, A. S., Canetta, P. A., Khairallah, P., Kiryluk, K., Lecker, S. H., McMahon, G. M., Palevsky, P. M., Parikh, S., Rosas, S. E., Tuttle, K., Vazquez, M. A., Vijayan, A., & Rovin, B. H. (2020). Systematic review and meta-analysis of native kidney biopsy complications. *Clinical Journal of the American Society of Nephrology*. https://doi.org/10.2215/CJN.04710420

Qiu, J., Sun, M., Wang, Y., & Chen, B. (2020). Identification of Hub Genes and Pathways in Gastric Adenocarcinoma Based on Bioinformatics Analysis. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research*, *26*, e920261. https://doi.org/10.12659/MSM.920261

Ras-Carmona, A., Gomez-Perosanz, M., & Reche, P. A. (2021). Prediction of unconventional protein secretion by exosomes. *BMC Bioinformatics*, *22*(1), 333. https://doi.org/10.1186/s12859-021-04219-z

Rathore, S., Arora, A., Choudhury, S., Tijare, P., & Raghava, G. P. S. (n.d.). *ToxinPred 3.0: An improved method for predicting the toxicity of peptides*. https://doi.org/10.1101/2023.08.11.552911

Ren, K., Zeng, Y., Cao, Z., & Zhang, Y. (2022). ID-RDRL: a deep reinforcement learning-based feature selection intrusion detection model. *Scientific Reports*, *12*(1), 15370. https://doi.org/10.1038/s41598-022-19366-3

Reynolds, S. J., & Muwonga, J. (2004). OraQuick® ADVANCE Rapid HIV-1/2 antibody test. *Expert Review of Molecular Diagnostics*, *4*(5), 587–591. https://doi.org/10.1586/14737159.4.5.587

Rhim, J., Baek, W., Seo, Y., & Kim, J. H. (2022). From Molecular Mechanisms to Therapeutics: Understanding MicroRNA-21 in Cancer. *Cells*, *11*(18). https://doi.org/10.3390/cells11182791

RNALocate: a resource for RNA subcellular localizations. (2016). *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw728

Robbins, P. D., & Morelli, A. E. (2014). Regulation of immune responses by extracellular vesicles. *Nature Reviews Immunology*, *14*(3), 195–208. https://doi.org/10.1038/nri3622

Roi, A., Rusu, L. C., Roi, C. I., Luca, R. E., Boia, S., & Munteanu, R. I. (2019). A New Approach for the Diagnosis of Systemic and Oral Diseases Based on Salivary Biomolecules. *Disease Markers*, *2019*, 8761860. https://doi.org/10.1155/2019/8761860

Rosa, N., Correia, M. J., Arrais, J. P., Lopes, P., Melo, J., Oliveira, J. L., & Barros, M. (2012). From the salivary proteome to the OralOme: Comprehensive molecular oral biology. *Archives of Oral Biology*. https://doi.org/10.1016/j.archoralbio.2011.12.010

Russo, F., Di Bella, S., Vannini, F., Berti, G., Scoyni, F., Cook, H. V, Santos, A., Nigita, G., Bonnici, V., Laganà, A., Geraci, F., Pulvirenti, A., Giugno, R., De Masi, F., Belling, K., Jensen, L. J., Brunak, S., Pellegrini, M., & Ferro, A. (2018). miRandola 2017: a curated knowledge base of non-invasive biomarkers. *Nucleic Acids Research*, *46*(D1), D354–D359. https://doi.org/10.1093/nar/gkx854

Saeki, N., Kuwahara, Y., Sasaki, H., Satoh, H., & Shiroishi, T. (2000). Gasdermin (Gsdm) localizing to mouse Chromosome 11 is predominantly expressed in upper

gastrointestinal tract but significantly suppressed in human gastric cancer cells. *Mammalian Genome : Official Journal of the International Mammalian Genome Society*, *11*(9), 718–724. https://doi.org/10.1007/s003350010138

Saeki, N., Usui, T., Aoyagi, K., Kim, D. H., Sato, M., Mabuchi, T., Yanagihara, K., Ogawa, K., Sakamoto, H., Yoshida, T., & Sasaki, H. (2009). Distinctive expression and function of four GSDM family genes (GSDMA-D) in normal and malignant upper gastrointestinal epithelium. *Genes, Chromosomes & Cancer*, *48*(3), 261–271. https://doi.org/10.1002/gcc.20636

Salehi, M., Kamali, M. J., Arab, D., Safaeian, N., Ashuori, Z., Maddahi, M., Latifi, N., & Jahromi, A. M. (2024a). Exosomal microRNAs in regulation of tumor cells resistance to apoptosis. *Biochemistry and Biophysics Reports*, *37*, 101644. https://doi.org/10.1016/j.bbrep.2024.101644

Salehi, M., Kamali, M. J., Arab, D., Safaeian, N., Ashuori, Z., Maddahi, M., Latifi, N., & Jahromi, A. M. (2024b). Exosomal microRNAs in regulation of tumor cells resistance to apoptosis. *Biochemistry and Biophysics Reports*, *37*, 101644. https://doi.org/10.1016/j.bbrep.2024.101644

Samsonov, R., Shtam, T., Burdakov, V., Glotov, A., Tsyrlina, E., Berstein, L., Nosov, A., Evtushenko, V., Filatov, M., & Malek, A. (2016). Lectin-induced agglutination method of urinary exosomes isolation followed by mi-RNA analysis: Application for prostate cancer diagnostic. *The Prostate*, *76*(1), 68–79. https://doi.org/10.1002/pros.23101

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database : The Journal of Biological Databases and Curation*, *2020*(D1), D608–D617. https://doi.org/10.1093/database/baaa062

Shah, S. (2018). Salivaomics: The current scenario. *Journal of Oral and Maxillofacial Pathology*, *22*(3), 375. https://doi.org/10.4103/jomfp.JOMFP_171_18

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*(5). https://doi.org/10.1093/bib/bbac174

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*(4). https://doi.org/10.1093/bib/bbaa294

Shegekar, T., Vodithala, S., & Juganavar, A. (2023). The Emerging Role of Liquid Biopsies in Revolutionising Cancer Diagnosis and Therapy. *Cureus*. https://doi.org/10.7759/cureus.43650

Shibata, C., Nakano, T., Yasumoto, A., Mitamura, A., Sawada, K., Ogawa, H., Miura, T., Ise, I., Takami, K., Yamamoto, K., & Katayose, Y. (2022). Comparison of CEA and CA19-9 as a predictive factor for recurrence after curative gastrectomy in gastric cancer. *BMC Surgery*, *22*(1), 213. https://doi.org/10.1186/s12893-022-01667-z

Simons, M., & Raposo, G. (2009). Exosomes--vesicular carriers for intercellular communication. *Current Opinion in Cell Biology*, *21*(4), 575–581. https://doi.org/10.1016/j.ceb.2009.03.007

Singh, S., Kumar, R., Payra, S., & Singh, S. K. (2023). Artificial Intelligence and Machine Learning in Pharmacological Research: Bridging the Gap Between Data and Drug Discovery. *Cureus*, *15*(8), e44359. https://doi.org/10.7759/cureus.44359

Siravegna, G., Marsoni, S., Siena, S., & Bardelli, A. (2017a). Integrating liquid biopsies into the management of cancer. *Nature Reviews Clinical Oncology*, *14*(9), 531–548. https://doi.org/10.1038/nrclinonc.2017.14

Siravegna, G., Marsoni, S., Siena, S., & Bardelli, A. (2017b). Integrating liquid biopsies into the management of cancer. *Nature Reviews Clinical Oncology*, *14*(9), 531–548. https://doi.org/10.1038/nrclinonc.2017.14

Sivadasan, P., Gupta, M. K., Sathe, G., Sudheendra, H. V., Sunny, S. P., Renu, D., Hari, P. S., Gowda, H., Suresh, A., Kuriakose, M. A., & Sirdeshmukh, R. (2020). Salivary proteins from dysplastic leukoplakia and oral squamous cell carcinoma and their potential for early detection. *Journal of Proteomics*, *212*, 103574. https://doi.org/10.1016/j.jprot.2019.103574

Skog, J., Würdinger, T., van Rijn, S., Meijer, D. H., Gainche, L., Curry, W. T., Carter, B. S., Krichevsky, A. M., & Breakefield, X. O. (2008). Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. *Nature Cell Biology*, *10*(12), 1470–1476. https://doi.org/10.1038/ncb1800

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Smyth, E. C., Nilsson, M., Grabsch, H. I., van Grieken, N. C., & Lordick, F. (2020). Gastric cancer. *Lancet (London, England)*, *396*(10251), 635–648. https://doi.org/10.1016/S0140-6736(20)31288-5

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic Emergency Medicine : Official Journal of the Society for Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Sugimoto, M., Wong, D. T., Hirayama, A., Soga, T., & Tomita, M. (2010). Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles. *Metabolomics : Official Journal of the Metabolomic Society*, *6*(1), 78–95. https://doi.org/10.1007/s11306-009-0178-y

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*(2), 1416–1425. https://doi.org/10.1002/jcp.26936

Surdu, A., Foia, L. G., Luchian, I., Trifan, D., Tatarciuc, M. S., Scutariu, M. M., Ciupilan, C., & Budala, D. G. (2025). Saliva as a Diagnostic Tool for Systemic Diseases-A Narrative Review. *Medicina (Kaunas, Lithuania)*, *61*(2). https://doi.org/10.3390/medicina61020243

Swarup, N., Cheng, J., Choi, I., Heo, Y. J., Kordi, M., Aziz, M., Arora, A., Li, F., Chia, D., Wei, F., Elashoff, D., Zhang, L., Kim, S., Kim, Y., & Wong, D. T. W. (2023). Multi-faceted attributes of salivary cell-free DNA as liquid biopsy biomarkers for gastric cancer detection. *Biomarker Research*, *11*(1), 90. https://doi.org/10.1186/s40364-023-00524-2

Taba, M., Kinney, J., Kim, A. S., & Giannobile, W. V. (2005). Diagnostic biomarkers for oral and periodontal diseases. *Dental Clinics of North America*, *49*(3), 551–571, vi. https://doi.org/10.1016/j.cden.2005.03.009

Tanikawa, C., Kamatani, Y., Toyoshima, O., Sakamoto, H., Ito, H., Takahashi, A., Momozawa, Y., Hirata, M., Fuse, N., Takai-Igarashi, T., Shimizu, A., Sasaki, M., Yamaji, T., Sawada, N., Iwasaki, M., Tsugane, S., Naito, M., Hishida, A., Wakai, K., … Matsuda, K. (2018). Genome-wide association study identifies gastric cancer susceptibility loci at 12q24.11-12 and 20q11.21. *Cancer Science*, *109*(12), 4015–4024. https://doi.org/10.1111/cas.13815

Tastan, B., Tarakcioglu, E., Birinci, Y., Park, Y., & Genc, S. (2022). Role of Exosomal MicroRNAs in Cell-to-Cell Communication. *Methods in Molecular Biology (Clifton, N.J.)*, *2257*, 269–292. https://doi.org/10.1007/978-1-0716-1170-8_14

Thakur, B. K., Zhang, H., Becker, A., Matei, I., Huang, Y., Costa-Silva, B., Zheng, Y., Hoshino, A., Brazier, H., Xiang, J., Williams, C., Rodriguez-Barrueco, R., Silva, J. M., Zhang, W., Hearn, S., Elemento, O., Paknejad, N., Manova-Todorova, K., Welte, K., … Lyden, D. (2014). Double-stranded DNA in exosomes: a novel biomarker in cancer detection. *Cell Research*, *24*(6), 766–769. https://doi.org/10.1038/cr.2014.44

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. https://doi.org/10.1002/0471143030.cb0322s30

Théry, C., Ostrowski, M., & Segura, E. (2009a). Membrane vesicles as conveyors of immune responses. *Nature Reviews Immunology*, *9*(8), 581–593. https://doi.org/10.1038/nri2567

Théry, C., Ostrowski, M., & Segura, E. (2009b). Membrane vesicles as conveyors of immune responses. *Nature Reviews Immunology*, *9*(8), 581–593. https://doi.org/10.1038/nri2567

Théry, C., Witwer, K. W., Aikawa, E., Alcaraz, M. J., Anderson, J. D., Andriantsitohaina, R., Antoniou, A., Arab, T., Archer, F., Atkin-Smith, G. K., Ayre, D. C., Bach, J., Bachurski, D., Baharvand, H., Balaj, L., Baldacchino, S., Bauer, N. N., Baxter, A. A., Bebawy, M., … Zuba-Surma, E. K. (2018). Minimal information for studies of extracellular vesicles 2018 (MISEV2018): a position statement of the International Society for Extracellular Vesicles and update of the MISEV2014 guidelines. *Journal of Extracellular Vesicles*, *7*(1). https://doi.org/10.1080/20013078.2018.1535750

Théry, C., Zitvogel, L., & Amigorena, S. (2002). Exosomes: composition, biogenesis and function. *Nature Reviews. Immunology*, *2*(8), 569–579. https://doi.org/10.1038/nri855

Thompson, A. G., Gray, E., Heman-Ackah, S. M., Mäger, I., Talbot, K., Andaloussi, S. El, Wood, M. J., & Turner, M. R. (2016). Extracellular vesicles in neurodegenerative disease — pathogenesis to biomarkers. *Nature Reviews Neurology*, *12*(6), 346–357. https://doi.org/10.1038/nrneurol.2016.68

Thrift, A. P., & El-Serag, H. B. (2020). Burden of Gastric Cancer. *Clinical Gastroenterology and Hepatology : The Official Clinical Practice Journal of the American Gastroenterological Association*, *18*(3), 534–542. https://doi.org/10.1016/j.cgh.2019.07.045

Tian, Y., Li, S., Song, J., Ji, T., Zhu, M., Anderson, G. J., Wei, J., & Nie, G. (2014). A doxorubicin delivery platform using engineered natural membrane vesicle exosomes for targeted tumor therapy. *Biomaterials*, *35*(7), 2383–2390. https://doi.org/10.1016/j.biomaterials.2013.11.083

Tian, Y., Ma, L., Gong, M., Su, G., Zhu, S., Zhang, W., Wang, S., Li, Z., Chen, C., Li, L., Wu, L., & Yan, X. (2018). Protein Profiling and Sizing of Extracellular Vesicles from Colorectal Cancer Patients *via* Flow Cytometry. *ACS Nano*, *12*(1), 671–680. https://doi.org/10.1021/acsnano.7b07782

Tkach, M., & Théry, C. (2016). Communication by Extracellular Vesicles: Where We Are and Where We Need to Go. *Cell*, *164*(6), 1226–1232. https://doi.org/10.1016/j.cell.2016.01.043

Tsering, T., Li, M., Chen, Y., Nadeau, A., Laskaris, A., Abdouh, M., Bustamante, P., & Burnier, J. V. (2022). EV-ADD, a database for EV-associated DNA in human liquid biopsy samples. *Journal of Extracellular Vesicles*, *11*(10), e12270. https://doi.org/10.1002/jev2.12270

UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Vakhshiteh, F., Hassani, S., Momenifar, N., & Pakdaman, F. (2021). Exosomal circRNAs: new players in colorectal cancer. *Cancer Cell International*, *21*(1), 483. https://doi.org/10.1186/s12935-021-02112-6

Valadi, H., Ekström, K., Bossios, A., Sjöstrand, M., Lee, J. J., & Lötvall, J. O. (2007). Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nature Cell Biology*, *9*(6), 654–659. https://doi.org/10.1038/ncb1596

van Niel, G., D'Angelo, G., & Raposo, G. (2018). Shedding light on the cell biology of extracellular vesicles. *Nature Reviews Molecular Cell Biology*, *19*(4), 213–228. https://doi.org/10.1038/nrm.2017.125

Vaz, S. N., Santana, D. S. de, Netto, E. M., Pedroso, C., Wang, W. K., Santos, F. D. A., & Brites, C. (2020). Saliva is a reliable, non-invasive specimen for SARS-CoV-2 detection. *Brazilian Journal of Infectious Diseases*. https://doi.org/10.1016/j.bjid.2020.08.001

Vens, C., Rosso, M.-N., & Danchin, E. G. J. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics (Oxford, England)*, *27*(9), 1231–1238. https://doi.org/10.1093/bioinformatics/btr110

Vila, T., Rizk, A. M., Sultan, A. S., & Jabra-Rizk, M. A. (2019). The power of saliva: Antimicrobial and beyond. *PLoS Pathogens*. https://doi.org/10.1371/journal.ppat.1008058

Vogels, C. B. F., Watkins, A. E., Harden, C. A., Brackney, D. E., Shafer, J., Wang, J., Caraballo, C., Kalinich, C. C., Ott, I. M., Fauver, J. R., Kudo, E., Lu, P., Venkataraman, A., Tokuyama, M., Moore, A. J., Muenker, M. C., Casanovas-Massana, A., Fournier, J., Bermejo, S., … Grubaugh, N. D. (2020). *SalivaDirect: A simplified and flexible platform to enhance SARS-CoV-2 testing capacity*. https://doi.org/10.1101/2020.08.03.20167791

Wan, J. C. M., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J. D., Caldas, C., Pacey, S., Baird, R., & Rosenfeld, N. (2017). Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer*, *17*(4), 223–238. https://doi.org/10.1038/nrc.2017.7

Wang, D., Zhang, Z., Jiang, Y., Mao, Z., Wang, D., Lin, H., & Xu, D. (2021). DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism. *Nucleic Acids Research*, *49*(8), e46. https://doi.org/10.1093/nar/gkab016

Wang, H., Ding, Y., Tang, J., Zou, Q., & Guo, F. (2021). Identify RNA-associated subcellular localizations based on multi-label learning using Chou's 5-steps rule. *BMC Genomics*, *22*(1), 56. https://doi.org/10.1186/s12864-020-07347-7

Wang, M., Yu, F., Ding, H., Wang, Y., Li, P., & Wang, K. (2019). Emerging Function and Clinical Values of Exosomal MicroRNAs in Cancer. *Molecular Therapy - Nucleic Acids*, *16*, 791–804. https://doi.org/10.1016/j.omtn.2019.04.027

Wang, X., Chai, Z., Pan, G., Hao, Y., Li, B., Ye, T., Li, Y., Long, F., Xia, L., & Liu, M. (2021). ExoBCD: a comprehensive database for exosomal biomarker discovery in breast cancer. *Briefings in Bioinformatics*, *22*(3). https://doi.org/10.1093/bib/bbaa088

Wang, X., Li, F., Xu, J., Rong, J., Webb, G. I., Ge, Z., Li, J., & Song, J. (2022). ASPIRER: a new computational approach for identifying non-classical secreted proteins based on deep learning. *Briefings in Bioinformatics*, *23*(2). https://doi.org/10.1093/bib/bbac031

Wang, Y., Springer, S., Mulvey, C. L., Silliman, N., Schaefer, J., Sausen, M., James, N., Rettig, E. M., Guo, T., Pickering, C. R., Bishop, J. A., Chung, C. H., Califano, J. A., Eisele, D. W., Fakhry, C., Gourin, C. G., Ha, P. K., Kang, H., Kiess, A., … Agrawal, N. (2015). Detection of somatic mutations and HPV in the saliva and plasma of patients with head and neck squamous cell carcinomas. *Science Translational Medicine*, *7*(293), 293ra104. https://doi.org/10.1126/scitranslmed.aaa8507

Wang, Z., Tan, W., Li, B., Zou, J., Li, Y., Xiao, Y., He, Y., Yoshida, S., & Zhou, Y. (2023). Exosomal non-coding RNAs in angiogenesis: Functions, mechanisms and potential clinical applications. *Heliyon*, *9*(8), e18626. https://doi.org/10.1016/j.heliyon.2023.e18626

Weber, J. A., Baxter, D. H., Zhang, S., Huang, D. Y., Huang, K. H., Lee, M. J., Galas, D. J., & Wang, K. (2010). The microRNA spectrum in 12 body fluids. *Clinical Chemistry*, *56*(11), 1733–1741. https://doi.org/10.1373/clinchem.2010.147405

Wei, F., Lin, C.-C., Joon, A., Feng, Z., Troche, G., Lira, M. E., Chia, D., Mao, M., Ho, C.-L., Su, W.-C., & Wong, D. T. W. (2014). Noninvasive Saliva-based *EGFR* Gene Mutation Detection in Patients with Lung Cancer. *American Journal of Respiratory and Critical Care Medicine*, *190*(10), 1117–1126. https://doi.org/10.1164/rccm.201406-1003OC

Whiteside, T. L. (2016). *Tumor-Derived Exosomes and Their Role in Cancer Progression* (pp. 103–141). https://doi.org/10.1016/bs.acc.2015.12.005

Wiklander, O. P. B., Nordin, J. Z., O'Loughlin, A., Gustafsson, Y., Corso, G., Mäger, I., Vader, P., Lee, Y., Sork, H., Seow, Y., Heldring, N., Alvarez-Erviti, L., Smith, C. E., Le Blanc, K., Macchiarini, P., Jungebluth, P., Wood, M. J. A., & Andaloussi, S. EL. (2015). Extracellular vesicle in vivo biodistribution is determined by cell source, route of administration and targeting. *Journal of Extracellular Vesicles*, *4*(1). https://doi.org/10.3402/jev.v4.26316

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., … Scalbert, A. (2018a). HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx1089

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., … Scalbert, A. (2018b). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*, *46*(D1), D608–D617. https://doi.org/10.1093/nar/gkx1089

Wong, D. T. W. (2012). Salivaomics. *The Journal of the American Dental Association*, *143*, 19S-24S. https://doi.org/10.14219/jada.archive.2012.0339

World Health Organization. (2023). *Noncommunicable diseases*. Https://Www.Who.Int/News-Room/Fact-Sheets/Detail/Noncommunicable-Diseases.

Wozniak, A. L., Adams, A., King, K. E., Dunn, W., Christenson, L. K., Hung, W.-T., & Weinman, S. A. (2020). The RNA binding protein FMR1 controls selective exosomal miRNA cargo loading during inflammation. *Journal of Cell Biology*, *219*(10). https://doi.org/10.1083/jcb.201912074

Wu, L., Wang, L., Hu, S., Tang, G., Chen, J., Yi, Y., Xie, H., Lin, J., Wang, M., Wang, D., Yang, B., & Huang, Y. (2025). RNALocate v3.0: Advancing the Repository of RNA Subcellular Localization with Dynamic Analysis and Prediction. *Nucleic Acids Research*, *53*(D1), D284–D292. https://doi.org/10.1093/nar/gkae872

Wu, Y., Ianakiev, K., & Govindaraju, V. (2002). Improved k-nearest neighbor classification. *Pattern Recognition*. https://doi.org/10.1016/S0031-3203(01)00132-7

Xu, M., Chen, Y., Xu, Z., Zhang, L., Jiang, H., & Pian, C. (2022). MiRLoc: predicting miRNA subcellular localization by incorporating miRNA–mRNA interactions and mRNA subcellular localization. *Briefings in Bioinformatics*, *23*(2). https://doi.org/10.1093/bib/bbac044

Yadav, R., Singh, A. V., Kushwaha, S., & Chauhan, D. S. (2024). Emerging role of exosomes as a liquid biopsy tool for diagnosis, prognosis & monitoring treatment

response of communicable & non-communicable diseases. *The Indian Journal of Medical Research*, *159*(2), 163–180. https://doi.org/10.4103/ijmr.ijmr_2344_22

Yakob, M., Fuentes, L., Wang, M. B., Abemayor, E., & Wong, D. T. W. (2014). Salivary biomarkers for detection of oral squamous cell carcinoma - current state and recent advances. *Current Oral Health Reports*, *1*(2), 133–141. https://doi.org/10.1007/s40496-014-0014-y

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., & Kanehisa, M. (2008). Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, *24*(13), i232–i240. https://doi.org/10.1093/bioinformatics/btn162

Yáñez-Mó, M., Siljander, P. R.-M., Andreu, Z., Zavec, A. B., Borràs, F. E., Buzas, E. I., Buzas, K., Casal, E., Cappello, F., Carvalho, J., Colás, E., Cordeiro-da Silva, A., Fais, S., Falcon-Perez, J. M., Ghobrial, I. M., Giebel, B., Gimona, M., Graner, M., Gursel, I., … De Wever, O. (2015). Biological properties of extracellular vesicles and their physiological functions. *Journal of Extracellular Vesicles*, *4*, 27066. https://doi.org/10.3402/jev.v4.27066

Yang, F., Tian, C., Zhou, L., Guan, T., Chen, G., Zheng, Y., & Cao, Z. (2024). The value of urinary exosomal <scp>microRNA</scp> -21 in the early diagnosis and prognosis of bladder cancer. *The Kaohsiung Journal of Medical Sciences*, *40*(7), 660–670. https://doi.org/10.1002/kjm2.12845

Yang, F.-K., Tian, C., Zhou, L.-X., Guan, T.-Y., Chen, G.-L., Zheng, Y.-Y., & Cao, Z.-G. (2024). The value of urinary exosomal microRNA-21 in the early diagnosis and prognosis of bladder cancer. *The Kaohsiung Journal of Medical Sciences*, *40*(7), 660–670. https://doi.org/10.1002/kjm2.12845

Yang, K. S., Im, H., Hong, S., Pergolini, I., del Castillo, A. F., Wang, R., Clardy, S., Huang, C.-H., Pille, C., Ferrone, S., Yang, R., Castro, C. M., Lee, H., del Castillo, C. F., & Weissleder, R. (2017). Multiparametric plasma EV profiling facilitates diagnosis of pancreatic malignancy. *Science Translational Medicine*, *9*(391). https://doi.org/10.1126/scitranslmed.aal3226

Yekula, A., Muralidharan, K., Kang, K. M., Wang, L., Balaj, L., & Carter, B. S. (2020). From laboratory to clinic: Translation of extracellular vesicle based cancer biomarkers. *Methods (San Diego, Calif.)*, *177*, 58–66. https://doi.org/10.1016/j.ymeth.2020.02.003

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. https://doi.org/10.3389/fonc.2022.966981

Yoshioka, Y., Kosaka, N., Konishi, Y., Ohta, H., Okamoto, H., Sonoda, H., Nonaka, R., Yamamoto, H., Ishii, H., Mori, M., Furuta, K., Nakajima, T., Hayashi, H., Sugisaki, H., Higashimoto, H., Kato, T., Takeshita, F., & Ochiya, T. (2014). Ultra-sensitive liquid biopsy of circulating extracellular vesicles using ExoScreen. *Nature Communications*, *5*(1), 3591. https://doi.org/10.1038/ncomms4591

Yoshizawa, J. M., Schafer, C. A., Schafer, J. J., Farrell, J. J., Paster, B. J., & Wong, D. T. W. (2013a). Salivary biomarkers: toward future clinical and diagnostic utilities. *Clinical Microbiology Reviews*, *26*(4), 781–791. https://doi.org/10.1128/CMR.00021-13

Yoshizawa, J. M., Schafer, C. A., Schafer, J. J., Farrell, J. J., Paster, B. J., & Wong, D. T. W. (2013b). Salivary biomarkers: Toward future clinical and diagnostic utilities. In *Clinical Microbiology Reviews*. https://doi.org/10.1128/CMR.00021-13

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*(4), 574–578. https://doi.org/10.1016/j.peptides.2009.12.026

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*(4), 466–477. https://doi.org/10.1016/j.annonc.2021.01.074

Zeng, M., Wu, Y., Li, Y., Yin, R., Lu, C., Duan, J., & Li, M. (2023). LncLocFormer: a Transformer-based deep learning model for multi-label lncRNA subcellular localization prediction by using localization-specific attention mechanism. *Bioinformatics*, *39*(12). https://doi.org/10.1093/bioinformatics/btad752

Zhang, J., Li, S., Li, L., Li, M., Guo, C., Yao, J., & Mi, S. (2015). Exosome and exosomal microRNA: trafficking, sorting, and function. *Genomics, Proteomics & Bioinformatics*, *13*(1), 17–24. https://doi.org/10.1016/j.gpb.2015.02.001

Zhang, L., Xiao, H., Karlan, S., Zhou, H., Gross, J., Elashoff, D., Akin, D., Yan, X., Chia, D., Karlan, B., & Wong, D. T. (2010). Discovery and preclinical validation of salivary transcriptomic and proteomic biomarkers for the non-invasive detection of breast cancer. *PloS One*, *5*(12), e15573. https://doi.org/10.1371/journal.pone.0015573

Zhang, L., Xiao, H., Zhou, H., Santiago, S., Lee, J. M., Garon, E. B., Yang, J., Brinkmann, O., Yan, X., Akin, D., Chia, D., Elashoff, D., Park, N.-H., & Wong, D. T. W. (2012). Development of transcriptomic biomarker signature in human saliva to detect lung cancer. *Cellular and Molecular Life Sciences*, *69*(19), 3341–3350. https://doi.org/10.1007/s00018-012-1027-0

Zhang, X., Yuan, X., Shi, H., Wu, L., Qian, H., & Xu, W. (2015). Exosomes in cancer: small particle, big player. *Journal of Hematology & Oncology*, *8*, 83. https://doi.org/10.1186/s13045-015-0181-x

Zhang, Z.-Y., Zhang, Z., Ye, X., Sakurai, T., & Lin, H. (2024). A BERT-based model for the prediction of lncRNA subcellular localization in Homo sapiens. *International Journal of Biological Macromolecules*, *265*, 130659. https://doi.org/10.1016/j.ijbiomac.2024.130659

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*(1), 19448. https://doi.org/10.1038/s41598-019-55351-z

Zheng, D., Huo, M., Li, B., Wang, W., Piao, H., Wang, Y., Zhu, Z., Li, D., Wang, T., & Liu, K. (2021). The Role of Exosomes and Exosomal MicroRNA in Cardiovascular Disease. *Frontiers in Cell and Developmental Biology*, *8*. https://doi.org/10.3389/fcell.2020.616161

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*(1), 144. https://doi.org/10.1038/s41392-020-00258-9

Zhou, X., Zhu, W., Li, H., Wen, W., Cheng, W., Wang, F., Wu, Y., Qi, L., Fan, Y., Chen, Y., Ding, Y., Xu, J., Qian, J., Huang, Z., Wang, T., Zhu, D., Shu, Y., & Liu, P. (2015). Diagnostic value of a plasma microRNA signature in gastric cancer: a microRNA expression analysis. *Scientific Reports*, *5*, 11251. https://doi.org/10.1038/srep11251

Zhu, L., Qin, J., Wang, J., Guo, T., Wang, Z., & Yang, J. (2016). Early Gastric Cancer: Current Advances of Endoscopic Diagnosis and Treatment. *Gastroenterology Research and Practice*, *2016*, 9638041. https://doi.org/10.1155/2016/9638041