

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Computer Science



Sequential homology of circular RNA

Master Thesis

Bsc. Eliška Sirůčková

Study programme: Open Informatics
Specialisation: Bioinformatics
Supervisor: Bc. Petr Ryšavý, MSc.

Prague, May 2021

Thesis Supervisor:

Bc. Petr Ryšavý, MSc.
Intelligent Data Analysis
Faculty of Electrical Engineering
Czech Technical University in Prague
Technická 2
160 00 Prague 6
Czech Republic

Declaration

I hereby declare that I have completed this thesis on my own and that all the used sources are included in the list of references, in accordance with the *Methodological instructions on ethical principles in the preparation of university theses*.

In Prague, May 21st, 2021

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s *Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací*.

V Praze dne 21. 5. 2021

.....
Bsc. Eliška Sirůčková

I. Personal and study details

Student's name: **Sirůčková Eliška** Personal ID number: **474568**
Faculty / Institute: **Faculty of Electrical Engineering**
Department / Institute: **Department of Computer Science**
Study program: **Open Informatics**
Specialisation: **Bioinformatics**

II. Master's thesis details

Master's thesis title in English:

Sequential homology of circular RNA

Master's thesis title in Czech:

Sekvenční homologie cirkulární RNA

Guidelines:

1. Sestrojte přehled toho, jak circRNA přispívá k regulaci v lidských buňkách. Zaměřte se především na situaci, kdy circRNA funguje jakožto konkurenční endogenní RNA a utlumuje aktivitu miRNA.
 2. Proveďte rešerši současné literatury a nástrojů, které se zabývají predikcí miRNA-circRNA interakcí. Rešerše by měla zahrnovat TargetScan [5] jakožto obecný nástroj pro predikci miRNA cílů a podobné metody (např. miRnada či RNAhybrid [6]). Zahrňte relevantní databáze se známými a predikovanými interakcemi (např. CircInteractome [3] či StarBase [7]).
 3. Navrhněte vlastní způsob predikce circRNA-miRNA interakcí.
 4. Navrženou metodu otestujte a porovnejte s referenčními metodami.
1. Provide a review of how circRNA contributes to regulatory mechanisms in a human cell. Pay special attention to the effect called miRNA sponging when competing endogenous RNA inhibits miRNA activity.
 2. Write a review of the current literature and tools that deal with the prediction of miRNA-circRNA interactions. The study should include TargetScan [5] as a general tool for miRNA target predictions and similar methods (such as miRanda or RNAhybrid [6]). Include relevant databases with known and predicted interactions (for example, CircInteractome [3] or StarBase [7]).
 3. Propose your method to predict circRNA-miRNA interactions.
 4. Test this method and compare it with other reference tools.

Bibliography / sources:

- [1] Mitra, A., Pfeifer, K., & Park, K. S. (2018). Circular RNAs and competing endogenous RNA (ceRNA) networks. *Translational cancer research*, 7(Suppl 5), S624–S628. <https://doi.org/10.21037/tcr.2018.05.12>
- [2] Lin, YC., Lee, YC., Chang, KL. et al. Analysis of common targets for circular RNAs. *BMC Bioinformatics* 20, 372 (2019). <https://doi.org/10.1186/s12859-019-2966-3>
- [3] Dawood B. Dudekula, Amaresh C. Panda, Ioannis Grammatikakis, Supriyo De, Kotb Abdelmohsen & Myriam Gorospe (2016) CircInteractome: A web tool for exploring circular RNAs and their interacting proteins and microRNAs, *RNA Biology*, 13:1, 34-42, DOI: 10.1080/15476286.2015.1128065
- [4] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, Anton J. Enright, miRBase: tools for microRNA genomics, *Nucleic Acids Research*, Volume 36, Issue suppl_1, 1 January 2008, Pages D154–D158, <https://doi.org/10.1093/nar/gkm952>
- [5] Agarwal, V., Bell, G. W., Nam, J. W., & Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *elife*, 4, e05005. <https://elifesciences.org/articles/05005>
- [6] Jan Krüger, Marc Rehmsmeier, RNAhybrid: microRNA target prediction easy, fast and flexible, *Nucleic Acids Research*, Volume 34, Issue suppl_2, 1 July 2006, Pages W451–W454, <https://doi.org/10.1093/nar/gkl243>
- [7] Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, Jian-Hua Yang, starBase v2.0:

decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data, *Nucleic Acids Research*, Volume 42, Issue D1, 1 January 2014, Pages D92–D97, <https://doi.org/10.1093/nar/gkt1248>

Name and workplace of master's thesis supervisor:

Bc. Petr Ryšavý, MSc., Intelligent Data Analysis, FEE

Name and workplace of second master's thesis supervisor or consultant:

Date of master's thesis assignment: **11.02.2021** Deadline for master's thesis submission: **21.05.2021**

Assignment valid until: **30.09.2022**

Bc. Petr Ryšavý, MSc.
Supervisor's signature

Head of department's signature

prof. Mgr. Petr Páta, Ph.D.
Dean's signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

Date of assignment receipt

Student's signature

Abstract / Abstrakt

Circular RNAs (circRNAs) are believed to play an important role in cellular functions via interactions with micro RNAs involved in regulation of gene expression. miRNA-circRNA interactions are rarely experimentally validated, so researchers rely on their predictions. The commonly used prediction tools have been developed for and mostly tailored to the much better understood and researched miRNA-mRNA interactions. The aim of this thesis was to design a novel approach to the analysis of available data to improve the identification of miRNA-circRNA interaction sites specifically. The focus was on the use of the secondary structure of circRNA as a novel parameter in prediction using machine learning. Several classifiers were rigorously tested to select the best representation and classifier combination. A novel approach to pairing prediction was also taken. Based on 15 nt of the primary sequence of circRNA, secondary structure-based classifiers and a simple neural network, the interactions were predicted with a perfect recall. Finally, to improve the low precision of the proposed method, an ensemble of existing tools was proposed, improving the prediction and recall of the combined ensemble beyond the capabilities of each individual tool.

Keywords: circRNA, miRNA, miRNA-circRNA interactions, seed sites, circRNA secondary structure, software, bioinformatics

Cirkulární RNA (cirkRNA) hrají podle dosavadních výzkumů důležitou roli v buňce skrze interakce s miRNA. Interakce miRNA-cirkRNA jsou zřídka experimentálně ověřené, takže se vědci spoléhají na jejich predikce. Běžně používané predikční nástroje byly vyvinuty a většinou přizpůsobeny pro mnohem lépe pochopené a prozkoumané interakce miRNA-mRNA. Cílem této práce bylo vytvořit nový přístup k dostupným datům, který by zlepšil identifikaci míst interakce konkrétně mezi miRNA a cirkRNA. Důraz byl kladen na použití sekundární struktury cirkRNA jako nového parametru v predikci pomocí strojového učení. Několik klasifikátorů bylo testováno, za účelem vybrání nejlepší kombinace reprezentace a klasifikátoru. Byl také použit nový přístup k predikci párování. Na základě 15ti nuklotidů primární sekvence cirkRNA, klasifikátorů založených na sekundární struktuře a pomocí jednoduché neurální sítě byly předpovězeny interakce s dokonalou výtěžností. Nakonec, aby se zlepšila nízká přesnost navrhované metody, byla navržena kombinace navrhované metody se existujícími nástroji, která zlepšila predikci a výtěžnost konečného kombinace metod nad možnosti metod braných jednotlivě.

Klíčová slova: cirkRNA, miRNA, miRNA-cirkRNA interakce, seed oblasti, sekundární struktura circRNA, software, bioinformatika

Acknowledgements

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would like to thank my supervisor, Msc. Petr Ryšavý, whose expertise was invaluable in solving the task ahead of me. Your feedback and never-ending optimism during times when my work did not head the way I planned was invaluable.

I would also like to thank my family for their wise counsel and continuous support throughout my education.

Finally, I could not have completed this thesis without Lukáš Fanta, who provided me with stimulating discussions and coding insights, and eased my mind during tough times.

Contents

Abstract	vii
Acknowledgements	viii
List of Tables	xii
List of Figures	xiv
List of Acronyms	xv
1 Introduction	1
1.1 Proposed Solution	2
1.2 Thesis Structure	3
2 Biological Background	4
2.1 Pre-mRNA and mRNA	4
2.2 miRNA	4
2.3 circRNA structure	5
2.4 circRNA functions	5
2.5 Secondary structures of RNA molecules	7
2.6 mRNA-miRNA-circRNA binding	8
3 Existing Tools - Databases	11
3.1 circRNADb	11
3.2 circBase	12
3.3 ENCORI: StarBase	12
3.4 CircInteractome	12
3.5 CircFunBase	13
3.6 CSCD - Cancer-Specific CircRNA Database	13
3.7 CIRCpedia	13
4 Existing Tools - Computational Tools	14
4.1 TargetScan	14
4.2 miRanda	16
4.3 RNAhybrid	17
4.4 Ensemble of TargetScan, RNAhybrid and miRanda	18

5	Technical Background	20
5.1	RNAfold	20
5.1.1	Dot-Bracket and Pseudo-Bracket Notations	21
5.1.1.1	RNAfold output example	21
5.2	Alignment (Biopython)	21
5.2.1	Substitution matrix	22
5.3	Classifiers Used with Secondary Structures	22
5.3.1	k-Nearest Neighbours Classifier	22
5.3.2	Decision Tree Classifier	22
5.3.3	Random Forest	23
5.3.4	AdaBoost	23
5.3.5	SVC	23
5.3.6	Complement Naive Bayes	24
5.4	Analysis Tools	24
5.4.1	PCA and ICA	24
5.4.2	GridSearch	25
5.5	MLP Classifier for prediction of interactions	25
5.6	Evaluation	26
5.6.1	TP,TN, FP,FN and confusion matrix	26
5.6.2	SE, SP, gmean	26
5.6.3	Precision, Recall and F1-score	27
6	Implementation	28
6.1	Data Extraction and Analysis	28
6.1.1	Obtaining Data	28
6.1.2	Creating datasets from secondary structure sequences	31
6.1.2.1	Create dataset from alignments	31
6.1.2.2	Labelling	32
6.2	Classification Based on Secondary Structures	34
6.3	Classification and Prediction of Interactions	35
6.4	Comparison with Reference Tools	36
7	Results	39
7.1	Data analysis	39
7.2	Secondary Structure-Based Classifiers	42
7.2.1	Initial Experiments	42
7.2.2	GridSearch	43
7.3	Classification of Interactions	44
7.4	Comparison with Reference Tools	45
7.4.1	Ensemble proposition	48
8	Evaluation and Discussion	49
8.1	Data Analysis	49
8.2	Secondary Structure-Based Classifiers	49
8.3	Classification of Interactions	50
8.4	Reference Tools Comparison and Ensemble Proposition	51
8.5	Discussion	52
8.5.1	Possible Improvements and Future Work	52

<i>CONTENTS</i>	xi
9 Conclusion	54
A	55
A.1 Data Analysis	55
A.1.1 PCA and ICA	55
A.2 Secondary Structure Based Classifiers	63
A.2.1 Initial experiments	63
A.2.2 GridSearch	80
A.3 Results for MLPClassifiers	93
B	95
B.1 CD contents	95
B.2 Used Graphical Programs	95
Bibliography	96

List of Tables

4.1	miRanda Parameter Scores	17
5.1	Pseudo bracket notation symbols explained.[1]	21
5.2	Substitution matrix for Watson-Crick pairing including G:U wobble pairs.[2]	22
5.3	Activation functions available for MLP Classifier by scikit-learn [3].	26
6.1	Datasets for classification based on secondary structures	32
6.2	An example of Initial experiments with <i>2Categs15Centroids</i>	36
6.3	Example of SVC GridSearch results for <i>2Categs15More</i> dataset	37
6.4	Best Secondary Structure-Based Classifiers	37
6.5	Experimental Setup for MLP Classifiers	37
6.6	Parameters of MLP classifiers	38
7.1	Initial Experiment 26. - <i>3Categs24</i> dataset	42
7.2	The best results obtained for each classifier.	43
7.3	The best MLP and its results for each experiment.	44
7.4	Results for MLPs based on pairing only.	44
7.5	Comparison with Reference Tools	45
A.1	Initial Experiment 1 - <i>2Categs15Centroids</i>	63
A.2	Initial Experiment 2 - <i>2Categs15Centroids</i>	64
A.3	Initial Experiment 3 - <i>2Categs15Centroids</i>	64
A.4	Initial Experiment 4 - <i>2Categs15Centroids</i>	65
A.5	Initial Experiment 5 - <i>2Categs15Less</i>	65
A.6	Initial Experiment 6 - <i>2Categs15Less</i>	66
A.7	Initial Experiment 7 - <i>2Categs15Less</i>	66
A.8	Initial Experiment 8 - <i>2Categs15Less</i>	67
A.9	Initial Experiment 9 - <i>2Categs15MFE</i>	67
A.10	Initial Experiment 10 - <i>2Categs15MFE</i>	68
A.11	Initial Experiment 11 - <i>2Categs15MFE</i>	68
A.12	Initial Experiment 12 - <i>2Categs15MFE</i>	69
A.13	Initial Experiment 13 - <i>2Categs15More</i>	69
A.14	Initial Experiment 14 - <i>2Categs15More</i>	70
A.15	Initial Experiment 15 - <i>2Categs15More</i>	70
A.16	Initial Experiment 16 - <i>2Categs15More</i>	71
A.17	Initial Experiment 17 - <i>3Categs15</i>	71
A.18	Initial Experiment 18 - <i>3Categs15</i>	72
A.19	Initial Experiment 19 - <i>3Categs15</i>	72
A.20	Initial Experiment 20 - <i>3Categs15</i>	73

A.21 Initial Experiment 21 - <i>5Categs15</i>	73
A.22 Initial Experiment 22 - <i>5Categs15</i>	74
A.23 Initial Experiment 23 - <i>5Categs15</i>	74
A.24 Initial Experiment 24 - <i>5Categs15</i>	75
A.25 Initial Experiment 25 - <i>3Categs24</i>	75
A.26 Initial Experiment 26 - <i>3Categs24</i>	76
A.27 Initial Experiment 27 - <i>3Categs24</i>	76
A.28 Initial Experiment 28 - <i>3Categs24</i>	77
A.29 Initial Experiment 29 - <i>5Categs24</i>	77
A.30 Initial Experiment 30 - <i>5Categs24</i>	78
A.31 Initial Experiment 31 - <i>5Categs24</i>	78
A.32 Initial Experiment 32 - <i>5Categs24</i>	79
A.33 GridSearch: SVC <i>2Categs15Centroids</i>	80
A.34 GridSearch: SVC <i>2Categs15Less</i>	80
A.35 GridSearch: SVC <i>2Categs15MFE</i>	81
A.36 GridSearch: SVC <i>2Categs15More</i>	81
A.37 GridSearch: Adaboost(DT) <i>2Categs15Centroids</i>	82
A.38 GridSearch: Adaboost(DT) <i>2Categs15Less</i>	83
A.39 GridSearch: Adaboost(DT) <i>2Categs15MFE</i>	84
A.40 GridSearch: Adaboost(DT) <i>2Categs15More</i>	85
A.41 GridSearch: Decision Tree <i>2Categs15Centroids</i>	86
A.42 GridSearch: Decision Tree <i>2Categs15Less</i>	86
A.43 GridSearch: Decision Tree <i>2Categs15MFE</i>	87
A.44 GridSearch: Decision Tree <i>2Categs15More</i>	87
A.45 GridSearch: k-NN <i>2Categs15Centroids</i>	87
A.46 GridSearch: k-NN <i>2Categs15Less</i>	88
A.47 GridSearch: k-NN <i>2Categs15MFE</i>	88
A.48 GridSearch: k-NN <i>2Categs15More</i>	88
A.49 GridSearch: Random Forest <i>2Categs15Centroids</i>	89
A.50 GridSearch: Random Forest <i>2Categs15Less</i>	90
A.51 GridSearch: Random Forest <i>2Categs15MFE</i>	91
A.52 GridSearch: Random Forest <i>2Categs15More</i>	92
A.53 Results of MLPs for <i>Experiment 1</i>	93
A.54 Results of MLPs for <i>Experiment 2</i>	93
A.55 Results of MLPs for <i>Experiment 3</i>	93
A.56 Results of MLPs for <i>Experiment 4</i>	94

List of Figures

2.1	Back-splicing (reprint from [4])	5
2.2	miRNA Sponging	6
2.3	miRNA interactions	7
2.4	miRNA secondary structures (reprint from [5])	8
2.5	miRNA pairing motifs (reprint from [6])	9
4.1	26 considered features, 14 highlighted selected features (reprint from [7]) .	16
5.1	Diagram of a neural network	25
6.1	Diagram showing the processes and relationships of proposed solution . . .	29
6.2	CircInteractome website example	30
6.3	Process of labelling secondary structure subsequences.	34
7.1	PCA and ICA for <i>2Categs15Less</i> dataset	40
7.2	PCA and ICA for <i>3Categs15</i> dataset	41
7.3	Venn Diagrams	46
7.4	Venn diagram for ensemble based on majority-voting of proposed method, TargetScan and RNAhybrid with seed.	48
A.1	PCA and ICA for <i>2Categs15Centroids</i> dataset	55
A.2	PCA and ICA for <i>2Categs15Less</i> dataset	56
A.3	PCA and ICA for <i>2Categs15MFE</i> dataset	57
A.4	PCA and ICA for <i>2Categs15More</i> dataset	58
A.5	PCA and ICA for <i>3Categs15</i> dataset	59
A.6	PCA and ICA for <i>3Categs24</i> dataset	60
A.7	PCA and ICA for <i>5Categs15</i> dataset	61
A.8	PCA and ICA for <i>5Categs24</i> dataset	62

List of Acronyms

- circRNA** circular RNA. 5
- CNB** Complement Naive Bayes. 23
- DT** Decision Tree Classifier. 21
- Gm** Geometric mean. 26
- ICA** Independent Component Analysis. 23
- k-NN** k-Nearest Neighbours. 21
- MREs** miRNA-response elements. 6
- mRNA** messenger RNA. 4
- PCA** Principal Component Analysis. 23
- pre-mRNA** precursor messenger RNA. 4
- RBPs** RNA-binding proteins. 6
- RF** Random Forest. 22
- SE** Sensitivity. 26
- SP** Specificity. 26
- SVC** C-Support Vector Classifier. 22
- 3'UTR** 3' untranslated region. 8
- nt** nucleotides. 20

Chapter 1

Introduction

CircRNAs have been known to biologists since the 1990s. Even though it has been around 30 years since their discovery, little is known about these molecules. The most interest has been focused on their role within cellular pathways, especially on the interactions with miRNA molecules. Having often the same primary structure as mRNAs, these molecules are believed to compete with mRNAs by sponging miRNAs, affecting the downstream pathways of mRNAs, including gene expression regulation. mRNAs are transcripts for protein synthesis. miRNAs are short RNA sequences that regulate activity of other RNAs such as mRNAs or circRNAs. Furthermore, circRNAs are often expressed in specific tissues, cell types or cancer, suggesting they have a significant function.

The traditional methods for studying RNA make it difficult to study circRNAs due to their similarity with mRNAs. Several independent attempts have been made to build a database, such as CircInteractome, of known circRNAs and their properties, leading to many potential sources with limited compatibility between them. In comparison, the tools for miRNA-circRNA specific interactions are scarce, so the tools designed for mRNA-miRNA interactions are usually used.

Although mRNA and circRNA have the same or very similar primary structure, two main differences should be considered when using miRNA-mRNA interaction predictors. First, mRNA is a linear molecule while circRNA is circular, suggesting that they form different secondary structures, which in turn defines their possible interactions with miRNA target subsequences. Second, tools for analysing miRNA-mRNA interactions tend to focus on the 3' untranslated region (3'UTR) of mRNA, whereas the miRNA-circRNA interactions have not been shown to have any strong preference in terms of specific binding regions.

Since no tools specific for the prediction of miRNA-circRNA interactions have been designed to date, new circRNA-tailored methods are needed. This thesis considers the secondary structure of circRNA. The secondary structure of mRNA has previously been

found to play a significant role in miRNA-mRNA interactions and has been utilised by TargetScan. The circRNA secondary structures are not experimentally known, which is the primary reason for them not being used in predictions. However, the secondary structures of circRNAs are expected to be different from the structures of miRNAs due to their circular form. A computational tool, such as RNAfold, can be used to predict the circular secondary structure, which has not been previously used to predict the miRNA-circRNA interactions.

Furthermore, no tool currently used for miRNA-circRNA interaction prediction uses machine learning or neural networks. Currently, the tools used (TargetScan Section 4.1, RNAhybrid Section 4.3) are based on series of assumptions made by observation of known interactions, especially of miRNA-mRNA. As a result, the proposed solution tries to teach machine learning classifiers the difference between interacting and non-interacting sequences while considering the secondary structure of given subsequences.

1.1 Proposed Solution

This thesis aims to implement and evaluate the use of secondary structure of circRNA in miRNA-circRNA interactions predictions. The task is split into four sections:

- data extraction and analysis Section 6.1,
- classification based on secondary structures,
- classification and prediction of interactions,
- evaluation and comparison with currently available tools.

In Section 6.1, available data will be extracted from various sources, and several datasets of different properties will be formed. The positive and negative class will be defined, and the samples will be split accordingly. The datasets will be formed by subsequences of the secondary structures of circRNAs. They will be further analysed for their potential to distinguish between positive and negative class using ICA and PCA.

In Section 6.2, all datasets with potential to be used in classification will be used with machine learning classifiers. The task will be to determine which datasets and which classifiers are most suitable for the prediction of the availability for complementary pairing with miRNA using the secondary structure information. The output of this section will be the best three combinations of classifiers and datasets.

In Section 6.3, a neural network will be created to evaluate the likelihood of a subsequence interacting with a miRNA. The input will include binary representations of

pairings between subsequences of individual circRNAs and selected miRNAs. Here only one dataset will be used to train the neural network consisting of all known combinations of circRNAs and miRNAs in humans. The outputs of classifiers from the previous section and the pairing data will extend the input to the neural network.

Last but not least, in Section 6.4, the results of previous experiments will be evaluated and compared to results of other predictive tools for miRNA-circRNA interactions. An ensemble combining the proposed method and two reference tools is also proposed (see Section 7.4.1).

1.2 Thesis Structure

This chapter introduces the overall topic and outlines the concept of the proposed solution. In Chapter 2, the biological concepts along with the importance of the issues are defined, while offering a deeper insight into the options that are currently available for new potential methods. Chapters 3 and 4 describe the current state of the art databases and prediction tools, respectively.

Next, Chapter 5, describes the algorithms, tools and methods that have been used for the implementation of the proposed solution. The implementation itself is described in Chapter 6 with the results discussed in Chapter 7. Both Chapter 6 and 7 follow the split of tasks outlined in Section 1.1.

Chapter 8, contains the evaluation and discussion of the results obtained, while the last Chapter 9 concludes the thesis and suggests possible improvements and further research.

Chapter 2

Biological Background

2.1 Pre-mRNA and mRNA

DNA is known to encode information required for protein synthesis and protein synthesis regulation. This information is relayed through transcription into various types of RNA. One such type is the precursor messenger RNA (pre-mRNA) which is a raw RNA transcript of a gene encoded by the DNA. As the name suggests, the pre-mRNA is a precursor of mRNA. The pre-mRNA consists of protein-coding and non-coding parts of a gene called exons and introns, respectively. To form messenger RNA (mRNA), the pre-mRNA can be spliced by a spliceosome, a large ribonucleoprotein complex. During this process the introns are removed and the exons are fused together to form mature mRNA, the final template for protein synthesis via translation.

2.2 miRNA

MiRNAs are a group of non-coding functional RNA molecules. Human cells are estimated to contain around 2500 different human miRNAs, and more have been identified in other mammals [8]. These 20-25 nucleotides long sequences [9], processed from a longer mRNA precursor are responsible for the regulation of up to 90% of human genes involved in proliferation, cell growth, cellular signalling, embryonic development, tissue differentiation and apoptosis. Varying expression patterns of miRNAs can be observed in different tissues at different times, such as during development, disease or in response to treatment [10]. MiRNA molecules are incorporated into a miRNA-induced silencing complex (RISC, details in Section 2.6) which then specifically targets mRNA molecules via complementary base-pairing to 5-7 nucleotides (a seed region), preventing translational machinery from translating the mRNA sequence into proteins, and/or marking the mRNA for degradation, resulting in efficient down-regulation of gene expression. Aberrant expressions of

miRNAs have been linked to various diseases and documented in the Human microRNA Disease Database (HMDD) [11].

2.3 circRNA structure

CircRNA is a continuous loop of single-stranded RNA. It is produced by back splice junction (back-splicing) of pre-mRNA (Figure 2.1) [4, 6] where the 5' and 3' ends of the linear transcript are covalently fused together, by the same spliceosome machinery that generates mRNA, to form a circular structure containing one or more exons [12]. CircRNAs, unlike linear mRNAs, lack terminal structures (5' caps and poly(A) tails) and are therefore resistant to exonucleases such as RNase R [13]. RNase R degrades mRNA and other linear RNAs, purifying circular RNA (circRNA) for subsequent sequencing. Because circRNAs lack the terminal structures, they are more stable *in vivo* than their linear counterparts. Furthermore, the levels of circRNA in different tissues in a single organism can vary, suggesting that circRNA may have tissue-specific functionality. The two properties make circRNAs interesting candidates for biomarkers for diseases such as cancer [14].

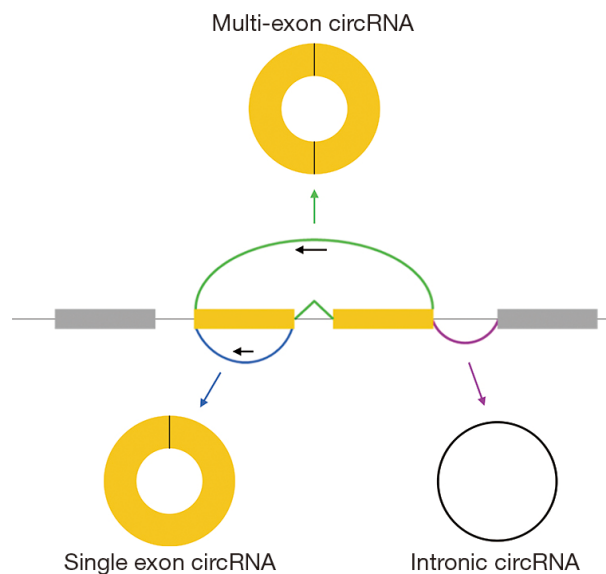


Figure 2.1: Back-splicing (reprint from [4])

2.4 circRNA functions

Although the exact function of most circRNAs is unknown, the structure of the few that are known can be used to predict the functions of the rest. About 40% of circRNAs contain the AUG start codon, a sequence marking an initiation point of synthesis of small

proteins. At least one such natural protein is synthesised this way [15]. Also, it has been shown that artificial circRNAs can interact with ribosomes, cell organelles that facilitate protein synthesis [16].

However, studies have shown that most circRNAs do not serve as a template for protein synthesis. Instead, the scientific focus has been aimed at their interaction with other RNAs and their role in the regulation of gene expression. Recent studies suggest that long non-coding RNAs are regulating each other and also regulating the amount of coding transcripts present in the cell via miRNAs [17]. A coding sequence, such as circRNA and mRNA, can have multiple miRNA binding sites and thus can be regulated by multiple different miRNAs. CircRNAs carry miRNA-response elements (MREs), sequences complementary to target miRNA, that allow circRNA to bind miRNA molecules and thereby reduce the availability of the target miRNA to bind and down-regulate protein-coding mRNA. This phenomenon (Figure 2.2) [18], called miRNA sponging, can also be observed with some other RNAs (Figure 2.3) [4]. The sponge effect is, however, only effective if there is a large amount of circRNA competing for the miRNA with mRNA template.

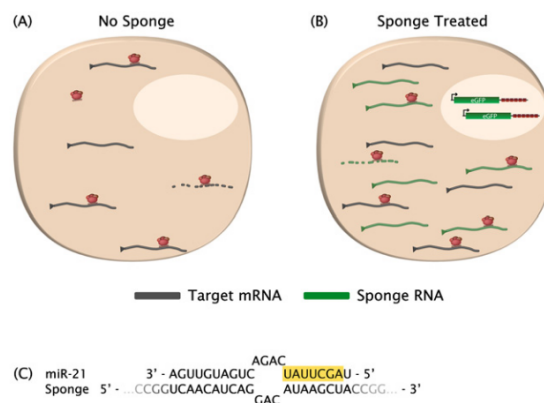


Figure 2.2: miRNA Sponging. Presence of sponge (circRNA) in (B) show a reduced amount of miRNA (red) available for binding with mRNA compared to binding in (A) (reprint from [18])

Some circRNAs also have binding sites for RNA-binding proteins (RBPs) and can therefore exert the same sponge effect on them. Recent cancer studies report that circRNAs play a role in epigenetic regulation via controlling the RNA splicing or transcription [15, 19, 20]. However, these other functions have not been studied as well as miRNA sponging. Reviews by Su et al. (2019) [21] and Zhao et al. (2019) [22] elaborate on this topic further.

As a circRNA can have multiple binding sites for various miRNAs, so can coding mRNAs. This fact creates a complex network of interactions too extensive for us to fully

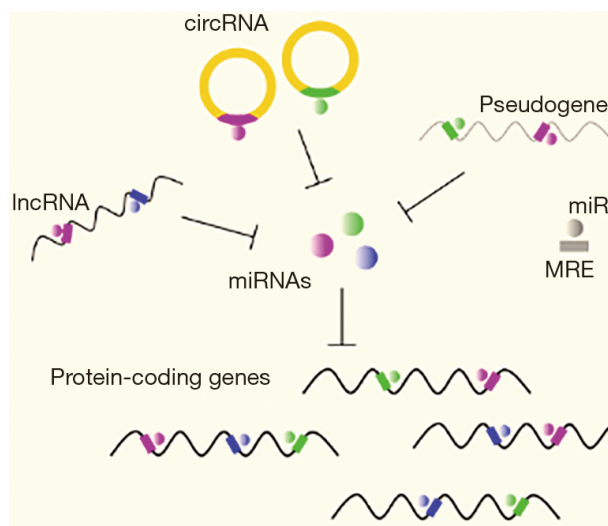


Figure 2.3: miRNA interactions. miRNAs can interact with various molecules. CircRNA, lncRNA and pseudogenes inhibit the miRNAs while miRNAs inhibit the protein-coding genes. When miRNAs are inhibited they cannot repress the protein coding genes (reprint from [4])

understand yet, however, using various computational tools for circRNA analysis may offer a starting point for navigating it.

2.5 Secondary structures of RNA molecules

The mRNA and circRNA forms a local secondary structure through binding of complementary parts of its sequence via Watson-Crick pairing. This secondary structure consists of various loops, bulges and stems and affects the ability of certain parts of the RNA to bind to other molecules by rendering the necessary bases inaccessible for binding. The main difference in secondary structure between mRNA and circRNA is that mRNA is linear and so has two “free” ends, whereas circRNA is circular. This is a significant difference, especially when trying to predict secondary structure computationally.

The inaccessibility of the binding site can also be caused by pseudoknots. The sites are primarily inaccessible because the energy required to break the existing bonds would be insufficiently compensated by new potential bonds with miRNA [23]. Thus, unstructured regions of RNA, accessible for binding, are favoured and efficiently regulated by miRNA.

Similar to mRNAs and circRNAs, the miRNAs form secondary structures. By adopting homo-duplex and/or hairpin structures [24]. Hairpin structure consists of a self-complementary stem and loop of free nucleotides and possibly free 3’ and/or 5’ ends. Homo-duplex refers to the Watson-Crick pairing with G:U wobble pairs between parts of two copies of miRNA (Figure 2.4). The transition between the two secondary structures is linked to the cellular environment, namely the concentration of miRNA in the cell and

the ionic conditions. When the concentration of miRNA is high, more homo-duplexes are observed. In low concentrations of miRNA the prevailing structure is hairpin [24, 5]. Depending on the secondary structure, the function of miRNA can be affected if the ‘seed’ region, important in miRNA-target interactions, is inaccessible in either of the two formations.

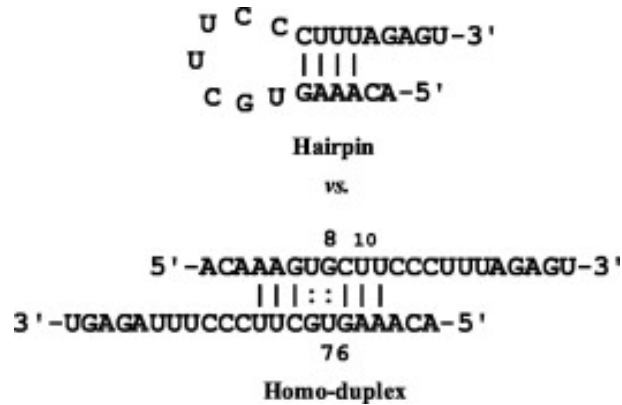


Figure 2.4: miRNA secondary structures (reprint from [5])

2.6 mRNA-miRNA-circRNA binding

In the majority of cases, the miRNAs are incorporated in the miRNA-induced silencing complex (RISC). This complex also includes a type of Argonaut protein, which exposes the ‘seed’ region of the miRNA that is thought to be essential in most miRNA-target interactions [25]. The ‘seed’ region refers to nucleotides at positions 2-8 in the 3’UTR of miRNA. Several classes of target sites have been identified based on conserved miRNA-pairing motifs [6] (also depicted in Figure 2.5):

- 8mer – complementary pairing with positions 2-8 of miRNA and with an A opposite the position 1
- 7mer-m8 – complementary pairing with positions 2-8 of miRNA
- 7mer-A1 – complementary pairing with positions 2-7 of miRNA and with an A opposite the position 1
- 6mer – complementary pairing with positions 2-7 of miRNA
- offset-6mer – position 3-8 match

A smaller number of 6 matching pairs in seed region can be compensated by supplementary base pairing with the remaining part of the miRNA but the evidence is observed

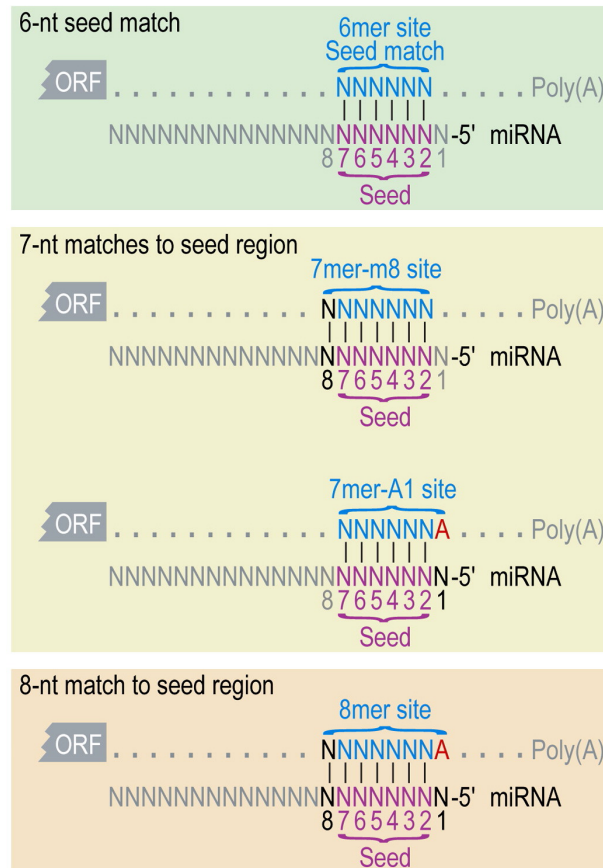


Figure 2.5: miRNA pairing motifs (reprint from [6])

in only 5% of the cases of mentioned pairing motifs [7]. A very extensive pairing with the 3' region of the miRNA sequence can completely substitute a lack of pairing in the seed region. However, it is sporadic to observe preferentially conserved pairing sites in the 3' region.

Additional Watson-Crick pairing at positions 12-17, especially nucleotides 13-16, further enhances targeting by miRNAs. Similar to the seed region, uninterrupted pairing at these positions by wobbles, bulges or other mismatches is preferred. These positions are the best conserved positions outside of the seed region.

The preferred nucleotides in the immediate vicinity of functional sites are A and U, even more for conserved sites. The experiments that lead to these findings were performed on the 3' UTR of mRNA because it was believed that the miRNA bound mRNA in this region only [6]. However, more recent findings have confirmed that miRNAs also bind in the 5'UTR and the open reading frames (ORFs) [26].

Although miRNAs are able to bind in other places than 3' UTRs, the sites in 3'UTR are more often effective at repressing mRNA than the rest of the options. Furthermore, having a high local A-U content and being further away from centres of long UTRs makes a site more likely to be effective [23]. The efficacy is also influenced by the distance to

neighbouring miRNA-binding sites and protein sites. Sites nearby (8 to 40 nucleotides) are cooperative, meaning the repression exerted is greater than the summation of two independent locations. The distance between could be greater, but the preference is up to 40 nucleotides. When two locations are less than eight nucleotides apart, they are competitive [6].

Articles by Grimson et al. (2007) [6] and Agarwal et al. (2015) [7] offer further insights into the topic.

Chapter 3

Existing Tools - Databases

CircRNA research generates large quantities of data because circRNA is a relatively new concept that fits in with many previously researched concepts. For this reason, there are several databases that store related data, including not only circRNA properties but also interactions with mRNA, miRNA or RBP. Different databases consider the relationships between circRNA expression and disease or other traits. Major drawback of large number of databases for circRNA is used of various genome assemblies, multiple naming systems and different approaches to numbering positions in miRNA-circRNA interactions.

Databases such as circNet and circ2Traits are mentioned in various articles and books [27] but are not available anymore. Most existing databases such as circBase (Section 3.2) or circRNADb (Section 3.1) are available but not recently updated, possibly because there were no more studies into circRNA that would be relevant for them. To fully explore data that are available for circRNA analysis, a database with information about miRNAs is needed. The best database for miRNAs is the miRBase (<http://www.mirbase.org/>). Further information about miRBase can be found in the most recent article by Kozomara et al. (2018)[28].

3.1 circRNADb

circRNADb (<http://reprod.njmu.edu.cn/circrnadb>)[29] is a database of circular RNAs identified in humans. As of May 2021, the database contains 32,914 annotated exonic circRNAs from large-scale studies attempting to identify human circRNAs [30, 31, 32, 33]. The database allows scrolling through all entries, browsing by Gene Symbol, Cell Type, PubMedID or Protein-coding potential, and searching by more advanced criteria including chromosome number or chromosome strand. The database also provides a downloadable version of the whole dataset. Although it does not seem to have been updated since 2016, the database includes the best transcript, other possible transcripts and information about

protein-coding potential of an explored circRNA.

3.2 circBase

CircBase (<http://www.circbase.org/>) is another circRNA database and, like circRNADb (Section 3.1), it has also not been updated since 2015. It is based on the same large-scale studies [32, 30, 31, 33] as circRNADb but provides different search options. CircBase also offers FASTA format exportation option for a given search result. To show individual sequences, the database redirects to a corresponding PubMed record. The database entries cannot be extracted all at once as the site only shows results of a narrowed-down search.

3.3 ENCORI: StarBase

The Encyclopedia of RNA Interactomes (ENCORI) also known as StarBase (starbase.sysu.edu.cn/)[34, 35] is a database of miRNA interactions including miRNA-circRNA and miRNA-mRNA. All entries in this database were subject to passing two constraints. First, the interaction between two miRNA and mRNA had to be predicted by at least 3 out of 5 computational tools (TargetScan[36, 6], miRanda[2], Pictar2 [37], PITA [38], and RNA22 [39]). The miRNA-circRNA interactions were only predicted using miRanda (Section 4.2). miRanda is the most general algorithm for identifying miRNA-mRNA and can therefore be easily used for identification of miRNA-circRNA interactions. Other tools are designed purely for miRNA-mRNA. Second, the interaction had to be at least once identified experimentally by sequencing of cross-linked immunoprecipitated RNA (CLIP-seqs) with Argonaute or other RBP protein, ensuring the findings are biologically relevant [40].

ENCORI database currently contains 149 miRNA-circRNA interactions for the human genome and 15 for the mouse genome. 11 circRNA-RBP interactions for mammals and 1 for the nematode. ENCORI has so far used over 700 CLIP-seq datasets. The database is connected to Genome Browser (<http://genome.ucsc.edu/>) and circBase (Section 3.2). Unfortunately, the database belonging to Sun Yat-sen University can be accessed by the public only in restricted hours.

3.4 CircInteractome

CircInteractome (circinteractome.nia.nih.gov/) is a database of predicted circRNA interactions with RBP and circRNA [41]. The database also stores information about individual circRNAs; however, circRNA entries can only be searched by name, gene symbol or

cell-line/tissue. Further, the information can only be extracted for each entry separately.

The interactions between miRNA and circRNA can be explored by entering miRNA or circRNA, or both. The results show the number of pairing sites that are shared between a particular pair of circRNA and miRNA. For each site, the expected pairing is shown alongside scores used to predict the site of interest. The scoring system devised by Grimson et al. [6] and generated by TargetScan is described in Section 4.1.

3.5 CircFunBase

CircFunBase (<http://bis.zju.edu.cn/CircFunBase/>) is a database for functional circular RNAs. This database contains 7000 circRNAs from various species including 3799 entries for human circRNAs. It essentially extracts or links information from other databases to allow users to access all information available on each circRNA from one database.

The database can be browsed by species or lineage, or searched by circRNA name, location, gene symbol or keyword. A detailed description of circRNA includes function, expression profile, sequence and most importantly predicted and experimentally validated interactions. The database adds links to miRBase and circBase to give further information on specific miRNA or circRNA, respectively.

3.6 CSCD - Cancer-Specific CircRNA Database

CSCD (<https://gb.whu.edu.cn/CSCD/>) [42] contains 272,152 cancer-specific circRNAs. The database contains annotations including predicted binding sites for miRNA (predicted using TargetScan [Section 4.1]).

3.7 CIRCpedia

CIRCpedia (<https://www.picb.ac.cn/rnomics/circpedia/>) [43] is a database containing 262,782 circRNA entries out of which 183,943 are human. The database provides general information about each circRNA, including information about conservation in human and mouse genomes.

Chapter 4

Existing Tools - Computational Tools

Due to their structure and similarity to mRNAs, circRNAs are very difficult to distinguish from mRNAs and therefore, it is also problematic to study circRNA experimentally. As a result, various computational tools emerged to help process available data and predict functional characteristics of individual circRNAs such as miRNA interactions, protein interactions, and the relation to diseases.

Currently, the only mediator of circRNA interactions with miRNA is the CircInteractome (Section 3.4), which uses TargetScan algorithm to identify potential interactions. TargetScan (Section 4.1) was initially created to identify miRNA-mRNA interactions, but its implementations worked on principles that were general enough to work with circRNA as well.

In comparison to miRNA-circRNA, there are more than 38 tools designed for the prediction of miRNA-mRNA interactions [44]. These tools can be split into two groups: heuristic and empirical. TargetScan, miRanda (Section 4.2) and RNAhybrid (Section 4.3) are all based on heuristic models. Empirical models generally use a machine learning-based approach, including the use of SVMs, decision trees, and artificial neural networks.

4.1 TargetScan

TargetScan (<http://www.targetscan.org/>) is one of the most common tools used for predicting miRNA-circRNA interactions. Since 2003 this tool has undergone many changes that eventually led to TargetScan v7.2 being currently the best tool for predicting target mRNA for miRNAs. However, as this tool has evolved, it focused more and more on mRNA-miRNA interactions, and the current version is no longer relevant to miRNA-circRNA predictions as these are treated as background noise.

Earlier versions of TargetScan focused on conserved seed-matches to the miRNA based on five genomes - human, mouse, rat, dog and chicken. The sites were scored based on

their context that included:

- type of seed-match,
- complementarity outside of the seed region,
- local A-U contribution,
- position contribution (distance to nearest UTR end of target);

and degree of conservation:

- highly conserved (across human, mouse, rat, dog, chicken),
- conserved (across human, mouse, rat, dog),
- poorly conserved (any other combination of species).

The initial assumption was that miRNA-target interactions with imperfect seed pairing require extensive pairing outside the seed, and therefore, would rarely occur because they would be challenging to conserve throughout evolution [36, 6].

Later, the probability of preferentially conserved targeting improved conservation scoring, increasing the number of miRNA target binding sites to >45,000 in human 3'UTRs [45]. The number of species was raised from 5 to 22. The differential ability of miRNA to repress mRNA was found to be linked to seed-pairing stability and high target-site abundance. TargetScan considers these two properties in its context-score model since 2011 [46].

The newest version implements many features specific to mRNAs but, with small adjustments, these could be applied to circRNAs. A list of the features selected, including a list of features considered, is summarized in Figure 4.1 [7].

Concerning mRNA-miRNA interactions, the most relevant improved feature is the predicted structural accessibility of the site. Previous versions only considered the accessibility of the binding site. When a score for probability of a region (14 nucleotides long centred on miRNA nucleotides 7 and 8) being unpaired was also considered, the results significantly improved.

When compared with 17 other miRNA target prediction tools and previous versions, TargetScan v7 performs the best in identifying miRNA-mRNA interaction sites. The newest version, however, is not tailored for prediction of miRNA-circRNA interaction sites. Although the code can be used to examine circRNA, the same quality of results cannot be expected.

Feature	Abbreviation	Description	Frequency chosen			
			8mer	7mer-m8	7mer-A1	6mer
miRNA						
3'-UTR target-site abundance	TA_3UTR	Number of sites in all annotated 3' UTRs (Arvey et al., 2010; Garcia et al., 2011)	100%	100%	100%	100%
ORF target-site abundance	TA_ORF	Number of sites in all annotated ORFs (Garcia et al., 2011)	9.4%	0.7%	68.1%	93.4%
Predicted seed-pairing stability	SPS	Predicted thermodynamic stability of seed pairing (Garcia et al., 2011)	100%	100%	100%	100%
sRNA position 1	sRNA1	Identity of nucleotide at position 1 of the sRNA	68%	100%	99.7%	97.7%
sRNA position 8	sRNA8	Identity of nucleotide at position 8 of the sRNA	0%	0.8%	100%	100%
Site						
Site position 1	site1	Identity of nucleotide at position 1 of the site	N/A	57.1%	N/A	2%
Site position 8	site8	Identity of nucleotide at position 8 of the site	0.8%	95.1%	99.4%	100%
Site position 9	site9	Identity of nucleotide at position 9 of the site (Lewis et al., 2005; Nielsen et al., 2007)	15.4%	7.1%	0.9%	93.7%
Site position 10	site10	Identity of nucleotide at position 10 of the site (Nielsen et al., 2007)	0.1%	100%	8.5%	26.3%
Local AU content	local_AU	AU content near the site (Grimson et al., 2007; Nielsen et al., 2007)	100%	100%	100%	100%
3' supplementary pairing	3P_score	Supplementary pairing at the miRNA 3' end (Grimson et al., 2007)	42.5%	100%	100%	100%
Distance from stop codon	dist_stop	\log_{10} (Distance of site from stop codon)	62.4%	10.8%	8.7%	25.7%
Predicted structural accessibility	SA	\log_{10}(Probability that a 14 nt segment centered on the match to sRNA positions 7 and 8 is unpaired)	100%	100%	100%	100%
Minimum distance	min_dist	\log_{10}(Minimum distance of site from stop codon or polyadenylation site) (Gaidatzis et al., 2007; Grimson et al., 2007; Majoros and Ohler, 2007)	99.9%	100%	87.4%	100%
Probability of conserved targeting	P_{CT}	Probability of site conservation, controlling for dinucleotide evolution and site context (Friedman et al., 2009)	100%	100%	100%	20.8%
mRNA						
5'-UTR length	len_5UTR	\log_{10} (Length of the 5' UTR)	98.2%	8.2%	4.6%	17.2%
ORF length	len_ORF	\log_{10}(Length of the ORF)	100%	100%	100%	100%
3'-UTR length	len_3UTR	\log_{10}(Length of the 3' UTR) (Hausser et al., 2009)	100%	100%	100%	100%
5'-UTR AU content	AU_5UTR	Fraction of AU nucleotides in the 5' UTR	13%	38.9%	91.1%	31.3%
ORF AU content	AU_ORF	Fraction of AU nucleotides in the ORF	1.2%	72.4%	28.4%	35.8%
3'-UTR AU content	AU_3UTR	Fraction of AU nucleotides in the 3' UTR (Robins and Press, 2005; Hausser et al., 2009)	5.4%	73.3%	65.3%	80.6%
3'-UTR offset-6mer sites	off6m	Number of offset-6mer sites in the 3' UTR (Friedman et al., 2009)	65.9%	89.6%	99.8%	100%
ORF 8mer sites	ORF8m	Number of 8mer sites in the ORF (Lewis et al., 2005; Reczko et al., 2012)	99.5%	99.1%	100%	100%
ORF 7mer-m8 sites	ORF7m8	Number of 7mer-m8 sites in the ORF (Reczko et al., 2012)	4.7%	4.3%	85.3%	100%
ORF 7mer-A1 sites	ORF7A1	Number of 7mer-A1 sites in the ORF (Reczko et al., 2012)	68.4%	34.2%	97.8%	98.4%
ORF 6mer sites	ORF6m	Number of 6mer sites in the ORF (Reczko et al., 2012)	91%	13.3%	0.7%	36.7%

Figure 4.1: 26 considered features, 14 highlighted selected features (reprint from [7])

4.2 miRanda

Probably the first tool for predicting miRNA-mRNA interactions is the miRanda. Although the website with the algorithm from 2003 [2] is not accessible anymore, miRanda remains one of the most mentioned algorithms in articles concerning miRNA-circRNA

interactions and is the only tool used in the ENCORI database.

The algorithm works in two parts. The first is similar to Smith-Waterman algorithm [47]; the dynamic programming is used to identify best possible non-overlapping hybridisation alignments. The scores used for different pairs and affine gap penalties are given in Table 4.1.

	A-U	C-G	G-U	Other pairing	Gap-opening	Gap-extension
Score	+5	+5	+2	-3	-8	-2

Table 4.1: Table showing scores for parameters of the first part of miRanda algorithm. First 4 columns are pairs formed between miRNA and target RNA, last two columns are scores for affine gap penalty.

The scores for the first eleven positions (counted from the 5' end of miRNA) are multiplied by a scaling factor of 2 to reflect the asymmetry of preferential binding in the seed region. In addition, there are four rules as described by Enright et al. (2003) [2]:

- mismatches at positions 2 to 4 are forbidden;
- less than five mismatches between positions 2 and 12 are allowed;
- at least one mismatch between position 9 to L-5 (where L is the total alignment length) is required;
- only fewer than two mismatches in the last five positions of the alignment are allowed.

The results from the first part of the algorithm are used to estimate the thermodynamic properties of predicted hybridisations. RNAlib (a folding prediction algorithm from the ViennaRNA package [48]) along with extended thermodynamic parameters from Mathews et al. (1999) [49] allow scoring of potential hybridisation sites by their folding energies. To calculate the minimum energy of the structure, a template RNA must be formed. miRanda joins miRNA with part of the complementary target RNA (the output of hybridisation in the first part of the algorithm) by a linker containing 8 artificial bases that cannot form base pairs.

4.3 RNAhybrid

RNAhybrid is another well-known tool for miRNA-mRNA prediction. Although this tool has not been upgraded since 2006, its unique method of searching for miRNA targets keeps it among the frequently used tools along with TargetScan.

To identify possible miRNA targets, RNAhybrid calculates the most energetically favourable hybridisation sites of miRNA in long RNAs such as mRNA or circRNA. Intramolecular hybridisations, the formation of complementary pairs between two parts of the same RNA, and overlapping hits are not allowed. The algorithms' time consumption is linear with respect to the length of the target RNA (circRNA, mRNA).

The algorithm is an extension of a general algorithm for prediction of RNA secondary structure [50] and uses energy parameters from Mathews et al. (1999) [49]. The algorithm uses Dynamic Programming to find all possible start positions in both RNAs to find the best hybridisation with minimum free energy (MFE) between binding pairs of the two RNAs. The algorithm is designed explicitly for RNA hybridisation and not RNA folding or pairwise sequence alignment.

The number and location of output sites are defined by constraints imposed by pre-defined or user-defined settings. The number of predicted sites can be limited by user-defined:

- number of optimal and additional suboptimal hits,
- free energy thresholds,
- or p-value thresholds.

Other options include:

- forcing hybridisations to contain only perfect helices in the seed region of miRNA (nucleotides 2 to 7),
- disallowing G:U base pairs in the seed region,
- or restricting maximum length of bulge loops (sequence of unpaired nucleotides in both strands) and internal loops (sequence of unpaired nucleotides in one strand).

Initially, the tool could be used only using the command line, however, since 2006 a web service is also available [51, 52].

4.4 Ensemble of TargetScan, RNAhybrid and miRanda

An ensemble has been proposed by Dori (2019) [53] for reduction of false positives identified by individual existing tools including TargetScan, RNAhybrid and miRanda. This is a "manual" method that has been so far performed on 100 randomly chosen mouse circRNAs.

The first step uses the three mentioned algorithms to predict miRNA-circRNA interaction sites. In the second step, all interactions predicted by only one algorithm were eliminated. Next, complementary experimental data was used. To pass the last step, the sites had to overlap with the binding sites of Argonaut proteins that are required for functional interactions.

So far the ensemble proposed was only performed on the mouse circRNA and has not been rigorously tested and evaluated.

Chapter 5

Technical Background

This chapter defines the existing algorithms, tools and methods that have been used during the implementation of the proposed solution.

5.1 RNAFold

RNAfold is a program from ViennaRNA package [54]. It is one of the few tools able to calculate secondary structure specific for circular RNA based on minimal free energy. The program can also compute a partition function, base pairing probabilities and generate a representation of probabilities of pairs in pseudo bracket notation.

Based on the command-line arguments, the program accepts either a file with primary sequences in text format (where one line consists of one sequence), a FASTA file, or a single sequence from stdin. **-c** option must be given to run this program specifically for circRNA sequences. To obtain pairing probabilities for each position of the sequence, option **-p** must also be used.

RNAfold generates three types of secondary structures and their free energy. The first structure is the minimum free energy (MFE) structure in dot-bracket notation (explained in Section 5.1.1). The second is a pairing probability denoted using pseudo-bracket notation (also in Section 5.1.1). The last is a centroid structure [55] which is the structure with "the smallest average base pair distance to all other structures in the ensemble" [54].

The asymptotic complexity of RNAfold is $\mathcal{O}(n^3)$ [54] where n is the length of the sequence. The complexity is not a problem because only sequences of a maximum length of 2000 nucleotides (nt) are used in this thesis.

For further information on available options, the reader is advised to visit the package manual at <https://www.tbi.univie.ac.at/RNA/RNAfold.1.html>.

5.2.1 Substitution matrix

A substitution matrix is used to evaluate the similarity between two nucleotide bases or amino acids. In our case, the matrix scores the log-likelihood of pairing between two nucleotides.

	A	U	C	G
A	-3	5	-3	-3
U	5	-3	-3	2
C	-3	-3	-3	5
G	-3	2	5	-3

Table 5.2: Substitution matrix for Watson-Crick pairing including G:U wobble pairs.[2]

5.3 Classifiers Used with Secondary Structures

In order to evaluate the information contained within different datasets of the secondary structure subsequences, various classifiers have been used. This section describes all classifiers tested for their ability to separate data into two classes as indicated by their labels. All classifiers used for this thesis and presented here are implemented in scikit-learn library [57].

5.3.1 k-Nearest Neighbours Classifier

The k-Nearest Neighbours (k-NN) splits data based on the majority voting of nearest neighbours to the queried sample. A point is classified as belonging to the class with the majority among k points. The value of k is specified by the user based on the data that are to be evaluated. Increasing k often leads to reduction of noise in the data, but the boundaries between categories become less distinct. By default, the votes have uniform weights. In some cases, it is better to use weights based on distance from the point queried.

5.3.2 Decision Tree Classifier

Decision Tree Classifier (DT) is a supervised non-parametric learning method [58, 59]. A model is built based on series of boolean decisions that can be inferred from the dataset. The performance of DT is affected by the maximum depth parameter set by the user. A high value might lead to overfitting on the training data, leading to poor generalisation with testing data.

5.3.3 Random Forest

Random Forest (RF) [60] uses a set of decision tree classifiers and their averaging to learn a model that would increase the accuracy of predictions and prevent overfitting. The main parameters that the user can set include the number of estimators (decision trees) and the maximum depth of the trees. If the bootstrap parameter is True, then each decision tree is trained on a subset of the dataset.

5.3.4 AdaBoost

AdaBoost [61, 62] is a meta-estimator, i.e., it is an ensemble formed by a set of many unreliable classifiers ("weak classifiers"). An example of a weak classifier is the Decision Tree Classifier. Multiple copies of the estimator are used to mitigate mistakes made by the original classifier by adjusting weights for incorrectly classified samples. As a result, AdaBoost is less susceptible to over-fitting. AdaBoost has been proven to converge into a strong classifier provided that all its weak classifiers are better than random guessing.

5.3.5 SVC

C-Support Vector Classifier (SVC) [63, 64] is a scikit-learn implementation of a support vector machine tailored to classification purposes. It uses points that are closest to the margin between two classes to decide the best splitting line. These points are called support vectors. The classifier tries to maximise the margin from splitting line to closest support vectors. For non-linear functions, this classifier can use a kernel function that takes training data and projects them into a higher dimension where the data are easier to separate with a plane. Most common kernel functions include radial basis function (rbf) kernel, polynomial kernel and sigmoid kernel. The classifier is effective even in high dimensional spaces.

Other parameters that the user can specify for an SVC include parameter C, degree (for polynomial kernel), kernel coefficient *gamma* or class-weight in case of unbalanced classes. Parameter C is a regularisation parameter responsible for a trade-off between correct classification and margin maximisation. The strength of regularisation is inversely proportional to the size of C. Parameter *gamma* is a kernel coefficient that defines the reach of the influence of a single training sample. The larger the *gamma*, the smaller the reach and the smaller the value, the larger the reach.

5.3.6 Complement Naive Bayes

Complement Naive Bayes (CNB) [65] is a modification of Naive Bayes adapting the classifier to work well with imbalanced datasets. Naive Bayes classifier applies the Bayes Theorem with assumption that for all pairs of features given the class conditional independence applies. From that we can infer following equation for classification:

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i|y)$$

This can be used to model conditional probabilities that can be further used for classification of data.

5.4 Analysis Tools

This section describes various tools used to visualise different datasets and to identify the most suitable datasets for classification and the most successful classifiers based on secondary structure. These tools have also been implemented in the scikit-learn library.

5.4.1 PCA and ICA

Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are statistical transformations of the secondary order and the higher order, respectively, with the main task to reduce dimensions. This property makes them especially valuable when attempting to visualise high dimensional data in two-dimensional graphs. The first two components of the two transformations allow visualisation of the highest variance in data when plotted. They are both class independent, meaning they do not attempt to separate the two classes but take the dataset as a single class.

PCA finds principal components (eigenvectors of the covariance matrix) of the dataset, meaning it finds a direction that explains the most variance. It removes correlations but not higher order dependence. The identified vectors are orthogonal, and the importance of found vectors can vary.

On the other hand, ICA is not orthogonal, all vectors have the same importance, and it can remove both correlation and higher order dependence. Unlike PCA, it works with non-Gaussian data. ICA tries to decompose independent components and assumes these are statistically independent.

5.4.2 GridSearch

GridSearch implementation from scikit-learn performs an exhaustive search over given parameter values to find the most suitable combination for given data and estimator.

5.5 MLP Classifier for prediction of interactions

Neural networks (NNs) [66, 67] (example shown in Figure 5.1) consist of nodes called neurons that pass on information to other connected neurons. One neuron has precisely one output that it can pass on to multiple other neurons. A neural network generally consists of an input layer, an arbitrary number of hidden layers and an output layer. The sizes of layers may vary, and consecutive layers are connected through the connections between neurons. Each connection is assigned a weight among other connections, and a weighted sum is computed by a propagation function. The hidden layers usually consist of linear functions interspersed by activation functions.

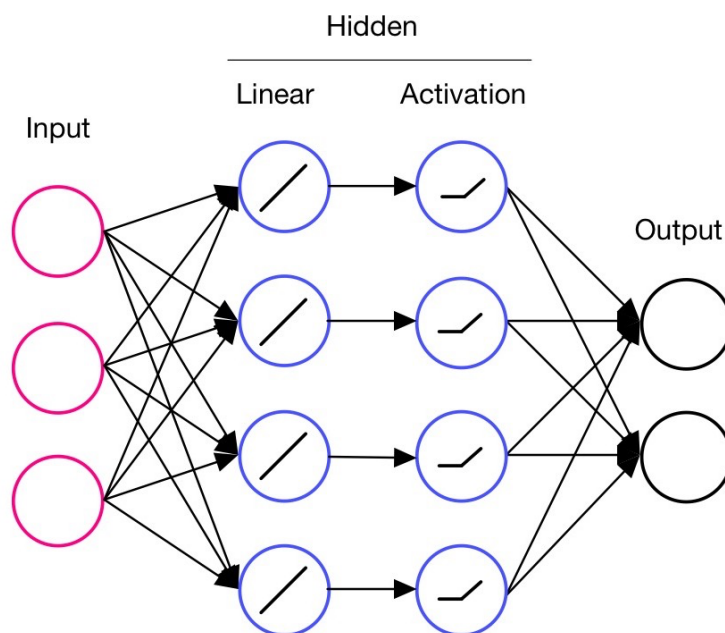


Figure 5.1: A neural network consisting of a single hidden layer formed of one linear and one active layer

For the prediction of interactions between circRNA and miRNA, a neural network model in the form of Multi-Layered Perceptron implementation from scikit-learn library has been selected. This model optimises the log-loss function, also known as Binary Cross-Entropy (see Equation 5.1 where y_i represents the actual class, and $P()$ is the probability

of that class)

$$-\frac{1}{N} \sum_{i=1}^N y_i * \log(P(y_i)) + (1 - y_i) * \log(1 - P(y_i)) \quad (5.1)$$

Weight optimiser options include LBFGS or stochastic gradient descent. LBFGS performs better with smaller datasets. For large datasets such as the ones used in this thesis, 'adam', the stochastic gradient-based optimiser is the most suitable, being superior in training time and resulting validation score.

MLP Classifier allows a user to specify, among other things, the number of hidden layers, penalty parameter alpha and most importantly, a selection of activation functions for hidden layers shown in Table 5.3.

Activation functions	
Identity function	$f(x) = x$
Logistic Sigmoid function	$f(x) = 1/(1 + \exp(-x))$
Hyperbolic Tan function	$f(x) = \tanh(x)$
Rectified Linear Unit function (ReLU)	$f(x) = \max(0, x)$

Table 5.3: Activation functions available for MLP Classifier by scikit-learn [3].

5.6 Evaluation

5.6.1 TP, TN, FP, FN and confusion matrix

To evaluate how well a classifier performs on a dataset, a confusion matrix might be constructed. The confusion matrix for binary classification consists of 4 values: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). These four values can be used to evaluate the predictions. Examples follow in Section 5.6.2 and Section 5.6.3.

5.6.2 SE, SP, gmean

Sensitivity (SE), Specificity (SP) and Geometric mean (Gm)) are all used to evaluate classifiers. Their computation is shown in the following equations:

$$SE = TP/(TP + FN)$$

$$SP = TN/(TN + FP)$$

$$Gm = \sqrt{SE * SP}$$

When assessing the performance of a model using accuracy, the best model is the one with the highest number of correctly identified samples. However, this approach in imbalanced datasets may lead to all samples being classified into negative class. For the classification of imbalanced datasets (such as datasets used in this thesis) where only a small proportion of the dataset belongs to the positive class and the majority belongs to the negative class, it is important to look for high SE in order to be able to identify novel positive samples. SP should also be as high as possible to make sure the number of false positives is minimised.

Gm evaluates the model considering both SE and SP. It is a good metric for assessing the performance of a model. However, even this metric can lead to suboptimal models as multiple models with varying SE and SP can lead to the same Gm.

The reader is advised to read [68] for more information on the evaluation of models with imbalanced data in bioinformatics and why standard methods (such as accuracy or ROC curve) are not suitable for representing the performance of machine learning models with imbalanced datasets.

5.6.3 Precision, Recall and F1-score

Precision says how many samples are positive from samples classified to that class. Recall reflects how many samples that should have been classified to a class have been classified correctly. F1-score represents the balance between precision and recall. Their computation is shown in the following equations.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-score} = 2 * \frac{precision * recall}{precision + recall}$$

Chapter 6

Implementation

As previously described in Section 1.1, the task was split into four sections: data extraction and analysis, classification based on secondary structures, classification and prediction of interactions, and evaluation and comparison with existing tools. In this section, the reader can find a detailed description of the implementation of these four tasks, including why specific decisions have been made. A diagram showing the processes and relationships between the first 3 sections is shown in Figure 6.1

6.1 Data Extraction and Analysis

6.1.1 Obtaining Data

First of all, general information about human circRNAs was downloaded from circBase database. The scope was limited to human circRNA as the central database with predicted miRNA-circRNA interactions (CircInteractome Section 3.4) is limited to it. The general information included genomic location (start, end, chromosome, strand), circRNA ID, genomic sequence length, mature sequence length and other less essential information.

Further, all experimentally validated human miRNA-circRNA interactions have been downloaded from ENCORI database. These later served as a basis for sample labelling in Section 6.1.2.2. The data downloaded also included a validated alignment between the two RNAs and miRNA sequence.

The dataset was limited by the intersection between circBase, CircInteractome and ENCORI downloads. The only circRNAs that were present in all three downloaded sets could be used from this point forward. Another restriction on the datasets was cast by the asymptotic complexity of RNAfold (Section 5.1). The mature sequences used a range of lengths between 100 and 2000 nt. Shorter sequences are unlikely to contain biologically relevant interactions, and longer sequences are less likely to be predicted by RNAfold in

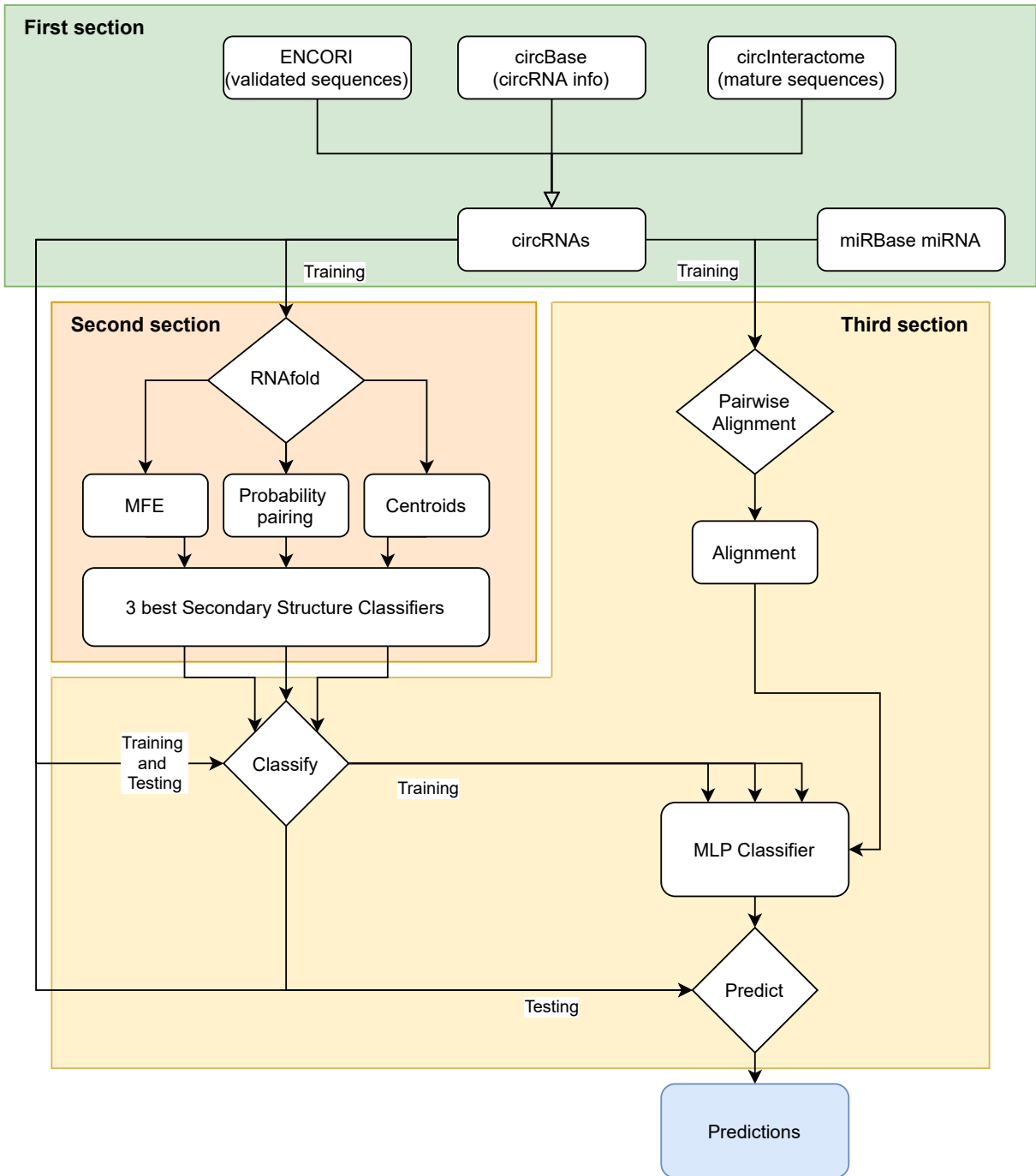


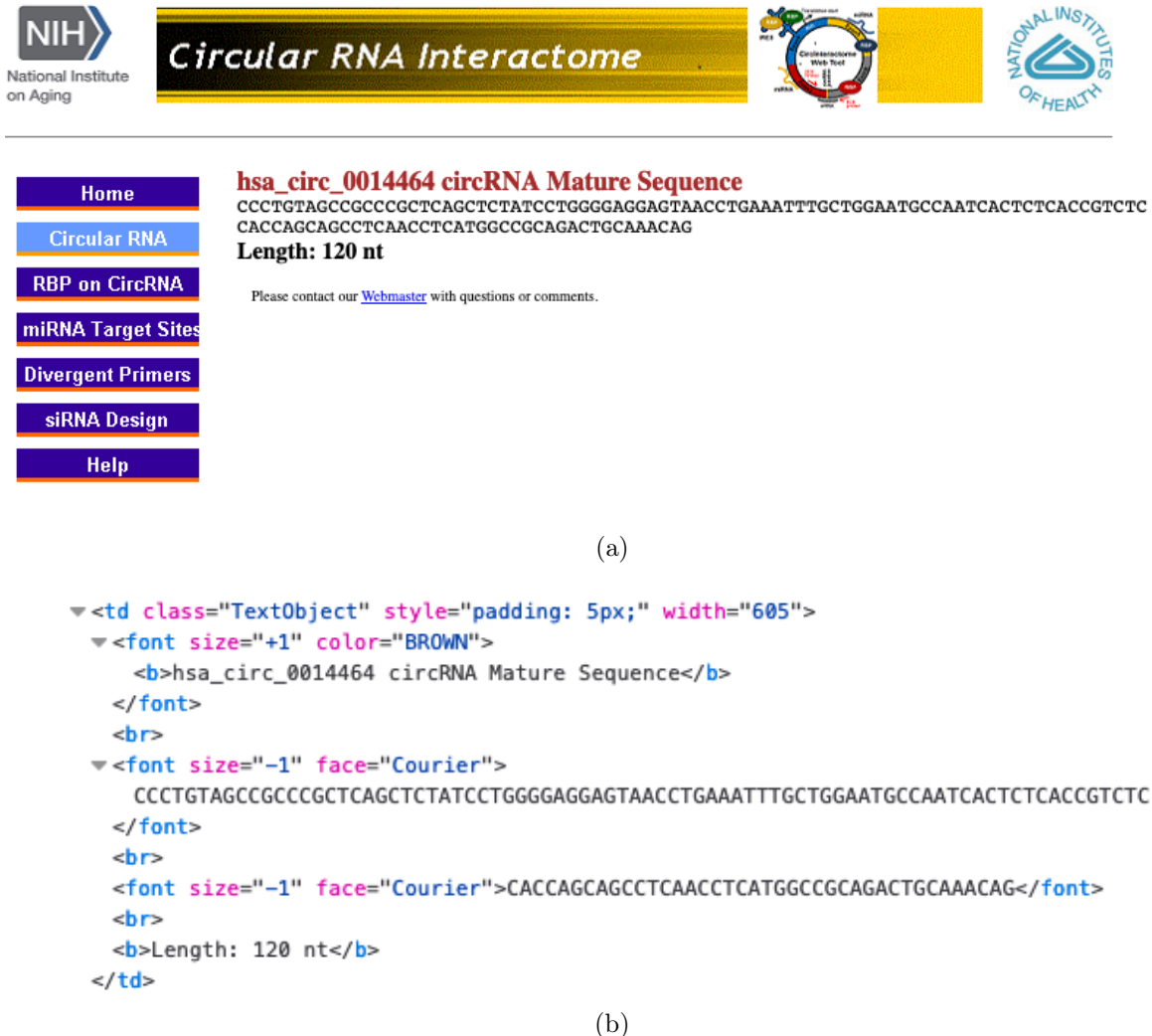
Figure 6.1: Diagram showing the processes and relationships between the main 3 sections leading to predictions by proposed method: data extraction, classification using secondary structures and classification and prediction of interactions.

a reasonable time. As a result of all these restrictions, the final dataset contains 1136 circRNAs.

Mature circRNA sequences that were later used as the basis for dataset creations were obtained from CircInteractome using a custom R script. CircInteractome is the only site containing transcripts of mature and genomic circRNA sequences on top of the genomic locations and sequence lengths. However, the transcripts are not available for download. A regex extraction

$$\{ier\} > [A|T|C|G]^+ < / \}$$

was used to isolate the sequences from HTML script (Figure 6.2). The extra characters were included and later removed to make sure only DNA sequences were captured from the website and that letters from other words of HTML were not included.



The figure shows a screenshot of the CircInteractome website. At the top, there are logos for NIH National Institute on Aging, the CircInteractome title, a circular diagram of a circRNA, and the National Institutes of Health logo. Below the navigation menu, the page displays the mature sequence for hsa_circ_0014464. The sequence is shown in two lines: CCCTGTAGCCGCCCGCTCAGCTCTATCCTGGGGAGGAGTAACCTGAAATTTGCTGGAATGCCAATCACTCTCACCGTCTC and CACCAGCAGCCTCAACCTCATGGCCGAGACTGCAAACAG. The length is noted as 120 nt. Below the sequence, there is a link to contact the webmaster. Part (b) shows the HTML code used to render this sequence, including tags for font size, color, and face.

(a)

```

<td class="TextObject" style="padding: 5px;" width="605">
  <font size="+1" color="BROWN">
    <b>hsa_circ_0014464 circRNA Mature Sequence</b>
  </font>
  <br>
  <font size="-1" face="Courier">
    CCCTGTAGCCGCCCGCTCAGCTCTATCCTGGGGAGGAGTAACCTGAAATTTGCTGGAATGCCAATCACTCTCACCGTCTC
  </font>
  <br>
  <font size="-1" face="Courier">CACCAGCAGCCTCAACCTCATGGCCGAGACTGCAAACAG</font>
  <br>
  <b>Length: 120 nt</b>
</td>

```

(b)

Figure 6.2: CircInteractome website example. a) An example circInteractome site with a mature sequence for hsa_circ_0014464, b) part of HTML of the same site as a) showing parts of sequences for extraction

In the future, additional circRNAs can be obtained for testing by downloading genomic sequences from ENSEMBL. The corresponding mature sequences can be obtained by splicing exons from their genomic counterparts using big bed files containing locations of splice-sites for each circRNA. Big bed files can be downloaded from <https://genome.mdc-berlin.de/> and converted using bigBedtoBed program [available at http://hgdownload.soe.ucsc.edu/admin/exe/macOSX.x86_64/bigBedToBed] to .bed format.

6.1.2 Creating datasets from secondary structure sequences

The data obtained was pre-processed into several datasets to find a representation of the secondary structure that best describes the data for classification and prediction. First, isolated mature circRNA sequences were used to obtain secondary structures using RNAfold (see Section 5.1). For each dataset, one of the 3 RNAfold outputs was selected (MFE, Pairing probability, Centroids) as a basis. From the selected output sequence, all possible subsequences of length 15 nt or 24 nt were produced. 15 nt was selected based on previous findings that show that the first 15 nt of the secondary structure of mRNA have the most significant influence on miRNA-mRNA interactions [7]. On the other hand, 24 nt long subsequences were selected to consider the full length of miRNAs. However, because the miRNAs vary in length, an average length of miRNA was selected.

The produced subsequences were then encoded based on the size of categories considered for the given dataset. The datasets used are described in Table 6.1. For 2 categories, the original subsequences were encoded as binary sequences. For 3 and 5 categories, one hot encoding was used (each category was represented as numerical array of ones). Due to the pseudo-bracket notation of probability pairings (Section 5.1.1) having inherently 3 categories (unpaired, weakly paired and strongly paired) when limited to only 2 categories (unpaired, paired), two new datasets were formed. The first considers weak pairs as unpaired, and the second considers weak pairs as paired. The sizes of the datasets range between 17,000 to 300,000 unique subsequences based on the complexity of the encoding.

6.1.2.1 Create dataset from alignments

The mature sequences were used again to create the *pairing dataset*. For simplicity, the length of all subsequences created from the primary sequence was equal to 15 nt. For each subsequence and each selected miRNA, the best alignment between the two sequences was selected. The alignment was established using PairwiseAligner (Section 5.2) from Biopython library. To be able to use the alignment of positive samples from ENCORI (Section 3.3), the same method for alignment was used. However, ENCORI alignment

Dataset	RNAfold	Length	# Categories	Encoding	Training set size
<i>2Categ15MFE</i>	MFE	15	2	(".") = 0 ("(") = 1	12,135
<i>2Categ15Centroids</i>	Centroids	15	2	(".") = 0 ("(") = 1	10,924
<i>2Categ15Less</i>	Probability	15	2	("." , { }) = 0 (" ()) = 1	19,652
<i>2Categ15More</i>	Probability	15	2	(".") = 0 (" , { } ()) = 1	13,418
<i>3Categs15</i>	Probability	15	3	OHE* ("." ,)	122,888
<i>3Categs24</i>	Probability	24	3	OHE* ("." ,)	248,536
<i>5Categs15</i>	Probability	15	5	OHE* ("." , ({ })	143,753
<i>5Categs24</i>	Probability	24	5	OHE* ("." , ({ })	267,040

Table 6.1: Datasets for classification based on secondary structures and their properties. The training set size excludes duplicated values. (*One Hot Encoder)

was found by miRanda software that is not available anymore, so the miRanda methods had to be replicated. The alignment was performed with the following scores:

- the substitution matrix from Section 5.2.1 was to score alignment
- -8 for opening a new gap in miRNA
- -2 for extending a gap in miRNA
- -100 for opening an internal gap in circRNA
- -100 for extending an internal gap in circRNA
- 0 for end gap in circRNA

In the end, the predicted alignment was converted to binary notation, encoding whether a position of original subsequence is paired or not.

Unlike in datasets for secondary structure based classification where only one set of subsequences from circRNA is necessary for all miRNAs, the *pairing dataset* includes aligned subsequences for each combination of circRNA and miRNA. The set, including all duplicates, was the size of 12,683,898 samples.

6.1.2.2 Labelling

Previously created prediction tools are based on observations from mRNA-miRNA interactions and predict sequences based on patterns such as seed pairing. As these interactions are predicted, it is not an ideal source for data labelling. Classifiers generally require samples from both positive and negative classes to learn to distinguish between them. For the positive samples, the experimentally validated interactions were selected and downloaded from ENCORI/StarBase database Section 3.3. The selection of negative samples

was more challenging as none have been previously generated. Two different approaches were selected.

The first approach is relevant to circRNA pairings. In this approach, all pairing structures that have not been labelled as positive were labelled as negative, assuming the following:

- If a sample has identical pairing as a positive sample, it should be labelled as a positive sample.
- For each miRNA-circRNA combination listed in ENCORI, all relevant and stable interactions have already been found and are contained in the database.
- Potential interactions not listed in ENCORI are insignificant (interact rarely or not at all).

The second approach is relevant to availability for interactions by secondary structure. Firstly, all positive samples were merged for each circRNA, making the sets of positive and negative samples independent of miRNAs. Secondly, the following additional assumptions to the assumptions made in the first approach were made:

- The secondary structure of positive samples allows them to interact with miRNAs more readily.
- If an unlabelled sample has identical properties as a positive sample, it should be labelled as a positive sample.
- If a positive sample is considered available for binding, we can assume the neighbouring samples (up to the width of subsequence) will not strictly be negative (these should be excluded from the dataset).
- Samples that do not fall under any of the previous categories are considered to be negative samples.

The process of labelling for secondary structures is also shown in Figure 6.3.

It is clear that just because a sample does not contain an interaction does not mean it is a negative sample. However, without any assumption about negative samples, it would be impossible to obtain a set of negative samples on which the classifier can train.

It is noteworthy that a sample from one dataset labelled as positive or negative does not necessarily have to have the same label in the other datasets. This is also true for the *pairing dataset*. Just because a sequence is available for pairing does not necessarily mean that the sequence would pair with every miRNA.

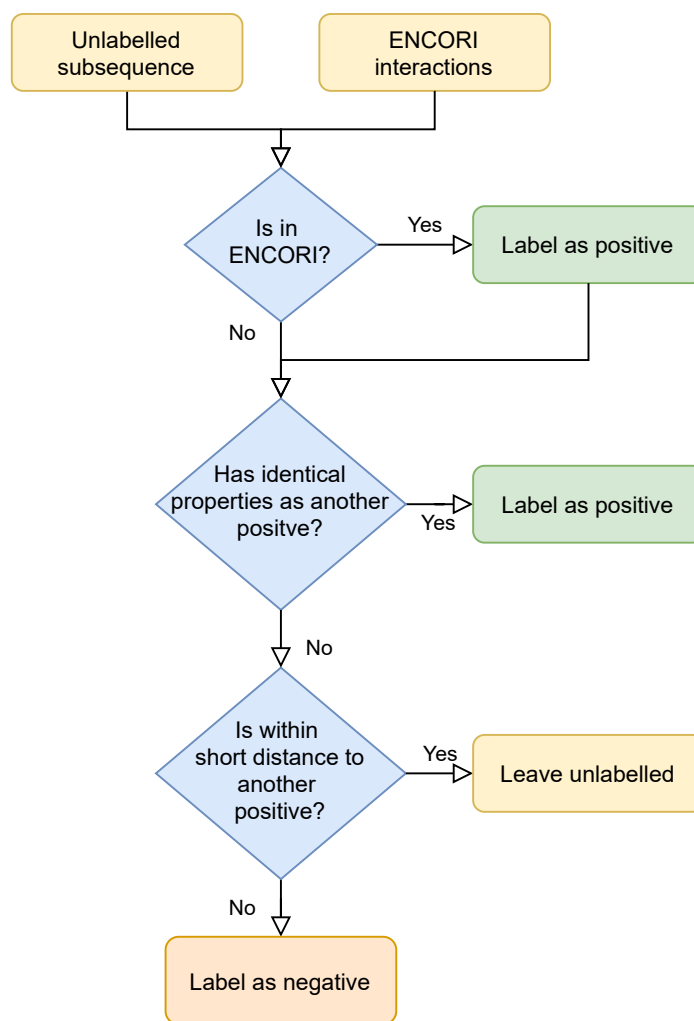


Figure 6.3: Process of labelling secondary structure subsequences.

6.2 Classification Based on Secondary Structures

The circRNA sequences were first separated into two sets: training (80%) and testing (20%). To ensure the results for classifiers were comparable, the training and the testing sets remained the same from this point forward.

With datasets showing promising signs, the secondary structure subsequences were ready for classification. A number of machine learning classifiers including k-NN (Section 5.3.1), DT (Section 5.3.2), RF (Section 5.3.3), Adaboost (Section 5.3.4), SVC (Section 5.3.5) and CNB (Section 5.3.6), were tested for their ability to learn to classify each dataset.

All classifiers were trained with the training set of circRNAs. The testing was done using some default and some randomly selected parameters. Furthermore, the tests were performed on both the training and testing sets. An example output is shown in Table 6.2.

Based on results of initial experiments, GridSearch with 5-fold cross-validation was used twice, first with random values of parameters. In the second round, the best random values were selected, and their neighbouring values were tested for improvements. The classifiers were evaluated using SE and Gm (Section 5.6.2) separately, and for each evaluation function, the best parameters were returned. SP was not used because it can be estimated given SE and Gm. An example outcome is shown in Table 6.3. The best parameters for each classifier were selected using the following method:

Algorithm 1: Selection of the optimal parameters for the classifiers

```

Input  : best SE and best Gm
Output: the best parameter setting
if parameters for best SE and best Gm are the same then
  | return parameters for best SE;
else
  | SE_mean = [];
  | Gm_mean = [];
  for set of parameters in [best SE and best Gm] do
    | get mean test score of parameters in SE and save it to SE_mean;
    | get mean test score of parameters in Gm and save it to Gm_mean;
  end
  if difference between values in SE_mean > difference between values in
    Gm_mean then
    | return parameters for best SE;
  else
    | return parameters for best Gm;
  end
end

```

Finally, four classifiers were selected along with three datasets. The combinations selected are shown in Table 6.4.

6.3 Classification and Prediction of Interactions

Three experiments were performed with sets of selected classifiers. The selected classifiers for each experiment are defined in Table 6.5. Subsequently, the classifiers were used to obtain labels for all of the circRNAs. The *pairing dataset* was combined with corresponding labels obtained from the classifiers, and the dataset was used for training of the MLP classifier. Because of the size of the training set, for training purposes duplicates were removed. Parameters of used MLP are shown in Table 6.6.

The experiments were tested on prepared testing data which, unlike the training data, included all subsequences from given circRNAs including duplicates. The confusion ma-

Dataset: <i>2Categs15Centroids</i>										
Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	2507	826	373	659	1199	0.27	0.09	0.64	0.75	0.69
k-NN	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Random Forest	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Adaboost	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	1781	1552	478	554	2030	0.47	0.11	0.54	0.53	0.54
Testing data										
SVM	1359	897	143	287	1040	0.39	0.03	0.67	0.60	0.63
k-NN	1616	640	185	245	825	0.31	0.04	0.57	0.72	0.64
Decision Tree	1568	688	192	238	880	0.33	0.04	0.55	0.70	0.62
Random Forest	1901	355	233	197	588	0.22	0.05	0.46	0.84	0.62
Adaboost	1575	681	196	234	877	0.33	0.04	0.54	0.70	0.62
Bayes Predictor	1138	1118	221	209	1339	0.50	0.05	0.49	0.50	0.50

Table 6.2: [

Initial classifier testing with mostly random parameter settings on *2Categs15Centroids* dataset]Initial classifier testing with mostly random parameter settings on *2Categs15Centroids* dataset. The first part shows the prediction results of training data. The second part shows the results for testing data. k-NN, DT, RF and Adaboost have very similar results, and with this setting, they all over-fit on training data. CNB in both instances can predict only slightly more than half of the data correctly, suggesting the data violate the main assumption of conditional independence. SVC ran with balanced classes (all positive samples and randomly selected negative samples of the same size). *AB = Adaboost

trices along with precision, recall and F1-scores, obtained for each MLPClassifier, were saved for later evaluation.

6.4 Comparison with Reference Tools

The proposed prediction method was compared with two reference tools - TargetScan and RNAhybrid. The same sets of pairs of miRNA and circRNA sequences were used to predict the miRNA-circRNA interactions. All predictions were recorded and compared with established ground truth labels. The RNAhybrid was used twice - first without any restrictions, second time forcing seed formation at positions 2-7. The tools were compared based on confusion matrices (further note in Section 7.4). The confusion matrices were formed by joining together all positive samples from all tools, including positive labels from ENCORI, hence the confusion matrices only contain samples that were classified as positive by at least one tool or ENCORI. The samples labelled as negative by all tools

SVC classifier									
Parameters			5-fold cross-validation					Score	
C	gamma	kernel	split0	split1	split2	split3	split4	mean	std
SE									
4	0.750	rbf	0.654	0.762	0.786	0.706	0.660	0.714	0.053
4	0.100	rbf	0.692	0.797	0.820	0.798	0.719	0.765	0.050
4	0.125	rbf	0.683	0.791	0.824	0.791	0.707	0.759	0.054
4	0.150	rbf	0.671	0.788	0.816	0.781	0.704	0.752	0.055
4	0.200	rbf	0.667	0.789	0.815	0.750	0.697	0.747	0.056
Gm									
4	0.750	rbf	0.685	0.734	0.745	0.751	0.699	0.723	0.026
4	0.100	rbf	0.736	0.791	0.806	0.792	0.754	0.776	0.026
4	0.125	rbf	0.734	0.789	0.806	0.791	0.752	0.774	0.027
4	0.150	rbf	0.734	0.787	0.806	0.788	0.755	0.774	0.026
4	0.200	rbf	0.729	0.785	0.801	0.782	0.741	0.767	0.028

Table 6.3: A part of final SVC GridSearch results for *2Categs15More* dataset. In this case, only the gamma values were tested. For each gamma value, 5-fold cross-validation (CV) was used, and results of each fold can be seen in columns "split". Based on mean and standard deviation, the performance with different gamma scores was ranked. In the first part of the table, the SE function was used to evaluate the folds, whereas, in the second part, the Gm function was used. The best score and corresponding parameters are highlighted.

Classifier	Parameters	Dataset	SE	Gm
SVC	C=10, class_weight=1: 2.930, gamma=0.1	2Categs15MFE	0.765	0.771
SVC	C=10, class_weight=1: 4.053, gamma=0.1	2Categs15Less	0.761	0.768
SVC	C=8, class_weight=1: 3.239, gamma=0.1	2Categs15Centroids	0.755	0.766
Adaboost	DT(max_depth=7), n_estimators=15	2Categs15MFE	0.563	0.706

Table 6.4: Four best performing classifiers, their parameter setting and dataset used to obtain the SE and Gm.

Experiment	Rows in Table 6.4
<i>Experiment 1</i>	1,2,3
<i>Experiment 2</i>	1,2,4
<i>Experiment 3</i>	1,3,4
<i>Experiment 4</i>	1,2,3,4

Table 6.5: Experimental setup for training and testing with MLP Classifier based on Table 6.4.

MLP Classifier	No of Layers	No of Nodes per Layer	No of iterations
<i>MLP_1_1</i>	1	1	1000
<i>MLP_1_12</i>	1	12	1000
<i>MLP_3_12</i>	3	12,12,12	1000
<i>MLP_3_D</i>	3	12,6,3	1000

Table 6.6: 8 sets of MLP parameters that were all used with all experimental setups from Table 6.5

were eliminated. Therefore, the true negatives contained within these confusion matrices consist only of samples that were considered as positive by one of the other tools. Apart from comparing the confusion matrices, execution times were also recorded.

Chapter 7

Results

In this chapter, all results obtained in different sections will be described.

7.1 Data analysis

In order to learn what kind of datasets have been created, each dataset (Table 6.1) has been visualised by plotting PCA and ICA (Section 5.4.1). Figure 7.1 and Figure 7.2 are selected examples of datasets showing graphs that were obtained. The rest of the graphs can be found in A.1.1

The graphs show us that the positive class forms a cluster within the dataset and it is not randomly scattered and thoroughly mixed with the negative results.

Apart from graphic visualisation, the sizes of datasets were also compared as shown in Table 6.1. The size of the dataset required to cover all possible subsequences can be calculated as $\#Categories^{length+1}$. For the smaller sets (2 Categories and Length of 15), between one-fifth to one-third of options were covered. For the larger sets, less than 1/350 were covered.

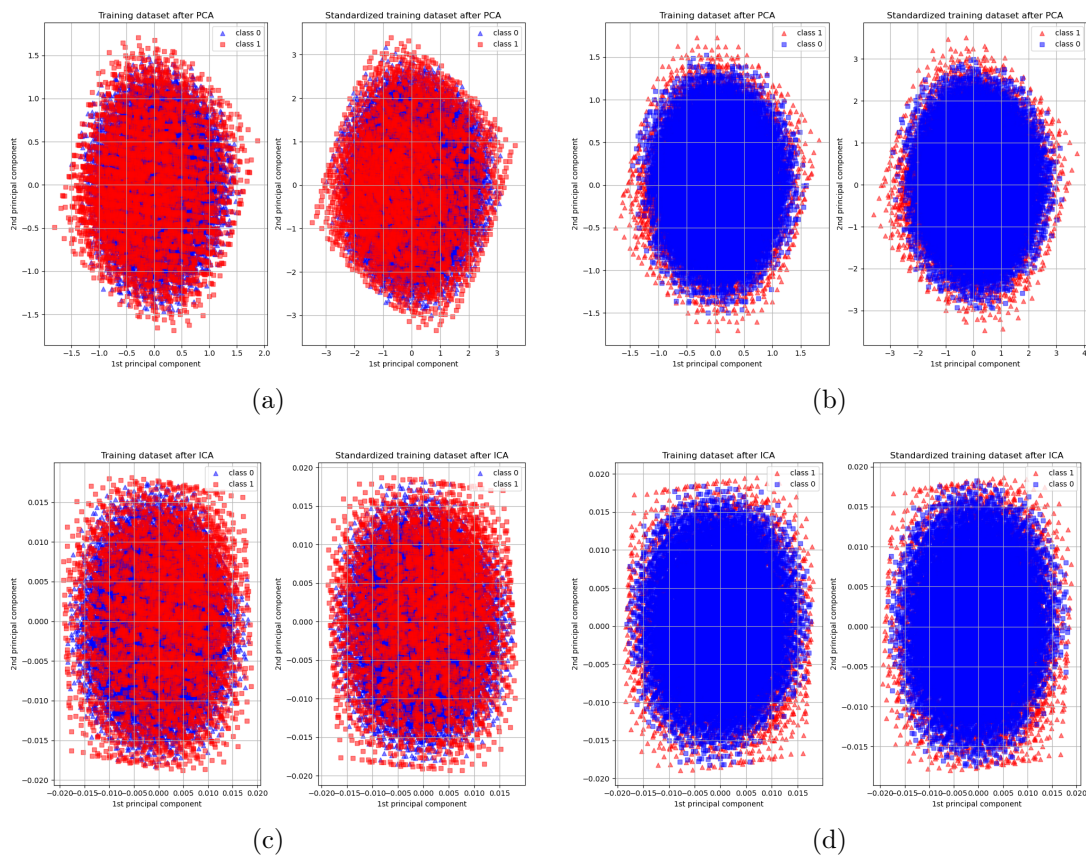


Figure 7.1: This set of graphs show PCA and ICA results for unmodified and standardized secondary structure-based *2Categs15Less* dataset. In all graphs red samples represent the positive class and blue represent the negative class. (a) and (b) contain each two graphs showing the unmodified and standardised dataset after PCA. The difference between (a) and (b) is caused by the order in which the two classes were added to the graph. The same is true for (c) and (d) except here the graphs show the ICA.

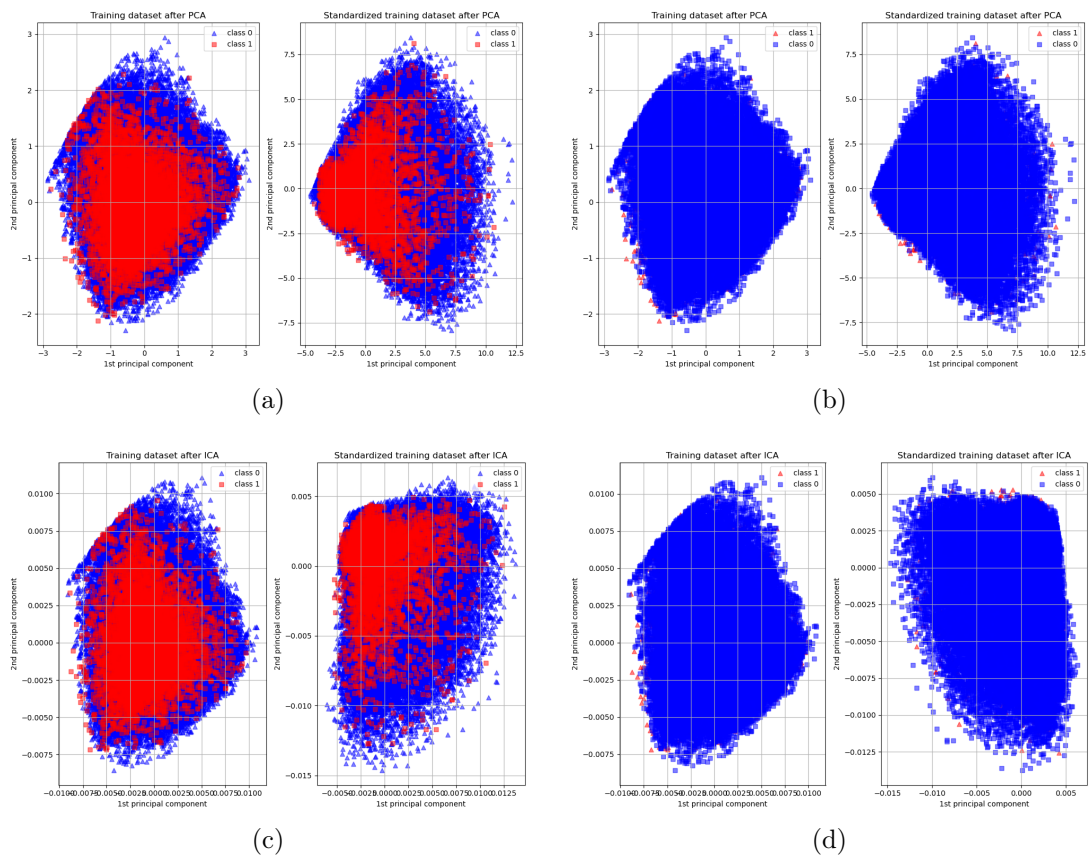


Figure 7.2: This set of graphs show PCA and ICA results for unmodified and standardized *3Cats15* dataset. In all graphs red samples represent the positive class and blue represent the negative class. As in Figure 7.1, (a) and (b) contain each two graphs showing the unmodified and standardised dataset after PCA. The difference between the two is caused by the order in which the two classes were added to the graph. The same is true for (c) and (d) except here the graphs show the ICA.

7.2 Secondary Structure-Based Classifiers

7.2.1 Initial Experiments

The initial experiments (example in Table 6.2) with smaller samples on secondary structure-based classifiers show that classifiers such as k-NN, DT, RF and Adaboost tend to over-fit on data. The comparison between the classification of training data and testing data supports this statement. When the SE, SP and Gm value all equal to 1 in training data, i.e. no mistakes were made, the SE, SP and Gm value variate around 0.72 for the testing data.

From the initial experiments, it can also be observed that the positive samples make up only around one-thousandth of the samples (including duplicates). Therefore, classifiers such as SVC need to consider the difference between the sizes of the two classes. Otherwise, these predictors prefer classifying all samples into one class (as seen in both training data and testing data results for 2Categs15 in Table 7.1) because the average number of mistakes is smaller than when the classifier is forced to classify at least some samples to another class. Increasing the number of mistakes, in this case, is desirable because the goal of the classification is to spot the rare cases of positive class at the cost of an increased number of false positives.

Dataset: 3Categs24

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	39591	56157	1179	2444	57336	0.58	0.01	0.67	0.41	0.53
k-NN	95748	0	0	3623	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	95747	1	45	3578	46	0.00	0.00	0.99	1.00	0.99
Random Forest	95748	0	45	3578	45	0.00	0.00	0.99	1.00	0.99
Adaboost	95748	0	0	3623	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	42468	53280	1357	2266	54637	0.55	0.01	0.63	0.44	0.53
Testing data										
SVM	9628	14926	313	559	15239	0.60	0.00	0.64	0.39	0.50
k-NN	24405	149	864	8	1013	0.04	0.01	0.01	0.99	0.10
Decision Tree	23070	1484	819	53	2303	0.09	0.01	0.06	0.94	0.24
Random Forest	24441	113	865	7	978	0.04	0.01	0.01	1.00	0.09
Adaboost	24320	234	858	14	1092	0.04	0.01	0.02	0.99	0.13
Bayes Predictor	10392	14162	325	547	14487	0.57	0.00	0.63	0.42	0.52

Table 7.1: Initial classifier testing with default parameter settings on secondary structure-based 3Categs24 dataset. The first part shows the prediction results of training data. The second part shows the results for testing data. Without some parameter setting, the classifiers often learn to return only the class with higher probability of being correct.

Table 6.2 is also an excellent example of why it is essential to look at SE, SP and Gm rather than at the percentage of mistakes. Here the SVC has made more mistakes than most other classifiers on both training and testing data. However, SVC is the most successful classifier when it comes to the correct identification of new positive data. A significantly lower SP value than most other classifiers suggests that the mistakes are generated mainly by false positives.

The best results are obtained with datasets limited to tens of thousands of samples for SVC or CNB. However, for *3Categs15* (the smallest large dataset), the Gm rarely exceeds 0.60 for testing data, and the maximum SE is 0.72. From the SE and Gm values, SP was estimated to be below 0.50. The low value of SP leads to a set of false positives that is ten times the size of true positives. The larger datasets perform even worse. A complete set of tables with results can be found in A.2.1.

7.2.2 GridSearch

At this point the datasets *3Categs15*, *3Categs24*, *5Categs15* and *5Categs24* were eliminated due to their size leading to time-consuming training of classifiers and low prediction scores. CNB was also eliminated from further experiments as none of the previous attempted experiments has shown SE or Gm higher than 0.60 making it unsuitable for the prediction of these data. The remaining datasets and classifiers were used in GridSearch. The results are reported in A.2.2.

The best results obtained for each classifier are summarised in Table 7.2. The *2Categs15MFE* dataset showed to be the best for the majority of classifiers. The best classifier is the SVC with the best results from both SE and Gm scores. The second best is the Adaboost classifier which is also the only other classifier that overcame the threshold of 0.6 for SE and 0.7 for Gm.

Classifier	Parameters	Dataset		Score	
		2Categs15	SE	SP	Gm
<i>SVC</i>	{'C': 10, 'class_weight': {1: 2.930}, 'gamma': 0.1}	<i>MFE</i>	0.777	0.782	0.775
<i>k-NN</i>	{'n_neighbors': 3, 'weights': 'distance'}	<i>Centroids</i>	0.585	0.716	0.645
<i>DT</i>	{'max_depth': 23}	<i>MFE</i>	0.513	0.784	0.632
<i>RF</i>	{'max_depth': 14, 'n_estimators': 3}	<i>MFE</i>	0.441	0.835	0.604
<i>Adaboost</i>	{'base_estimator': DT(max_depth=7), 'n_estimators': 15}	<i>MFE</i>	0.616	0.876	0.732

Table 7.2: The best results obtained for each classifier.

It is also noteworthy that SVC performed better than the other classifiers on all datasets. Suggesting a combination of SVC classifiers and different datasets may be the best option for representing the availability of secondary structure in the final stage of the proposed method. Based on the obtained results, k-NN, DT, and RF were not

used from this point forward, having a low score for Gm and especially SE. However, Adaboost will be tested along with various SVC to find the best combination of classifiers for final prediction because it has a high SP and, therefore, could partake in reducing false positives.

7.3 Classification of Interactions

The results for MLP Classifiers described in Table 6.6 can be found in A.3. The results of MLPs with different parameters are very similar. The number of false positives and number of false negatives is inversely proportional. The only difference between the experiments is the MLP structure required for the MLP to learn to distinguish between positive and negative class with minimal false negatives.

Experiment	Classifier	Confusion matrix				Precision		Recall		F1-Support	
		TN	FP	FN	TP	1	0	1	0	1	0
1	<i>MLP_1_1</i>	3015896	3265	5	3654	0.53	1.00	1.00	1.00	0.69	1.00
2	<i>MLP_1_12</i>	3015893	3266	3	3656	0.53	1.00	1.00	1.00	0.69	1.00
3	<i>MLP_3_12</i>	3015887	3272	0	3659	0.53	1.00	1.00	1.00	0.69	1.00
4	<i>MLP_1_12</i>	3015914	3245	17	3642	0.53	1.00	1.00	1.00	0.69	1.00

Table 7.3: The best MLP and its results for each experiment.

Table 7.3 shows the best MLP for each experiment. Only for *Experiment 1* was one node enough for the MLP to recognise the majority of positive samples correctly. The least successful was *Experiment 4* with the combination of all four classifiers. Due to the size of the negative dataset, the precision and recall always equal 1.00; therefore, it is more informative to look at the number of false positives compared to the true positives directly. The precision and recall for all MLP that have managed to classify data into two classes is the same, with values 0.53 and 1, respectively. As a result, the F1-score is also always the same, with a value of 0.69. In cases where the MLP did not learn to distinguish the two classes, precision and recall are 0.

Experiment	Classifier	Confusion matrix				Precision		Recall		F1-Support	
		TN	FP	FN	TP	1	0	1	0	1	0
<i>Pairing only</i>	<i>MLP_1_1</i>	3019159	0	3659	0	0.00	1.00	0.00	1.00	0.00	1.00
<i>Pairing only</i>	<i>MLP_1_12</i>	3015887	3272	0	3659	0.53	1.00	1.00	1.00	0.69	1.00
<i>Pairing only</i>	<i>MLP_3_12</i>	3015887	3272	0	3659	0.53	1.00	1.00	1.00	0.69	1.00
<i>Pairing only</i>	<i>MLP_3_D</i>	3019159	0	3659	0	0.00	1.00	0.00	1.00	0.00	1.00

Table 7.4: Results for MLPs based on pairing only.

Based on the similarity of the results for MLPs with different secondary structure-

based classifiers, an MLP that did not consider the secondary structure was also constructed. This classifier only considered the primary structure-based pairing with miRNA. The results of this MLP are shown in Table 7.4. The MLP based on only pairing either learns to distinguish all true positives or does not distinguish positives at all, e.g. when the MLP has only one layer. In comparison, the *Experiment 1* performs the best with one node and one layer. In addition, MLP with the secondary structure-based classifiers only was also constructed. The when duplicates were eliminated such that occurrences labelled as positive were kept, all samples were classified by neural network as positive.

7.4 Comparison with Reference Tools

The confusion matrices are shown in Table 7.5. It is important to point out that the testing dataset was the same one as in MLP classification, but the confusion matrix was created using only samples that were labelled as positive by at least one of the tools or were labelled positive by ENCORI. Based on the ENCORI labelling, the proposed method with the recall of 0.999 is able to capture more true positives than any other compared tool. However, compared to TargetScan and RNAhybrid with seed with the precision of 0.667 and 0.637, respectively, the proposed method has a significantly lower precision of 0.528.

Tool	TP	FN	FP	TN	Precision	Recall
<i>Proposed method</i>	3655	4	3267	5496	0.528	0.999
<i>TargetScan</i>	2771	888	1383	7380	0.667	0.757
<i>RNAhybrid w. seed</i>	2211	1448	1256	7507	0.637	0.604
<i>RNAhybrid wo. seed</i>	29	3630	3442	5321	0.008	0.008

Table 7.5: Confusion matrices for compared tools. The confusion matrices were created only with samples that were classified as positive by at least one tool or by ENCORI. Precision and recall are stated for the positive class only.

Venn diagrams have been created to visualise the overlap between the proposed solution and various combinations of tools and positive samples established by ENCORI and can be found in Figure 7.3. Figure 7.3 (a) RNAhybrid without forced seed generates sequences that are primarily without seed. The sequences that have seed are not always found in ENCORI positive results. (b) Samples in the intersection between RNAhybrid with forced seed region and proposed solution are mostly intersecting with ENCORI positive samples, therefore, correctly classified. Similarly, between the proposed method and TargetScan in (c). (e) and (f) show that RNAhybrid without seed finds very different results from the other tools. Taking into consideration (b) and (c), (d) suggests that the

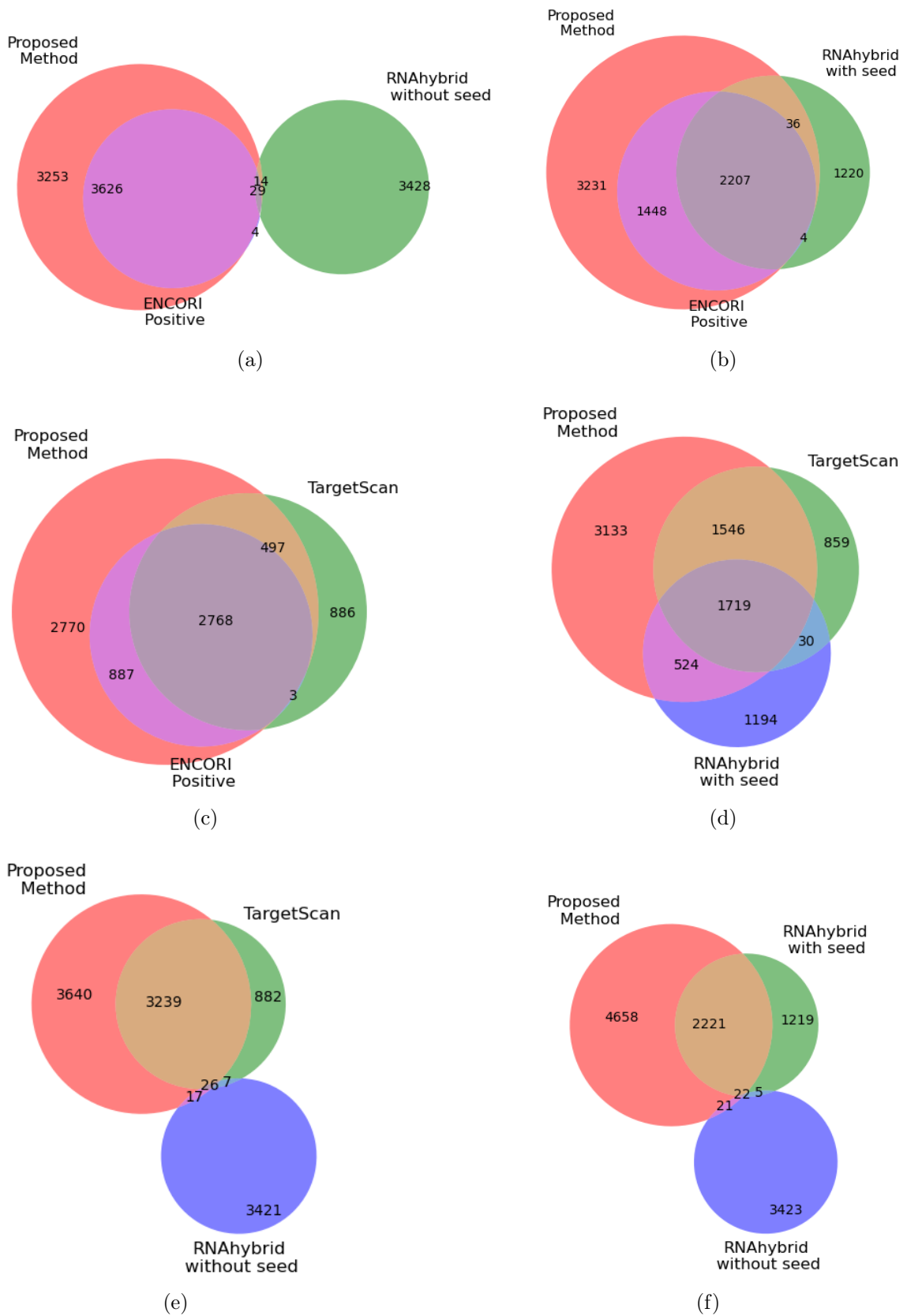
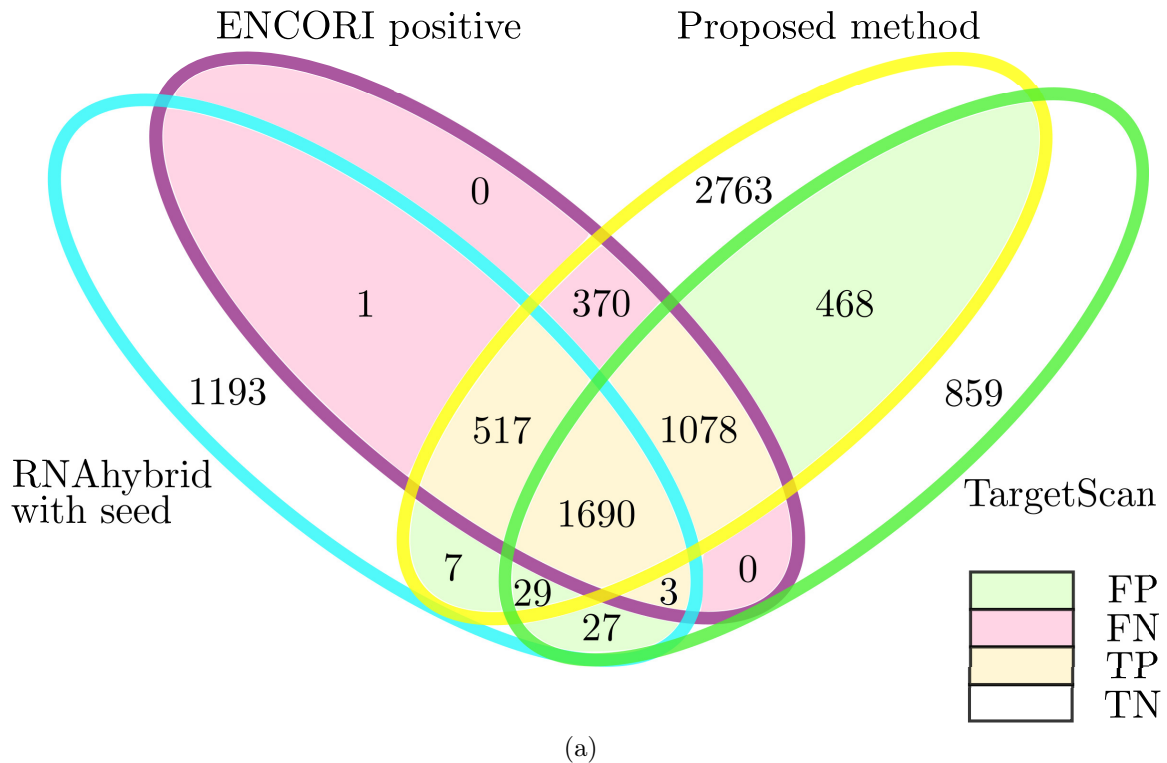


Figure 7.3: Subsections (a), (b) and (c) show proposed solution, ENCORI positives and one of the reference tools. (d), (e) and (f) show combinations of proposed method and two reference tools.

number of correctly classified positive samples could increase based on majority voting between the proposed method, TargetScan and RNAhybrid with seed.

7.4.1 Ensemble proposition

An ensemble of the three tools (Proposed method, TargetScan and RNAhybrid with seed) was considered based on observations from Figure 7.3. A Venn diagram representing these tools along with the ENCORI positive set can be found in Figure 7.4 (a). The corresponding confusion matrix, precision and recall are in Table 7.4 (b). A sample was classified to the positive class if predicted by at least two tools and to the negative class otherwise. The results show the ensemble would be a better predictor than any tool independently with the precision of 0.839 and the recall of 0.882.



Tool	TP	FN	FP	TN	Precision	Recall
<i>Ensemble</i>	2771	371	531	4815	0.839	0.882

(b)

Figure 7.4: Venn diagram for ensemble based on majority-voting of proposed method, TargetScan and RNAhybrid with seed.

Chapter 8

Evaluation and Discussion

In this chapter, the results from the previous chapter will be evaluated, and their relevance and importance will be discussed. The chapter will be concluded with remarks on any future outlooks.

8.1 Data Analysis

Based on the PCA results in Section 7.1, it was assumed that the datasets (see Table 6.1) contain information that has the potential to be used for the classification of interactions between circRNA and miRNA. Generally, not much difference was observed between the graphs generated for a set of datasets with the same number of categories. The patterns observed in the datasets within the same category (2,3,5) show little variation, and it can be thus predicted that the data will be split similarly by classifiers.

The larger datasets may perform better as the classes seem to be less intertwined. However, the greater ratio between the positive class and the negative class suggests otherwise. The imbalanced datasets may perform worse during classification.

8.2 Secondary Structure-Based Classifiers

The initial experiments have shown that only some of the datasets have suitable representation for classification. As predicted during data analysis, the larger datasets based on 3 and 5 Categories are strongly unsuitable for intended classification. The one-hot encoding with the combination of a large number of sequence positions resulted in multi-dimensional spaces that cannot be well described by the size of training data available with the current state of research. In comparison, the *2Categs* datasets have shown a potential that has been further increased with better-selected classifier parameters during GridSearch.

Along with the datasets, the classifiers were also pruned to eliminate those not suitable for classification. Low performance was detected in k-NN, DT, RF and CNB. CNB is most likely unsuitable because the data in the datasets is not normally distributed. The remaining classifiers are unsuitable because the data is not easily separable and cannot be easily generalised.

The Adaboost, although performing on a similar principle as DT and RF, performs much better. The success can be attributed to the ability to learn by combining several classifiers with different weights, leading to less generalisation, which seems to be important with this type of data. The general success of SVC is not surprising. It can adapt to the varying proportions of classes allowing it to maintain the ratio during testing as well. Further high parameter C pushes to avoid misclassification while low gamma prevents the classifier from over-fitting.

Combining SVCs with 2Categs datasets and potentially Adaboost with 2CategsMFE are the only classifier-dataset combinations suitable for further use in prediction. If the limit was set lower, it would be less likely that the information obtained from these classifiers would be relevant. The problem has already been generalised to only 2Categories, and 15 nt of circRNA sequence and lower SE and SP could further reduce the nuances between negative and positive classes. If the nuances were maintained using the highest possible SE and SP remained to be seen at this point.

8.3 Classification of Interactions

Seeing how similar the MLP results are, cross-validation for each dataset was not performed because it would not bring any new information. The differences between MLPs with different parameters are not significant, so any significant difference between MLPs with the same parameters but varying data cannot be expected. The lack of differences also suggests that the data are relatively simple.

Comparing the MLPs that include classifiers with MLP that do not include classifiers suggests that at least one SVC adds extra information. The fact that the one-layer MLPs without classifiers are not able to split the data into two classes suggests the SVCs are adding the extra information to do this.

However, the extra information added by the SVC does not reduce the number of false positives of miRNA-circRNA interactions. It only allows a simple MLP to find very similar results to a more complex MLP without it. The experimental results suggest the influence of seed is stronger than the influence of secondary structure-based classifiers, so the desired reduction of false positives was not observed. The generalisation and correlation to the size of the datasets and error rate may be too large to maintain the

nuances between the positive and negative dataset. Furthermore, the classifiers are at most 4, leaving $2^4 = 16$ combinations of four binary outputs. As we have seen in the results based on secondary structure-based classifiers only, none of the combinations fall strictly into one class, decreasing the differences between the two classes.

8.4 Reference Tools Comparison and Ensemble Proposition

The main differences observed between the three tools that all consider seed can be attributed to their approach towards implementation of the seed constrain. TargetScan takes on the input miRNAs with only seven positions of seed region, disregarding potential binding of sequences out of the seed region. All identified sequences are then rated by additional characteristics to rank the interactions by a score. However, none are eliminated, so all are considered as positives in this thesis.

On the other hand, RNAhybrid considers the entire length of miRNA sequence in establishing pairing with circRNAs. The scoring of sequences is based on the minimum free energy of the miRNA-circRNA compound. Additionally, a threshold limits the number of interactions on the output.

The proposed method in its approach considers only 15 nt of a circRNA with the entire length of miRNA. The seed is not forced as with the other two tools, and all sequences are considered equally. The seed can be observed in all positive results because ENCORI data were used as a template for the datasets. If the experimentally validated miRNA-circRNA interactions included non-canonical binding sites, the proposed method (unlike the other tools) would potentially be able to pick up these sites as well.

Unlike the other tools, the proposed method can correctly detect all ENCORI positive samples at the expense of an almost equal-sized number of false positives. Even though the number of false positives for the other two methods is significantly lower, so is the number of true positives. The decrease in the number of false positives is at the expense of an increase in false negatives, which the proposed method does not have.

The opposing shortcomings of the tools can nicely complement each other leading to an unexpected outcome. The proposed combination of tools significantly decreases the numbers of both false negatives and false positives. A further improvement might be achieved using logistic regression to give weights to the individual classifiers instead of just applying majority voting.

8.5 Discussion

The secondary structure-based classification comes relatively short of its expectations. Some of the limitations are caused by the excessive generalisation that results from the complexity of secondary structure and the lack of positive examples. This leads to poor description of the multi-dimensional space, which does not allow clear separation of positive class from negative. The new samples cannot be easily approximated. The imbalanced datasets (over 3 million negatives to 3,500 positives in the final testing set) ruled out many evaluation methods along with machine learning classifiers that would otherwise be able to split the data.

Furthermore, several assumptions had to be made to allow for data prediction using machine learning classifiers. For once, the miRNA-circRNA interaction data has never before been separated into two classes. The separation itself is based on many assumptions that do not necessarily have to be accurate, such that if a type of pairing is known to be positive in one place, it may not necessarily have the same properties in another. The pairing has also been generalised. Although G:U wobble pairs were considered by *PairwiseAligner*, for simplicity, the subsequent evaluation treated G:U wobble pairs as unpaired sequences. The secondary structure adds another assumption as it is not experimentally validated but only predicted. Such assumptions align with the biological knowledge we currently have, but the generalisation of the problem for computational purposes is far too great for the secondary structure to significantly impact the prediction.

The final results are nonetheless relevant. Although unforeseen, the findings regarding the MLP classifier and the possibility of an ensemble with existing tools have the potential to improve the current state of miRNA-circRNA interactions prediction.

8.5.1 Possible Improvements and Future Work

Several options could be further developed to find out whether the secondary structure could be used in a different form. If the *2Categs* datasets were too generalised and the *3Categs* and *5Categs* were too large to cover all possible combinations with the amount of data available, then *3Categs* or *5Categs* with only length of the seed could be a sufficient solution.

Another potential approach towards the problem would be to use the MLP with primary sequence only and test the secondary structure of the positively labelled samples. TP and FP of the MLP would be used as the positive and the negative class, respectively. Because the Precision of MLP is 0.53, two classes would be almost balanced out, removing many of the obstacles overcome during experiments considered in this thesis. The

secondary structure could be used to eliminate some of the false positives using the established classifiers with decisions based on majority voting, or the secondary structure-based classifiers could be trained directly on this data.

A different approach towards the problem would be through free energy. Knowing the secondary structure of circRNA, the free energy of subsequences could be calculated as well as the free energy of a potential complex with miRNA. The likelihood of the interaction would be established based on a difference between the two complexes. Such approach could be an extension to our proposed method or work as a stand-alone method.

Last but not least, the datasets are represented as binary strings or one-hot encoding and both only work with 0s and 1s. As such, the space considered is a hypercube and could be potentially described by logical formulas. For example, we could describe the interactions as sequences with a match at positions 2-7, while positions 9-10 would be forbidden to bind. Apart from the knowledge we already have about the miRNA-circRNA interacting sequences, we could form more logical formulas and conditions based on the available data. A machine-learning algorithm could then be used to check whether these conditions apply. This approach would affirm and extend the works of Agarwal et al. (2015) [7] on miRNA-mRNA interactions for TargetScan. Other methods that could take advantage of this representation or even of the original Pseudo-Bracket Notation would be Markov Logic Networks, which applies the Markov Network to probabilistic first-order logic, or Logical Neural Networks which connect the traits of neural networks and symbolical logic.

Chapter 9

Conclusion

This thesis aimed to identify a new method for the prediction of miRNA-circRNA interactions. Based on research experiments conducted with secondary structure-based classifiers and a simple neural network, it can be concluded that the predicted secondary structure of circRNAs is not a significant factor in the prediction of miRNA-circRNA interactions. However, the proposed method is an important factor in ensemble with TargetScan and RNAhybrid. The results indicate that the ensemble of the three methods performs better than any of the methods individually in both, precision and recall.

This study clearly illustrates that neural networks based on 15 nt subsequences can, with low precision, detect all known interactions from the ENCORI database. This raises the question of whether this method has the potential to unravel the non-canonical sites as well. Further research in this area is still required.

The secondary structure-based classifiers have not shown strong influence in the final neural network. The experiments show that the secondary structure is very complex. Without methods that would allow for less generalisation, the nuances between binding and non-binding sites are mostly lost.

The findings challenge the proposed ensemble described in Section 4.4. However, to make the two methods comparable, this ensemble would have to be tested on the mouse genome as the experiments in this thesis are limited to the human genome only. With the growing size of known miRNA-circRNA interactions, the proposed method can be further improved and make an important addition to the analysis of miRNA-circRNA interactions.

Appendix A

A.1 Data Analysis

A.1.1 PCA and ICA

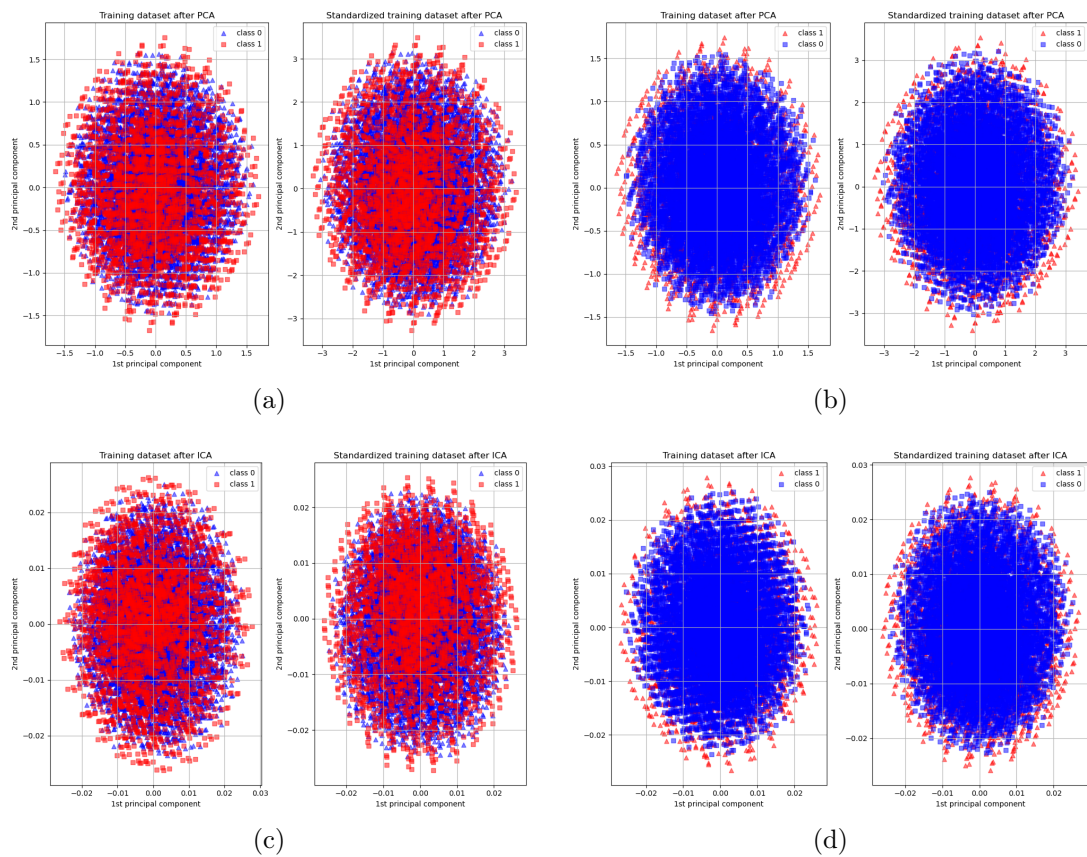


Figure A.1: This set of graphs show PCA and ICA results for unmodified and standardized *2Categs15Centroids* dataset. In all graphs red samples represent the positive class and blue represent the negative class. As in Figure 7.1, (a) and (b) contain each two graphs showing the unmodified and standardised dataset after PCA. The difference between the two is caused by the order in which the two classes were added to the graph. The same is true for (c) and (d) except here the graphs show the ICA.

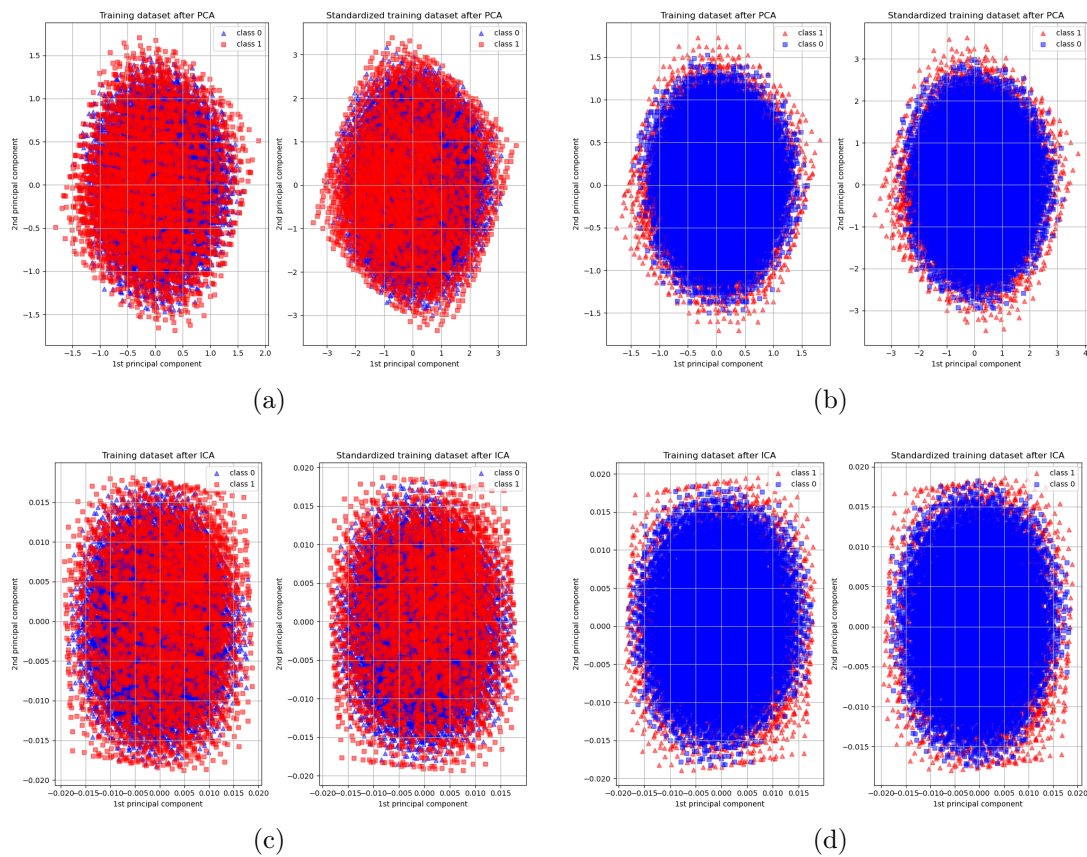


Figure A.2: This set of graphs show PCA and ICA results for unmodified and standardized *2Cats15Less* dataset. In all graphs red samples represent the positive class and blue represent the negative class. As in Figure 7.1, (a) and (b) contain each two graphs showing the unmodified and standardised dataset after PCA. The difference between the two is caused by the order in which the two classes were added to the graph. The same is true for (c) and (d) except here the graphs show the ICA.

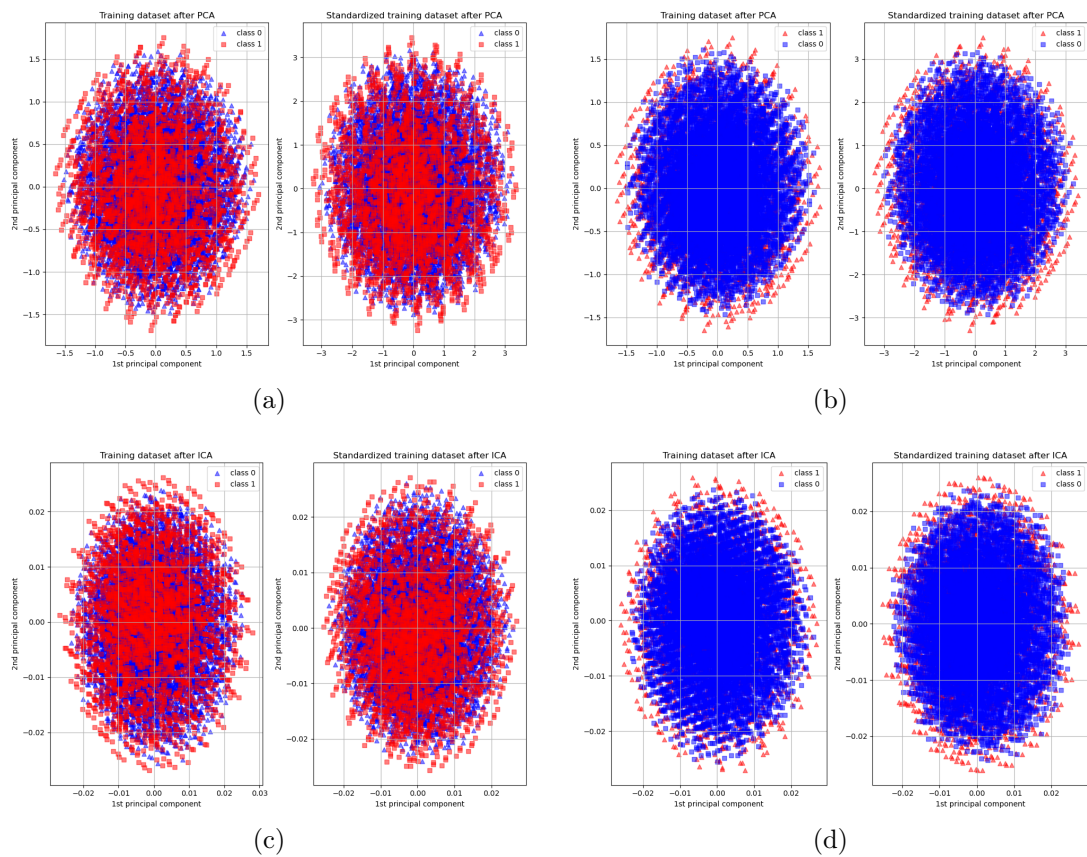


Figure A.3: This set of graphs show PCA and ICA results for unmodified and standardized *2Categs15MFE* dataset. In all graphs red samples represent the positive class and blue represent the negative class. As in Figure 7.1, (a) and (b) contain each two graphs showing the unmodified and standardised dataset after PCA. The difference between the two is caused by the order in which the two classes were added to the graph. The same is true for (c) and (d) except here the graphs show the ICA.

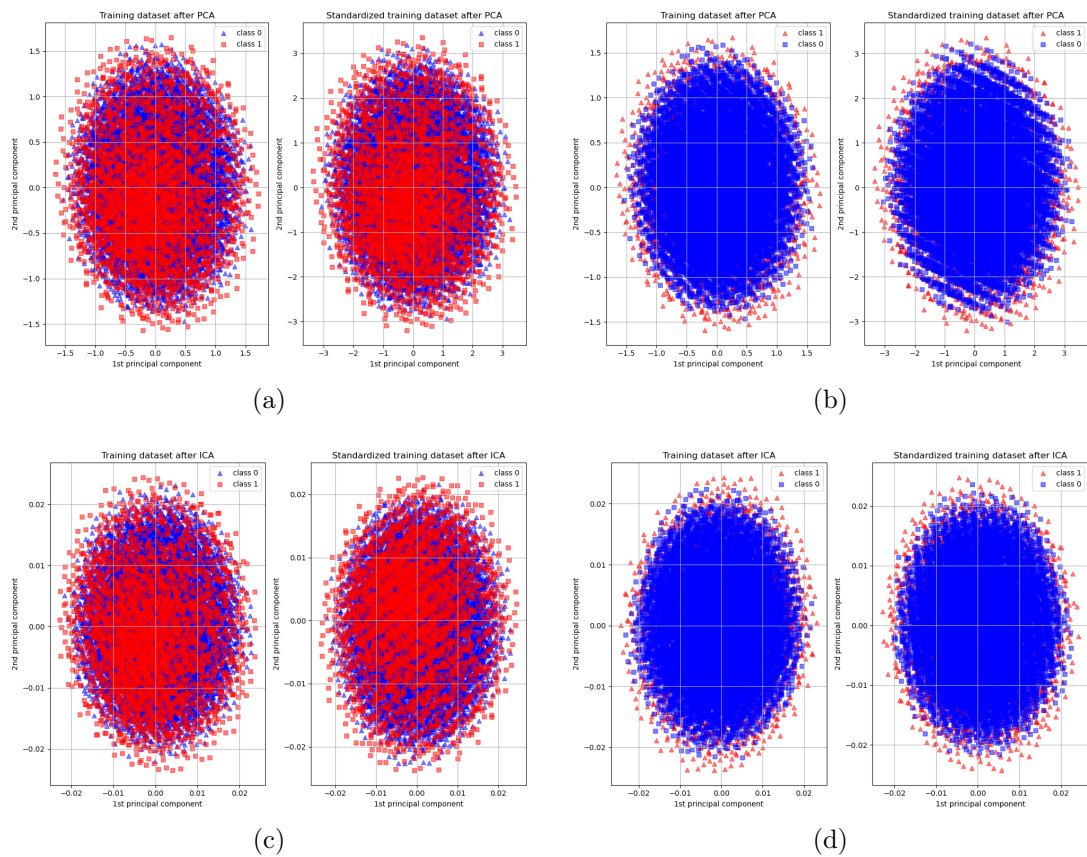


Figure A.4: This set of graphs show PCA and ICA results for unmodified and standardized *2Categs15More* dataset. In all graphs red samples represent the positive class and blue represent the negative class. As in Figure 7.1, (a) and (b) contain each two graphs showing the unmodified and standardised dataset after PCA. The difference between the two is caused by the order in which the two classes were added to the graph. The same is true for (c) and (d) except here the graphs show the ICA.

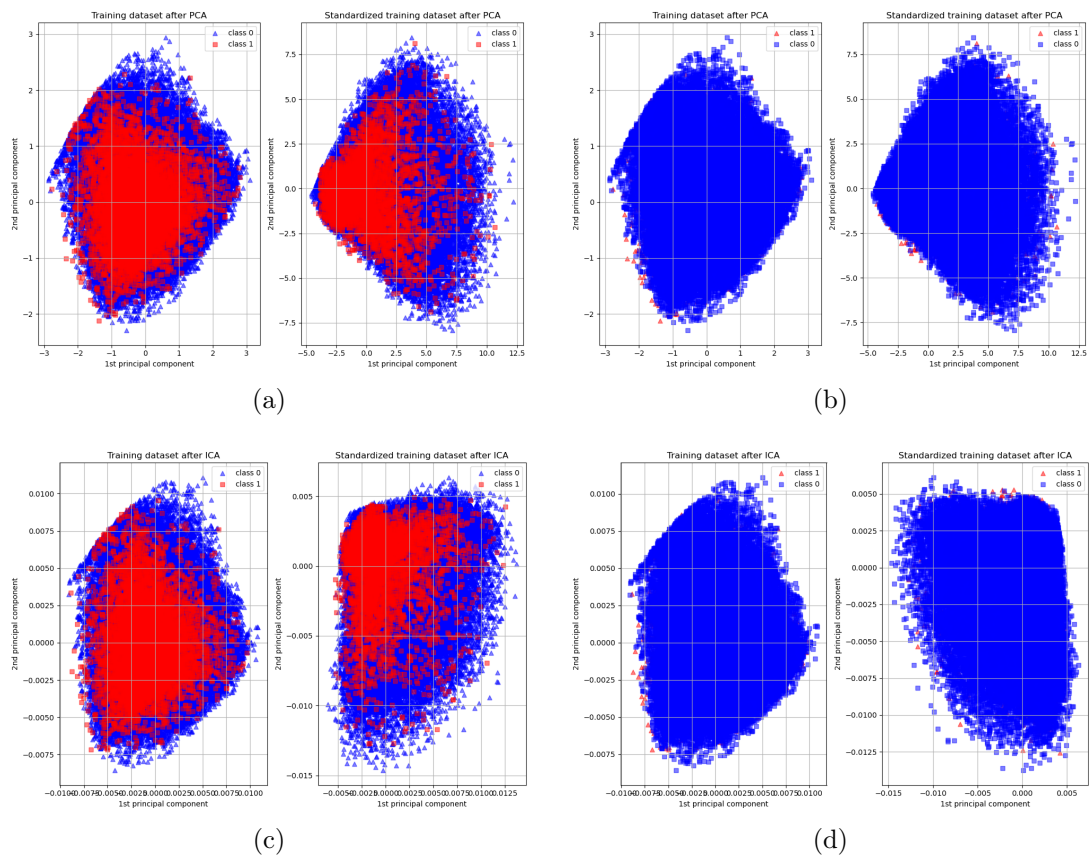


Figure A.5: This set of graphs show PCA and ICA results for unmodified and standardized *3Categs15* dataset. In all graphs red samples represent the positive class and blue represent the negative class. As in Figure 7.1, (a) and (b) contain each two graphs showing the unmodified and standardised dataset after PCA. The difference between the two is caused by the order in which the two classes were added to the graph. The same is true for (c) and (d) except here the graphs show the ICA.

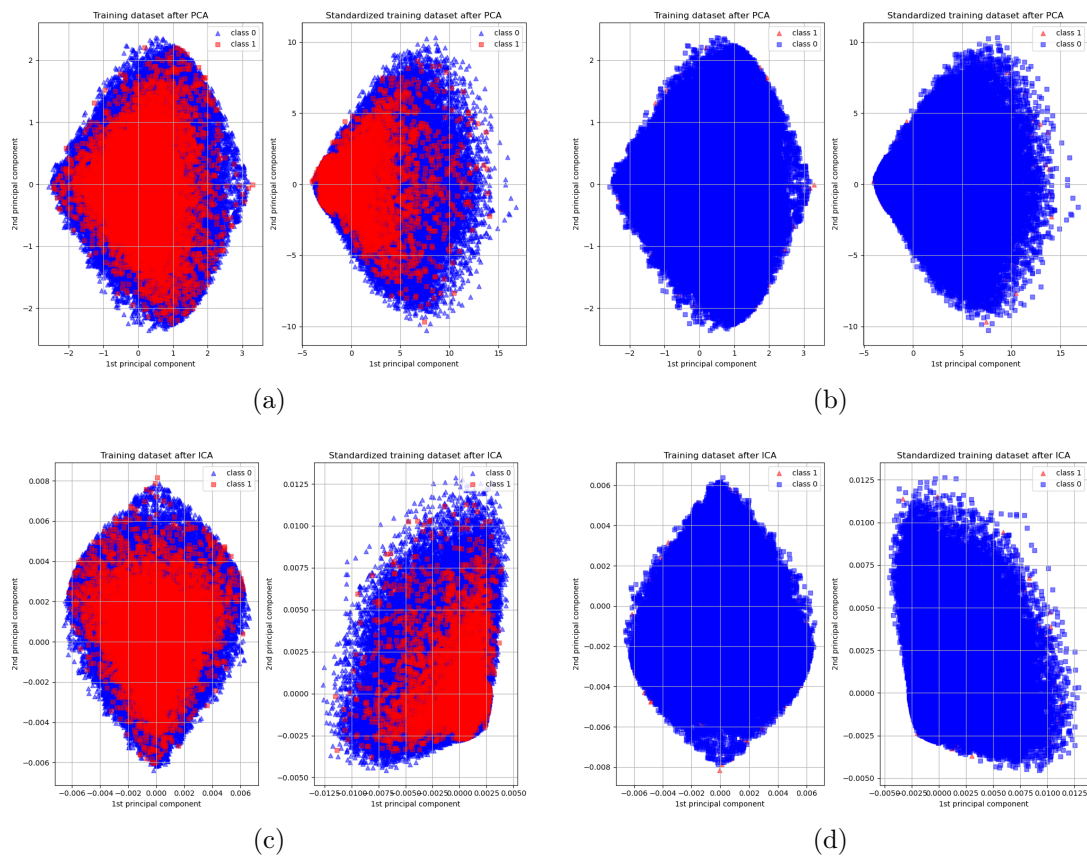


Figure A.6: This set of graphs show PCA and ICA results for unmodified and standardized *3Categs24* dataset. In all graphs red samples represent the positive class and blue represent the negative class. As in Figure 7.1, (a) and (b) contain each two graphs showing the unmodified and standardised dataset after PCA. The difference between the two is caused by the order in which the two classes were added to the graph. The same is true for (c) and (d) except here the graphs show the ICA.

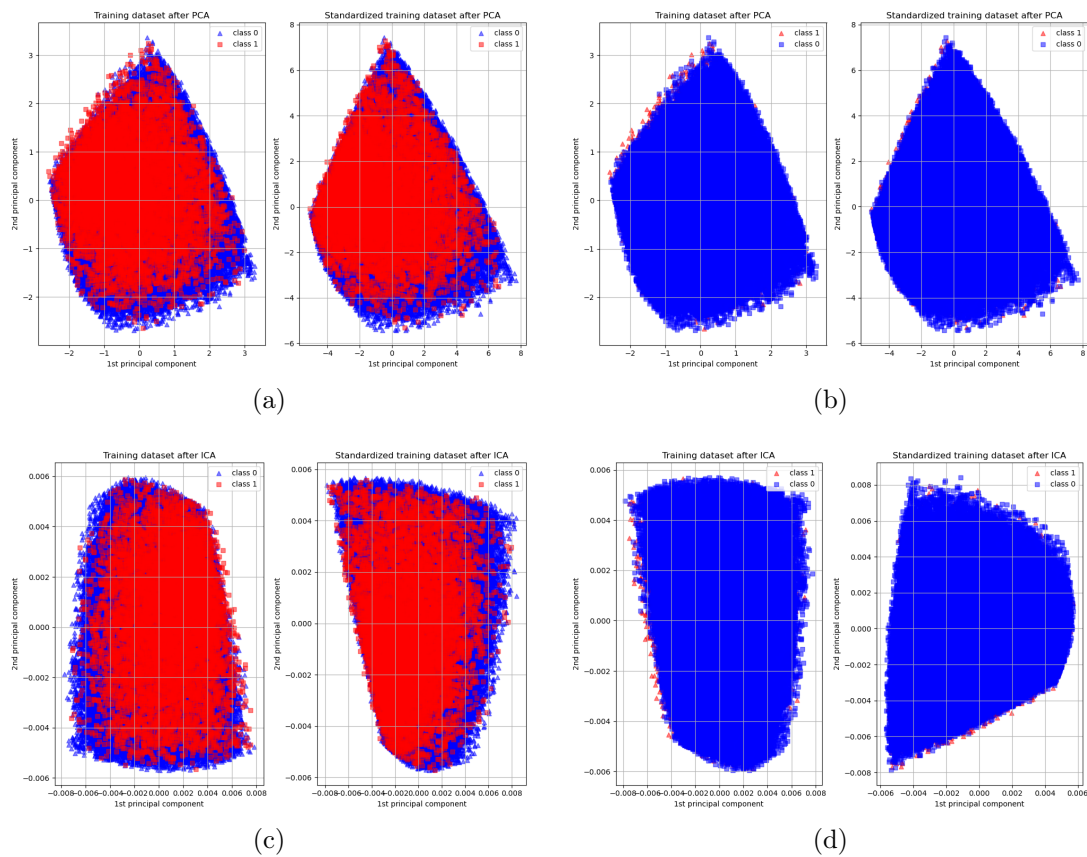


Figure A.7: This set of graphs show PCA and ICA results for unmodified and standardized *5Cats15* dataset. In all graphs red samples represent the positive class and blue represent the negative class. As in Figure 7.1, (a) and (b) contain each two graphs showing the unmodified and standardised dataset after PCA. The difference between the two is caused by the order in which the two classes were added to the graph. The same is true for (c) and (d) except here the graphs show the ICA.

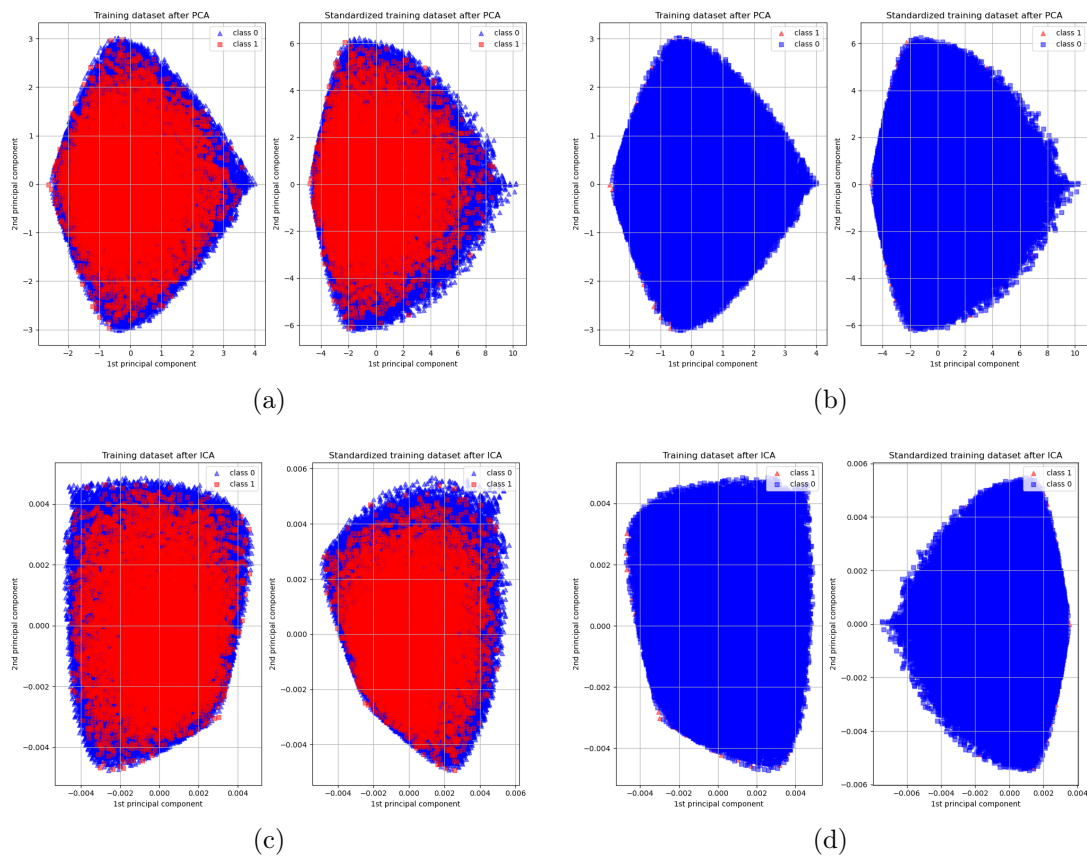


Figure A.8: This set of graphs show PCA and ICA results for unmodified and standardized *5Cats24* dataset. In all graphs red samples represent the positive class and blue represent the negative class. As in Figure 7.1, (a) and (b) contain each two graphs showing the unmodified and standardised dataset after PCA. The difference between the two is caused by the order in which the two classes were added to the graph. The same is true for (c) and (d) except here the graphs show the ICA.

A.2 Secondary Structure Based Classifiers

A.2.1 Initial experiments

All tables in this section follow the same concept. The tables are split in two parts. First part shows the results of classifiers for data that were used to train them. The second part is than set of new data that has never been seen by the classifier. All results contain confusion matrix and SE, SP and Gm scores. Another column called Misclassifications is also included. The sub-column Count is the sum of FP and FN. The sub-column "%" shows the percentage the number of counts make up from the full tested set. The %FN shows the percentage the FN makes up from the fill tested set.

Dataset: <i>2Categs15Centroids</i>										
Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	2414	919	336	696	1255	0.29	0.08	0.67	0.72	0.70
k-NN	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	3250	83	907	125	990	0.23	0.21	0.12	0.98	0.34
Random Forest	3333	0	1031	1	1031	0.24	0.24	0.00	1.00	0.03
Adaboost	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	1781	1552	478	554	2030	0.47	0.11	0.54	0.53	0.54
Testing data										
SVM	1284	972	120	310	1092	0.41	0.03	0.72	0.57	0.64
k-NN	1762	494	230	200	724	0.27	0.05	0.47	0.78	0.60
Decision Tree	2140	116	374	56	490	0.18	0.09	0.13	0.95	0.35
Random Forest	2256	0	430	0	430	0.16	0.10	0.00	1.00	0.00
Adaboost	1657	599	171	259	770	0.29	0.04	0.60	0.73	0.67
Bayes Predictor	1138	1118	221	209	1339	0.50	0.05	0.49	0.50	0.50

Table A.1: Initial Experiment 1 - *2Categs15Centroids*. Parameters: kNN(n_neighbours=2), DT(max_depth=7), RT(max_depth=7), Adaboost(max_depth=7). Further description can be found at the beginning of A.2.1

Dataset: *2Categs15Centroids*

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	2507	826	373	659	1199	0.27	0.09	0.64	0.75	0.69
k-NN	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Random Forest	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Adaboost	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	1781	1552	478	554	2030	0.47	0.11	0.54	0.53	0.54
Testing data										
SVM	1359	897	143	287	1040	0.39	0.03	0.67	0.60	0.63
k-NN	1616	640	185	245	825	0.31	0.04	0.57	0.72	0.64
Decision Tree	1568	688	192	238	880	0.33	0.04	0.55	0.70	0.62
Random Forest	1901	355	233	197	588	0.22	0.05	0.46	0.84	0.62
Adaboost	1575	681	196	234	877	0.33	0.04	0.54	0.70	0.62
Bayes Predictor	1138	1118	221	209	1339	0.50	0.05	0.49	0.50	0.50

Table A.2: Initial Experiment 2 - *2Categs15Centroids*. Parameters: kNN(n_neighbours=3), DT(max_depth=15), RF(max_depth=15), Adaboost(max_depth=15). Further description can be found at the beginning of A.2.1

Dataset: *2Categs15Centroids*

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	2417	916	340	692	1256	0.29	0.08	0.67	0.73	0.70
k-NN	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Random Forest	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Adaboost	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	1781	1552	478	554	2030	0.47	0.11	0.54	0.53	0.54
Testing data										
SVM	1297	959	130	300	1089	0.41	0.03	0.70	0.57	0.63
k-NN	1698	558	188	242	746	0.28	0.04	0.56	0.75	0.65
Decision Tree	1574	682	199	231	881	0.33	0.05	0.54	0.70	0.61
Random Forest	1901	355	233	197	588	0.22	0.05	0.46	0.84	0.62
Adaboost	1575	681	196	234	877	0.33	0.04	0.54	0.70	0.62
Bayes Predictor	1138	1118	221	209	1339	0.50	0.05	0.49	0.50	0.50

Table A.3: Initial Experiment 3 - *2Categs15Centroids*. Parameters: kNN(n_neighbours=5), DT(max_depth=50), RF(max_depth=50), Adaboost(max_depth=50). Further description can be found at the beginning of A.2.1

Dataset: *2Categs15Centroids*

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	2423	910	343	689	1253	0.29	0.08	0.67	0.73	0.70
k-NN	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Random Forest	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Adaboost	3333	0	0	1032	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	1781	1552	478	554	2030	0.47	0.11	0.54	0.53	0.54
Testing data										
SVM	1300	956	121	309	1077	0.40	0.03	0.72	0.58	0.64
k-NN	1812	444	221	209	665	0.25	0.05	0.49	0.80	0.62
Decision Tree	1573	683	194	236	877	0.33	0.04	0.55	0.70	0.62
Random Forest	1901	355	233	197	588	0.22	0.05	0.46	0.84	0.62
Adaboost	1575	681	196	234	877	0.33	0.04	0.54	0.70	0.62
Bayes Predictor	1138	1118	221	209	1339	0.50	0.05	0.49	0.50	0.50

Table A.4: Initial Experiment 4 - *2Categs15Centroids*. Parameters: kNN(n_neighbours=15), DT(max_depth=100), RF(max_depth=100), Adaboost(max_depth=100). Further description can be found at the beginning of A.2.1

Dataset: *2Categs15Less*

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	4935	1312	388	1118	1700	0.22	0.05	0.74	0.79	0.77
k-NN	6247	0	0	1506	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	6162	85	1377	129	1462	0.19	0.18	0.09	0.99	0.29
Random Forest	6247	0	1506	0	1506	0.19	0.19	0.00	1.00	0.00
Adaboost	6247	0	0	1506	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	3325	2922	701	805	3623	0.47	0.09	0.53	0.53	0.53
Testing data										
SVM	2390	1727	130	482	1857	0.39	0.02	0.79	0.58	0.68
k-NN	3338	779	345	267	1124	0.24	0.04	0.44	0.81	0.59
Decision Tree	4000	117	561	51	678	0.14	0.07	0.08	0.97	0.28
Random Forest	4117	0	612	0	612	0.13	0.08	0.00	1.00	0.00
Adaboost	3167	950	240	372	1190	0.25	0.03	0.61	0.77	0.68
Bayes Predictor	2118	1999	306	306	2305	0.49	0.04	0.50	0.51	0.51

Table A.5: Initial Experiment 5 - *2Categs15Less*. Parameters: kNN(n_neighbours=2), DT(max_depth=7), RT(max_depth=7), Adaboost(max_depth=7). Further description can be found at the beginning of A.2.1

Dataset: <i>2Categs15Less</i>										
Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	4832	1415	372	1134	1787	0.23	0.05	0.75	0.77	0.76
k-NN	6247	0	0	1506	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	6247	0	0	1506	0	0.00	0.00	1.00	1.00	1.00
Random Forest	6247	0	1	1505	1	0.00	0.00	1.00	1.00	1.00
Adaboost	6247	0	0	1506	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	3325	2922	701	805	3623	0.47	0.09	0.53	0.53	0.53
Testing data										
SVM	2326	1791	127	485	1918	0.41	0.02	0.79	0.56	0.67
k-NN	3122	995	285	327	1280	0.27	0.04	0.53	0.76	0.64
Decision Tree	3069	1048	275	337	1323	0.28	0.04	0.55	0.75	0.64
Random Forest	3558	559	356	256	915	0.19	0.05	0.42	0.86	0.60
Adaboost	3076	1041	284	328	1325	0.28	0.04	0.54	0.75	0.63
Bayes Predictor	2118	1999	306	306	2305	0.49	0.04	0.50	0.51	0.51

Table A.6: Initial Experiment 6 - *2Categs15Less*. Parameters: kNN(n_neighbours=3), DT(max_depth=15), RF(max_depth=15), Adaboost(max_depth=15). Further description can be found at the beginning of A.2.1

Dataset: <i>2Categs15Less</i>										
Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	4872	1375	378	1128	1753	0.23	0.05	0.75	0.78	0.76
k-NN	6247	0	0	1506	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	6247	0	0	1506	0	0.00	0.00	1.00	1.00	1.00
Random Forest	6247	0	1	1505	1	0.00	0.00	1.00	1.00	1.00
Adaboost	6247	0	0	1506	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	3325	2922	701	805	3623	0.47	0.09	0.53	0.53	0.53
Testing data										
SVM	2349	1768	123	489	1891	0.40	0.02	0.80	0.57	0.68
k-NN	3259	858	307	305	1165	0.25	0.04	0.50	0.79	0.63
Decision Tree	3086	1031	280	332	1311	0.28	0.04	0.54	0.75	0.64
Random Forest	3558	559	356	256	915	0.19	0.05	0.42	0.86	0.60
Adaboost	3076	1041	284	328	1325	0.28	0.04	0.54	0.75	0.63
Bayes Predictor	2118	1999	306	306	2305	0.49	0.04	0.50	0.51	0.51

Table A.7: Initial Experiment 7 - *2Categs15Less*. Parameters: kNN(n_neighbours=5), DT(max_depth=50), RF(max_depth=50), Adaboost(max_depth=50). Further description can be found at the beginning of A.2.1

Dataset: <i>2Categs15Less</i>										
Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	4906	1341	379	1127	1720	0.22	0.05	0.75	0.79	0.77
k-NN	6247	0	0	1506	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	6247	0	0	1506	0	0.00	0.00	1.00	1.00	1.00
Random Forest	6247	0	1	1505	1	0.00	0.00	1.00	1.00	1.00
Adaboost	6247	0	0	1506	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	3325	2922	701	805	3623	0.47	0.09	0.53	0.53	0.53
Testing data										
SVM	2349	1768	130	482	1898	0.40	0.02	0.79	0.57	0.67
k-NN	3409	708	326	286	1034	0.22	0.04	0.47	0.83	0.62
Decision Tree	3071	1046	282	330	1328	0.28	0.04	0.54	0.75	0.63
Random Forest	3558	559	356	256	915	0.19	0.05	0.42	0.86	0.60
Adaboost	3076	1041	284	328	1325	0.28	0.04	0.54	0.75	0.63
Bayes Predictor	2118	1999	306	306	2305	0.49	0.04	0.50	0.51	0.51

Table A.8: Initial Experiment 8 - *2Categs15Less*. Parameters: kNN(n_neighbours=15), DT(max_depth=100), RF(max_depth=100), Adaboost(max_depth=100). Further description can be found at the beginning of A.2.1

Dataset: <i>2Categs15MFE</i>										
Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	2729	815	366	918	1181	0.24	0.08	0.71	0.77	0.74
k-NN	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	3266	278	932	352	1210	0.25	0.19	0.27	0.92	0.50
Random Forest	3544	0	1267	17	1267	0.26	0.26	0.01	1.00	0.12
Adaboost	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	1864	1680	612	672	2292	0.47	0.13	0.52	0.53	0.52
Testing data										
SVM	1471	1085	129	404	1214	0.39	0.03	0.76	0.58	0.66
k-NN	1972	584	237	296	821	0.27	0.05	0.56	0.77	0.65
Decision Tree	2204	352	396	137	748	0.24	0.08	0.26	0.86	0.47
Random Forest	2556	0	519	14	519	0.17	0.11	0.03	1.00	0.16
Adaboost	1846	710	160	373	870	0.28	0.03	0.70	0.72	0.71
Bayes Predictor	1300	1256	269	264	1525	0.49	0.06	0.50	0.51	0.50

Table A.9: Initial Experiment 9 - *2Categs15MFE*. Parameters: kNN(n_neighbours=2), DT(max_depth=7), RT(max_depth=7), Adaboost(max_depth=7). Further description can be found at the beginning of A.2.1

Dataset: 2Categs15MFE										
Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	2761	783	389	895	1172	0.24	0.08	0.70	0.78	0.74
k-NN	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Random Forest	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Adaboost	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	1864	1680	612	672	2292	0.47	0.13	0.52	0.53	0.52
Testing data										
SVM	1507	1049	138	395	1187	0.38	0.03	0.74	0.59	0.66
k-NN	1756	800	176	357	976	0.32	0.04	0.67	0.69	0.68
Decision Tree	1802	754	180	353	934	0.30	0.04	0.66	0.71	0.68
Random Forest	2081	475	222	311	697	0.23	0.05	0.58	0.81	0.69
Adaboost	1780	776	176	357	952	0.31	0.04	0.67	0.70	0.68
Bayes Predictor	1300	1256	269	264	1525	0.49	0.06	0.50	0.51	0.50

Table A.10: Initial Experiment 10 - *2Categs15MFE*. Parameters: kNN(n_neighbours=3), DT(max_depth=15), RF(max_depth=15), Adaboost(max_depth=15). Further description can be found at the beginning of A.2.1

Dataset: 2Categs15MFE										
Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	2721	823	364	920	1187	0.25	0.08	0.72	0.77	0.74
k-NN	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Random Forest	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Adaboost	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	1864	1680	612	672	2292	0.47	0.13	0.52	0.53	0.52
Testing data										
SVM	1448	1108	136	397	1244	0.40	0.03	0.74	0.57	0.65
k-NN	1822	734	186	347	920	0.30	0.04	0.65	0.71	0.68
Decision Tree	1786	770	172	361	942	0.30	0.04	0.68	0.70	0.69
Random Forest	2081	475	222	311	697	0.23	0.05	0.58	0.81	0.69
Adaboost	1780	776	176	357	952	0.31	0.04	0.67	0.70	0.68
Bayes Predictor	1300	1256	269	264	1525	0.49	0.06	0.50	0.51	0.50

Table A.11: Initial Experiment 11 - *2Categs15MFE*. Parameters: kNN(n_neighbours=5), DT(max_depth=50), RF(max_depth=50), Adaboost(max_depth=50). Further description can be found at the beginning of A.2.1

Dataset: 2Categs15MFE

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	2773	771	393	891	1164	0.24	0.08	0.69	0.78	0.74
k-NN	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Random Forest	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Adaboost	3544	0	0	1284	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	1864	1680	612	672	2292	0.47	0.13	0.52	0.53	0.52
Testing data										
SVM	1515	1041	137	396	1178	0.38	0.03	0.74	0.59	0.66
k-NN	1947	609	214	319	823	0.27	0.04	0.60	0.76	0.68
Decision Tree	1785	771	183	350	954	0.31	0.04	0.66	0.70	0.68
Random Forest	2081	475	222	311	697	0.23	0.05	0.58	0.81	0.69
Adaboost	1780	776	176	357	952	0.31	0.04	0.67	0.70	0.68
Bayes Predictor	1300	1256	269	264	1525	0.49	0.06	0.50	0.51	0.50

Table A.12: Initial Experiment 12 - *2Categs15MFE*. Parameters: kNN(n_neighbours=15), DT(max_depth=100), RF(max_depth=100), Adaboost(max_depth=100). Further description can be found at the beginning of A.2.1

Dataset: 2Categs15More

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	3245	924	369	812	1293	0.24	0.07	0.69	0.78	0.73
k-NN	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	4064	105	1010	171	1115	0.21	0.19	0.14	0.97	0.38
Random Forest	4169	0	1163	18	1163	0.22	0.22	0.02	1.00	0.12
Adaboost	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	2182	1987	536	645	2523	0.47	0.10	0.55	0.52	0.53
Testing data										
SVM	1724	1108	97	395	1205	0.36	0.02	0.80	0.61	0.70
k-NN	2350	482	254	238	736	0.22	0.05	0.48	0.83	0.63
Decision Tree	2675	157	407	85	564	0.17	0.08	0.17	0.94	0.40
Random Forest	2832	0	483	9	483	0.15	0.09	0.02	1.00	0.14
Adaboost	2169	663	176	316	839	0.25	0.03	0.64	0.77	0.70
Bayes Predictor	1418	1414	243	249	1657	0.50	0.05	0.51	0.50	0.50

Table A.13: Initial Experiment 13 - *2Categs15More*. Parameters: kNN(n_neighbours=2), DT(max_depth=7), RT(max_depth=7), Adaboost(max_depth=7). Further description can be found at the beginning of A.2.1

Dataset: <i>2Categs15More</i>										
Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	3191	978	365	816	1343	0.25	0.07	0.69	0.77	0.73
k-NN	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Random Forest	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Adaboost	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	2182	1987	536	645	2523	0.47	0.10	0.55	0.52	0.53
Testing data										
SVM	1674	1158	98	394	1256	0.38	0.02	0.80	0.59	0.69
k-NN	2163	669	203	289	872	0.26	0.04	0.59	0.76	0.67
Decision Tree	2103	729	188	304	917	0.28	0.04	0.62	0.74	0.68
Random Forest	2445	387	231	261	618	0.19	0.04	0.53	0.86	0.68
Adaboost	2108	724	192	300	916	0.28	0.04	0.61	0.74	0.67
Bayes Predictor	1418	1414	243	249	1657	0.50	0.05	0.51	0.50	0.50

Table A.14: Initial Experiment 14 - *2Categs15More*. Parameters: kNN(n_neighbours=3), DT(max_depth=15), RF(max_depth=15), Adaboost(max_depth=15). Further description can be found at the beginning of A.2.1

Dataset: <i>2Categs15More</i>										
Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	3267	902	369	812	1271	0.24	0.07	0.69	0.78	0.73
k-NN	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Random Forest	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Adaboost	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	2182	1987	536	645	2523	0.47	0.10	0.55	0.52	0.53
Testing data										
SVM	1755	1077	103	389	1180	0.35	0.02	0.79	0.62	0.70
k-NN	2247	585	208	284	793	0.24	0.04	0.58	0.79	0.68
Decision Tree	2110	722	185	307	907	0.27	0.03	0.62	0.75	0.68
Random Forest	2445	387	231	261	618	0.19	0.04	0.53	0.86	0.68
Adaboost	2108	724	192	300	916	0.28	0.04	0.61	0.74	0.67
Bayes Predictor	1418	1414	243	249	1657	0.50	0.05	0.51	0.50	0.50

Table A.15: Initial Experiment 15 - *2Categs15More*. Parameters: kNN(n_neighbours=5), DT(max_depth=50), RF(max_depth=50), Adaboost(max_depth=50). Further description can be found at the beginning of A.2.1

Dataset: *2Categs15More*

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	3230	939	361	820	1300	0.24	0.07	0.69	0.77	0.73
k-NN	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Random Forest	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Adaboost	4169	0	0	1181	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	2182	1987	536	645	2523	0.47	0.10	0.55	0.52	0.53
Testing data										
SVM	1696	1136	100	392	1236	0.37	0.02	0.80	0.60	0.69
k-NN	2392	440	223	269	663	0.20	0.04	0.55	0.84	0.68
Decision Tree	2117	715	197	295	912	0.27	0.04	0.60	0.75	0.67
Random Forest	2445	387	231	261	618	0.19	0.04	0.53	0.86	0.68
Adaboost	2108	724	192	300	916	0.28	0.04	0.61	0.74	0.67
Bayes Predictor	1418	1414	243	249	1657	0.50	0.05	0.51	0.50	0.50

Table A.16: Initial Experiment 16 - *2Categs15More*. Parameters: kNN(n_neighbours=15), DT(max_depth=100), RF(max_depth=100), Adaboost(max_depth=100). Further description can be found at the beginning of A.2.1

Dataset: *3Categs15*

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	25810	20941	775	1921	21716	0.44	0.02	0.71	0.55	0.63
k-NN	46751	0	0	2696	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	46750	1	2687	9	2688	0.05	0.05	0.00	1.00	0.06
Random Forest	46751	0	2696	0	2696	0.05	0.05	0.00	1.00	0.00
Adaboost	46621	130	954	1742	1084	0.02	0.02	0.65	1.00	0.80
Bayes Predictor	25076	21675	839	1857	22514	0.46	0.02	0.69	0.54	0.61
Testing data										
SVM	6372	7757	242	543	7999	0.54	0.00	0.69	0.45	0.56
k-NN	13207	922	686	99	1608	0.11	0.01	0.13	0.93	0.34
Decision Tree	14119	10	785	0	795	0.05	0.02	0.00	1.00	0.00
Random Forest	14129	0	785	0	785	0.05	0.02	0.00	1.00	0.00
Adaboost	13620	509	685	100	1194	0.08	0.01	0.13	0.96	0.35
Bayes Predictor	6459	7670	261	524	7931	0.53	0.01	0.67	0.46	0.55

Table A.17: Initial Experiment 17 - *3Categs15*. Parameters: kNN(n_neighbours=2), DT(max_depth=7), RT(max_depth=7), Adaboost(max_depth=7). Further description can be found at the beginning of A.2.1

Dataset: 3Categs15

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	25867	20884	798	1898	21682	0.44	0.02	0.70	0.55	0.62
k-NN	46751	0	0	2696	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	46736	15	2162	534	2177	0.04	0.04	0.20	1.00	0.44
Random Forest	46751	0	2587	109	2587	0.05	0.05	0.04	1.00	0.20
Adaboost	46751	0	0	2696	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	25076	21675	839	1857	22514	0.46	0.02	0.69	0.54	0.61
Testing data										
SVM	6414	7715	245	540	7960	0.53	0.00	0.69	0.45	0.56
k-NN	13089	1040	665	120	1705	0.11	0.01	0.15	0.93	0.38
Decision Tree	13892	237	781	4	1018	0.07	0.02	0.01	0.98	0.07
Random Forest	14129	0	785	0	785	0.05	0.02	0.00	1.00	0.00
Adaboost	13939	190	755	30	945	0.06	0.02	0.04	0.99	0.19
Bayes Predictor	6459	7670	261	524	7931	0.53	0.01	0.67	0.46	0.55

Table A.18: Initial Experiment 18 - *3Categs15*. Parameters: kNN(n_neighbours=3), DT(max_depth=15), RF(max_depth=15), Adaboost(max_depth=15). Further description can be found at the beginning of A.2.1

Dataset: 3Categs15

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	25604	21147	782	1914	21929	0.44	0.02	0.71	0.55	0.62
k-NN	46751	0	0	2696	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	46750	1	1	2695	2	0.00	0.00	1.00	1.00	1.00
Random Forest	46751	0	0	2696	0	0.00	0.00	1.00	1.00	1.00
Adaboost	46751	0	0	2696	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	25076	21675	839	1857	22514	0.46	0.02	0.69	0.54	0.61
Testing data										
SVM	6277	7852	240	545	8092	0.54	0.00	0.69	0.44	0.56
k-NN	13513	616	715	70	1331	0.09	0.01	0.09	0.96	0.29
Decision Tree	12865	1264	649	136	1913	0.13	0.01	0.17	0.91	0.40
Random Forest	14111	18	778	7	796	0.05	0.02	0.01	1.00	0.09
Adaboost	13881	248	742	43	990	0.07	0.02	0.05	0.98	0.23
Bayes Predictor	6459	7670	261	524	7931	0.53	0.01	0.67	0.46	0.55

Table A.19: Initial Experiment 19 - *3Categs15*. Parameters: kNN(n_neighbours=5), DT(max_depth=50), RF(max_depth=50), Adaboost(max_depth=50). Further description can be found at the beginning of A.2.1

Dataset: 3Categs15

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	25029	21722	767	1929	22489	0.45	0.02	0.72	0.54	0.62
k-NN	46751	0	0	2696	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	46751	0	0	2696	0	0.00	0.00	1.00	1.00	1.00
Random Forest	46751	0	2	2694	2	0.00	0.00	1.00	1.00	1.00
Adaboost	46751	0	0	2696	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	25076	21675	839	1857	22514	0.46	0.02	0.69	0.54	0.61
Testing data										
SVM	6143	7986	230	555	8216	0.55	0.00	0.71	0.43	0.55
k-NN	14081	48	777	8	825	0.06	0.02	0.01	1.00	0.10
Decision Tree	12809	1320	658	127	1978	0.13	0.01	0.16	0.91	0.38
Random Forest	14116	13	778	7	791	0.05	0.02	0.01	1.00	0.09
Adaboost	12822	1307	659	126	1966	0.13	0.01	0.16	0.91	0.38
Bayes Predictor	6459	7670	261	524	7931	0.53	0.01	0.67	0.46	0.55

Table A.20: Initial Experiment 20 - *3Categs15*. Parameters: kNN(n_neighbours=15), DT(max_depth=100), RF(max_depth=100), Adaboost(max_depth=100). Further description can be found at the beginning of A.2.1

Dataset: 5Categs15

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	31798	22908	1178	1885	24086	0.42	0.02	0.62	0.58	0.60
k-NN	54706	0	0	3063	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	54706	0	3063	0	3063	0.05	0.05	0.00	1.00	0.00
Random Forest	54706	0	3063	0	3063	0.05	0.05	0.00	1.00	0.00
Adaboost	54640	66	519	2544	585	0.01	0.01	0.83	1.00	0.91
Bayes Predictor	29846	24860	1164	1899	26024	0.45	0.02	0.62	0.55	0.58
Testing data										
SVM	8773	8407	298	522	8705	0.48	0.01	0.64	0.51	0.57
k-NN	16418	762	762	58	1524	0.08	0.01	0.07	0.96	0.26
Decision Tree	17180	0	820	0	820	0.05	0.01	0.00	1.00	0.00
Random Forest	17180	0	820	0	820	0.05	0.01	0.00	1.00	0.00
Adaboost	16625	555	750	70	1305	0.07	0.01	0.09	0.97	0.29
Bayes Predictor	8269	8911	278	542	9189	0.51	0.00	0.66	0.48	0.56

Table A.21: Initial Experiment 21 - *5Categs15*. Parameters: kNN(n_neighbours=2), DT(max_depth=7), RT(max_depth=7), Adaboost(max_depth=7). Further description can be found at the beginning of A.2.1

Dataset: 5Categs15

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	31800	22906	1179	1884	24085	0.42	0.02	0.62	0.58	0.60
k-NN	54706	0	0	3063	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	54656	50	1627	1436	1677	0.03	0.03	0.47	1.00	0.68
Random Forest	54706	0	2694	369	2694	0.05	0.05	0.12	1.00	0.35
Adaboost	54706	0	0	3063	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	29846	24860	1164	1899	26024	0.45	0.02	0.62	0.55	0.58
Testing data										
SVM	8764	8416	294	526	8710	0.48	0.01	0.64	0.51	0.57
k-NN	16508	672	750	70	1422	0.08	0.01	0.09	0.96	0.29
Decision Tree	16543	637	795	25	1432	0.08	0.01	0.03	0.96	0.17
Random Forest	17180	0	819	1	819	0.05	0.01	0.00	1.00	0.03
Adaboost	17117	63	811	9	874	0.05	0.01	0.01	1.00	0.10
Bayes Predictor	8269	8911	278	542	9189	0.51	0.00	0.66	0.48	0.56

Table A.22: Initial Experiment 22 - 5Categs15. Parameters: kNN(n_neighbours=3), DT(max_depth=15), RF(max_depth=15), Adaboost(max_depth=15). Further description can be found at the beginning of A.2.1

Dataset: 5Categs15

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	31838	22868	1193	1870	24061	0.42	0.02	0.61	0.58	0.60
k-NN	54706	0	0	3063	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	54706	0	0	3063	0	0.00	0.00	1.00	1.00	1.00
Random Forest	54706	0	2	3061	2	0.00	0.00	1.00	1.00	1.00
Adaboost	54706	0	0	3063	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	29846	24860	1164	1899	26024	0.45	0.02	0.62	0.55	0.58
Testing data										
SVM	8783	8397	303	517	8700	0.48	0.01	0.63	0.51	0.57
k-NN	16889	291	782	38	1073	0.06	0.01	0.05	0.98	0.21
Decision Tree	15730	1450	732	88	2182	0.12	0.01	0.11	0.92	0.31
Random Forest	17174	6	817	3	823	0.05	0.01	0.00	1.00	0.06
Adaboost	15706	1474	723	97	2197	0.12	0.01	0.12	0.91	0.33
Bayes Predictor	8269	8911	278	542	9189	0.51	0.00	0.66	0.48	0.56

Table A.23: Initial Experiment 23 - 5Categs15. Parameters: kNN(n_neighbours=5), DT(max_depth=50), RF(max_depth=50), Adaboost(max_depth=50). Further description can be found at the beginning of A.2.1

Dataset: 5Categs15

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	30919	23787	1154	1909	24941	0.43	0.02	0.62	0.57	0.59
k-NN	54706	0	0	3063	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	54706	0	0	3063	0	0.00	0.00	1.00	1.00	1.00
Random Forest	54706	0	2	3061	2	0.00	0.00	1.00	1.00	1.00
Adaboost	54706	0	0	3063	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	29846	24860	1164	1899	26024	0.45	0.02	0.62	0.55	0.58
Testing data										
SVM	8544	8636	294	526	8930	0.50	0.01	0.64	0.50	0.56
k-NN	17166	14	816	4	830	0.05	0.01	0.00	1.00	0.07
Decision Tree	15749	1431	724	96	2155	0.12	0.01	0.12	0.92	0.33
Random Forest	17174	6	817	3	823	0.05	0.01	0.00	1.00	0.06
Adaboost	15706	1474	723	97	2197	0.12	0.01	0.12	0.91	0.33
Bayes Predictor	8269	8911	278	542	9189	0.51	0.00	0.66	0.48	0.56

Table A.24: Initial Experiment 24 - *5Categs15*. Parameters: kNN(n_neighbours=15), DT(max_depth=100), RF(max_depth=100), Adaboost(max_depth=100). Further description can be found at the beginning of A.2.1

Dataset: 3Categs24

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	39886	55862	1195	2428	57057	0.57	0.01	0.67	0.42	0.53
k-NN	95748	0	0	3623	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	95744	4	3603	20	3607	0.04	0.04	0.01	1.00	0.07
Random Forest	95748	0	3623	0	3623	0.04	0.04	0.00	1.00	0.00
Adaboost	95693	55	2443	1180	2498	0.03	0.02	0.33	1.00	0.57
Bayes Predictor	42468	53280	1357	2266	54637	0.55	0.01	0.63	0.44	0.53
Testing data										
SVM	9731	14823	310	562	15133	0.60	0.00	0.64	0.40	0.51
k-NN	23730	824	832	40	1656	0.07	0.01	0.05	0.97	0.21
Decision Tree	24543	11	872	0	883	0.03	0.01	0.00	1.00	0.00
Random Forest	24554	0	872	0	872	0.03	0.01	0.00	1.00	0.00
Adaboost	24371	183	865	7	1048	0.04	0.01	0.01	0.99	0.09
Bayes Predictor	10392	14162	325	547	14487	0.57	0.00	0.63	0.42	0.52

Table A.25: Initial Experiment 25 - *3Categs24*. Parameters: kNN(n_neighbours=2), DT(max_depth=7), RT(max_depth=7), Adaboost(max_depth=7). Further description can be found at the beginning of A.2.1

Dataset: 3Categs24

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	42600	53148	1245	2378	54393	0.55	0.01	0.66	0.44	0.54
k-NN	95748	0	0	3623	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	95731	17	2843	780	2860	0.03	0.03	0.22	1.00	0.46
Random Forest	95748	0	3539	84	3539	0.04	0.04	0.02	1.00	0.15
Adaboost	95748	0	0	3623	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	42468	53280	1357	2266	54637	0.55	0.01	0.63	0.44	0.53
Testing data										
SVM	10524	14030	328	544	14358	0.56	0.00	0.62	0.43	0.52
k-NN	23318	1236	820	52	2056	0.08	0.01	0.06	0.95	0.24
Decision Tree	24302	252	864	8	1116	0.04	0.01	0.01	0.99	0.10
Random Forest	24554	0	872	0	872	0.03	0.01	0.00	1.00	0.00
Adaboost	24176	378	849	23	1227	0.05	0.01	0.03	0.98	0.16
Bayes Predictor	10392	14162	325	547	14487	0.57	0.00	0.63	0.42	0.52

Table A.26: Initial Experiment 26 - 3Categs24. Parameters: kNN(n_neighbours=3), DT(max_depth=15), RF(max_depth=15), Adaboost(max_depth=15) Further description can be found at the beginning of A.2.1

Dataset: 3Categs24

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	39591	56157	1179	2444	57336	0.58	0.01	0.67	0.41	0.53
k-NN	95748	0	0	3623	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	95747	1	45	3578	46	0.00	0.00	0.99	1.00	0.99
Random Forest	95748	0	45	3578	45	0.00	0.00	0.99	1.00	0.99
Adaboost	95748	0	0	3623	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	42468	53280	1357	2266	54637	0.55	0.01	0.63	0.44	0.53
Testing data										
SVM	9628	14926	313	559	15239	0.60	0.00	0.64	0.39	0.50
k-NN	24405	149	864	8	1013	0.04	0.01	0.01	0.99	0.10
Decision Tree	23070	1484	819	53	2303	0.09	0.01	0.06	0.94	0.24
Random Forest	24441	113	865	7	978	0.04	0.01	0.01	1.00	0.09
Adaboost	24320	234	858	14	1092	0.04	0.01	0.02	0.99	0.13
Bayes Predictor	10392	14162	325	547	14487	0.57	0.00	0.63	0.42	0.52

Table A.27: Initial Experiment 27 - 3Categs24. Parameters: kNN(n_neighbours=5), DT(max_depth=50), RF(max_depth=50), Adaboost(max_depth=50). Further description can be found at the beginning of A.2.1

Dataset: 3Categs24

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	40108	55640	1194	2429	56834	0.57	0.01	0.67	0.42	0.53
k-NN	95748	0	0	3623	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	95748	0	0	3623	0	0.00	0.00	1.00	1.00	1.00
Random Forest	95748	0	7	3616	7	0.00	0.00	1.00	1.00	1.00
Adaboost	95748	0	0	3623	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	42468	53280	1357	2266	54637	0.55	0.01	0.63	0.44	0.53
Testing data										
SVM	9851	14703	308	564	15011	0.59	0.00	0.65	0.40	0.51
k-NN	24434	120	864	8	984	0.04	0.01	0.01	1.00	0.10
Decision Tree	23036	1518	811	61	2329	0.09	0.01	0.07	0.94	0.26
Random Forest	24432	122	863	9	985	0.04	0.01	0.01	1.00	0.10
Adaboost	23042	1512	819	53	2331	0.09	0.01	0.06	0.94	0.24
Bayes Predictor	10392	14162	325	547	14487	0.57	0.00	0.63	0.42	0.52

Table A.28: Initial Experiment 28 - 3Categs24. Parameters: kNN(n_neighbours=15), DT(max_depth=100), RF(max_depth=100), Adaboost(max_depth=100). Further description can be found at the beginning of A.2.1

Dataset: 5Categs24

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	53306	50130	1576	1976	51706	0.48	0.01	0.56	0.52	0.54
k-NN	103436	0	0	3552	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	103436	0	3552	0	3552	0.03	0.03	0.00	1.00	0.00
Random Forest	103436	0	3552	0	3552	0.03	0.03	0.00	1.00	0.00
Adaboost	103344	92	1989	1563	2081	0.02	0.02	0.44	1.00	0.66
Bayes Predictor	53720	49716	1630	1922	51346	0.48	0.02	0.54	0.52	0.53
Testing data										
SVM	13054	12588	448	456	13036	0.49	0.00	0.50	0.51	0.51
k-NN	25041	601	875	29	1476	0.06	0.01	0.03	0.98	0.18
Decision Tree	25642	0	904	0	904	0.03	0.01	0.00	1.00	0.00
Random Forest	25642	0	904	0	904	0.03	0.01	0.00	1.00	0.00
Adaboost	25433	209	895	9	1104	0.04	0.01	0.01	0.99	0.10
Bayes Predictor	13135	12507	460	444	12967	0.49	0.00	0.49	0.51	0.50

Table A.29: Initial Experiment 29 - 5Categs24. Parameters: kNN(n_neighbours=2), DT(max_depth=7), RT(max_depth=7), Adaboost(max_depth=7). Further description can be found at the beginning of A.2.1

Dataset: 5Categs24

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	47024	56412	1349	2203	57761	0.54	0.01	0.62	0.45	0.53
k-NN	103436	0	0	3552	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	103382	54	2011	1541	2065	0.02	0.02	0.43	1.00	0.66
Random Forest	103436	0	3374	178	3374	0.03	0.03	0.05	1.00	0.22
Adaboost	103436	0	0	3552	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	53720	49716	1630	1922	51346	0.48	0.02	0.54	0.52	0.53
Testing data										
SVM	11461	14181	402	502	14583	0.55	0.00	0.56	0.45	0.50
k-NN	24864	778	863	41	1641	0.06	0.01	0.05	0.97	0.21
Decision Tree	25031	611	875	29	1486	0.06	0.01	0.03	0.98	0.18
Random Forest	25642	0	904	0	904	0.03	0.01	0.00	1.00	0.00
Adaboost	25499	143	897	7	1040	0.04	0.01	0.01	0.99	0.09
Bayes Predictor	13135	12507	460	444	12967	0.49	0.00	0.49	0.51	0.50

Table A.30: Initial Experiment 30 - 5Categs24. Parameters: kNN(n_neighbours=3), DT(max_depth=15), RF(max_depth=15), Adaboost(max_depth=15). Further description can be found at the beginning of A.2.1

Dataset: 5Categs24

Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	53439	49997	1534	2018	51531	0.48	0.01	0.57	0.52	0.54
k-NN	103436	0	0	3552	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	103436	0	0	3552	0	0.00	0.00	1.00	1.00	1.00
Random Forest	103436	0	9	3543	9	0.00	0.00	1.00	1.00	1.00
Adaboost	103436	0	0	3552	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	53720	49716	1630	1922	51346	0.48	0.02	0.54	0.52	0.53
Testing data										
SVM	13081	12561	459	445	13020	0.49	0.00	0.49	0.51	0.50
k-NN	25591	51	901	3	952	0.04	0.01	0.00	1.00	0.06
Decision Tree	24252	1390	849	55	2239	0.08	0.01	0.06	0.95	0.24
Random Forest	25597	45	902	2	947	0.04	0.01	0.00	1.00	0.05
Adaboost	24244	1398	850	54	2248	0.08	0.01	0.06	0.95	0.24
Bayes Predictor	13135	12507	460	444	12967	0.49	0.00	0.49	0.51	0.50

Table A.31: Initial Experiment 31 - 5Categs24. Parameters: kNN(n_neighbours=5), DT(max_depth=50), RF(max_depth=50), Adaboost(max_depth=50). Further description can be found at the beginning of A.2.1

Dataset: 5Categs24										
Classifier	Confusion matrix				Misclassifications			Scores		
	TN	FP	FN	TP	Count	%	%FN	SE	SP	Gm
Training data										
SVM	57382	46054	1717	1835	47771	0.45	0.02	0.52	0.55	0.54
k-NN	103436	0	0	3552	0	0.00	0.00	1.00	1.00	1.00
Decision Tree	103436	0	0	3552	0	0.00	0.00	1.00	1.00	1.00
Random Forest	103436	0	9	3543	9	0.00	0.00	1.00	1.00	1.00
Adaboost	103436	0	0	3552	0	0.00	0.00	1.00	1.00	1.00
Bayes Predictor	53720	49716	1630	1922	51346	0.48	0.02	0.54	0.52	0.53
Testing data										
SVM	14352	11290	505	399	11795	0.44	0.00	0.44	0.56	0.50
k-NN	25599	43	902	2	945	0.04	0.01	0.00	1.00	0.05
Decision Tree	24283	1359	850	54	2209	0.08	0.01	0.06	0.95	0.24
Random Forest	25597	45	902	2	947	0.04	0.01	0.00	1.00	0.05
Adaboost	24244	1398	850	54	2248	0.08	0.01	0.06	0.95	0.24
Bayes Predictor	13135	12507	460	444	12967	0.49	0.00	0.49	0.51	0.50

Table A.32: Initial Experiment 32 - 5Categs24. Parameters: kNN(n_neighbours=15), DT(max_depth=100), RF(max_depth=100), Adaboost(max_depth=100). Further description can be found at the beginning of A.2.1

A.2.2 GridSearch

Modified tables from GridSearch show performance of classifiers with given parameters measured by 3 scoring functions SE, SP and Gm. The best scores for each scoring function are highlighted to make it easier to pick the best parameters. There is a table for each classifier combined with each dataset that have passed the initial experiments .

weightedSVC <i>2Categs15Centroids</i>									
C	Parameters			SE		SP		Gm	
	class_weight	gamma	kernel	mean	std	mean	std	mean	std
4	3.239	0.750	rbf	0.596	0.096	0.848	0.073	0.706	0.038
4	3.239	0.100	rbf	0.766	0.064	0.771	0.128	0.764	0.045
4	3.239	0.125	rbf	0.766	0.060	0.775	0.127	0.765	0.047
6	3.239	0.750	rbf	0.584	0.096	0.852	0.069	0.701	0.039
6	3.239	0.100	rbf	0.765	0.061	0.775	0.127	0.765	0.046
6	3.239	0.125	rbf	0.764	0.059	0.776	0.123	0.765	0.046
8	3.239	0.750	rbf	0.579	0.091	0.851	0.068	0.698	0.036
8	3.239	0.100	rbf	0.768	0.056	0.775	0.125	0.767	0.047
8	3.239	0.125	rbf	0.760	0.062	0.778	0.121	0.764	0.043
10	3.239	0.750	rbf	0.577	0.089	0.850	0.067	0.696	0.035
10	3.239	0.100	rbf	0.765	0.058	0.776	0.124	0.766	0.047
10	3.239	0.125	rbf	0.756	0.062	0.777	0.119	0.762	0.042

Table A.33: GridSearch: SVC *2Categs15Centroids*. Selected parameters are: C=8 and gamma=0.1. Further description can be found at the beginning of A.2.2

weightedSVC <i>2Categs15Less</i>									
C	Parameters			SE		SP		Gm	
	class_weight	gamma	kernel	mean	std	mean	std	mean	std
4	4.053	0.750	rbf	0.572	0.081	0.857	0.072	0.696	0.038
4	4.053	0.100	rbf	0.769	0.055	0.787	0.137	0.773	0.056
4	4.053	0.125	rbf	0.769	0.058	0.788	0.137	0.774	0.056
6	4.053	0.750	rbf	0.553	0.077	0.864	0.068	0.687	0.036
6	4.053	0.100	rbf	0.766	0.058	0.787	0.138	0.771	0.056
6	4.053	0.125	rbf	0.767	0.059	0.786	0.138	0.771	0.054
8	4.053	0.750	rbf	0.538	0.077	0.864	0.066	0.678	0.034
8	4.053	0.100	rbf	0.769	0.060	0.788	0.138	0.773	0.055
8	4.053	0.125	rbf	0.764	0.061	0.784	0.137	0.769	0.054
10	4.053	0.750	rbf	0.535	0.081	0.864	0.065	0.676	0.037
10	4.053	0.100	rbf	0.769	0.060	0.786	0.139	0.772	0.056
10	4.053	0.125	rbf	0.763	0.062	0.783	0.136	0.767	0.053

Table A.34: GridSearch: SVC *2Categs15Less*. Selected parameters are: C=4 and gamma=0.125. Further description can be found at the beginning of A.2.2

weightedSVC 2Categs15MFE

Parameters				SE		SP		Gm	
C	class_weight	gamma	kernel	mean	std	mean	std	mean	std
4	2.930	0.750	rbf	0.626	0.075	0.845	0.073	0.724	0.028
4	2.930	0.100	rbf	0.778	0.048	0.777	0.131	0.773	0.049
4	2.930	0.125	rbf	0.776	0.050	0.781	0.131	0.774	0.050
6	2.930	0.750	rbf	0.608	0.081	0.851	0.069	0.715	0.032
6	2.930	0.100	rbf	0.777	0.050	0.780	0.131	0.774	0.050
6	2.930	0.125	rbf	0.775	0.051	0.783	0.132	0.774	0.049
8	2.930	0.750	rbf	0.601	0.082	0.851	0.068	0.711	0.033
8	2.930	0.100	rbf	0.775	0.049	0.781	0.133	0.774	0.051
8	2.930	0.125	rbf	0.773	0.051	0.784	0.131	0.774	0.048
10	2.930	0.750	rbf	0.594	0.085	0.850	0.068	0.706	0.036
10	2.930	0.100	rbf	0.777	0.050	0.782	0.132	0.775	0.049
10	2.930	0.125	rbf	0.773	0.050	0.784	0.131	0.774	0.048

Table A.35: GridSearch: SVC 2Categs15MFE. Selected parameters are: C=10 and gamma=0.1. Further description can be found at the beginning of A.2.2

weightedSVC 2Categs15More

Parameters				SE		SP		Gm	
C	class_weight	gamma	kernel	mean	std	mean	std	mean	std
4	3.516	0.750	rbf	0.570	0.105	0.851	0.074	0.690	0.045
4	3.516	0.100	rbf	0.755	0.079	0.783	0.147	0.762	0.057
4	3.516	0.125	rbf	0.752	0.082	0.786	0.141	0.762	0.054
6	3.516	0.750	rbf	0.552	0.106	0.853	0.071	0.680	0.047
6	3.516	0.100	rbf	0.753	0.081	0.784	0.145	0.762	0.057
6	3.516	0.125	rbf	0.747	0.079	0.787	0.139	0.760	0.053
8	3.516	0.750	rbf	0.543	0.113	0.854	0.068	0.674	0.053
8	3.516	0.100	rbf	0.752	0.082	0.786	0.142	0.762	0.054
8	3.516	0.125	rbf	0.746	0.077	0.785	0.135	0.759	0.050
10	3.516	0.750	rbf	0.543	0.112	0.853	0.068	0.674	0.052
10	3.516	0.100	rbf	0.749	0.078	0.786	0.139	0.761	0.052
10	3.516	0.125	rbf	0.741	0.078	0.783	0.135	0.756	0.049

Table A.36: GridSearch: SVC 2Categs15More. Selected parameters are: C=10 and gamma=0.1. Further description can be found at the beginning of A.2.2

AdaBoost(DT) 2Categs15Centroids							
Parameters		SE		SP		Gm	
max_depth	n_estimators	mean	std	mean	std	mean	std
None	5	0.454	0.069	0.767	0.038	0.587	0.037
None	7	0.454	0.069	0.767	0.038	0.587	0.037
None	9	0.454	0.069	0.767	0.038	0.587	0.037
None	11	0.454	0.069	0.767	0.038	0.587	0.037
None	13	0.454	0.069	0.767	0.038	0.587	0.037
None	15	0.454	0.069	0.767	0.038	0.587	0.037
5	5	0.435	0.102	0.942	0.042	0.634	0.067
5	7	0.503	0.110	0.932	0.053	0.679	0.061
5	9	0.531	0.103	0.924	0.056	0.695	0.054
5	11	0.554	0.110	0.922	0.064	0.709	0.056
5	13	0.565	0.108	0.915	0.070	0.713	0.052
5	15	0.568	0.110	0.912	0.070	0.714	0.052
7	5	0.522	0.105	0.898	0.053	0.679	0.053
7	7	0.548	0.102	0.888	0.061	0.692	0.045
7	9	0.560	0.099	0.882	0.062	0.698	0.045
7	11	0.567	0.096	0.883	0.062	0.703	0.040
7	13	0.570	0.097	0.883	0.064	0.705	0.045
7	15	0.564	0.087	0.880	0.060	0.700	0.036
9	5	0.556	0.091	0.844	0.059	0.680	0.034
9	7	0.556	0.096	0.844	0.063	0.680	0.037
9	9	0.550	0.093	0.845	0.060	0.677	0.039
9	11	0.523	0.092	0.847	0.061	0.661	0.038
9	13	0.520	0.095	0.850	0.057	0.660	0.042
9	15	0.516	0.087	0.854	0.054	0.659	0.038
11	5	0.478	0.064	0.838	0.048	0.630	0.031
11	7	0.475	0.069	0.840	0.045	0.629	0.032
11	9	0.482	0.065	0.841	0.047	0.634	0.028
11	11	0.496	0.067	0.841	0.048	0.643	0.030
11	13	0.503	0.068	0.841	0.048	0.648	0.028
11	15	0.507	0.076	0.841	0.050	0.649	0.032

Table A.37: GridSearch: Adaboost(DT) 2Categs15Centroids. Selected parameters are: max_depth=5 and n_estimators=15. Further description can be found at the beginning of A.2.2

<i>AdaBoost(DT) 2Categs15Less</i>							
Parameters		SE		SP		Gm	
max_depth	n_estimators	mean	std	mean	std	mean	std
None	5	0.431	0.056	0.803	0.042	0.586	0.026
None	7	0.431	0.056	0.803	0.042	0.586	0.026
None	9	0.431	0.056	0.803	0.042	0.586	0.026
None	11	0.431	0.056	0.803	0.042	0.586	0.026
None	13	0.431	0.056	0.803	0.042	0.586	0.026
None	15	0.431	0.056	0.803	0.042	0.586	0.026
5	5	0.378	0.094	0.957	0.039	0.596	0.068
5	7	0.434	0.098	0.950	0.050	0.637	0.066
5	9	0.474	0.094	0.945	0.054	0.665	0.060
5	11	0.482	0.088	0.938	0.060	0.668	0.053
5	13	0.496	0.085	0.937	0.061	0.677	0.051
5	15	0.499	0.084	0.934	0.062	0.679	0.050
7	5	0.486	0.097	0.929	0.056	0.667	0.061
7	7	0.510	0.086	0.919	0.062	0.681	0.050
7	9	0.520	0.083	0.913	0.065	0.685	0.047
7	11	0.520	0.083	0.912	0.066	0.685	0.048
7	13	0.532	0.086	0.911	0.067	0.693	0.050
7	15	0.532	0.083	0.909	0.068	0.692	0.048
9	5	0.498	0.074	0.884	0.060	0.660	0.042
9	7	0.503	0.082	0.880	0.060	0.662	0.048
9	9	0.508	0.081	0.877	0.060	0.664	0.047
9	11	0.511	0.083	0.873	0.058	0.664	0.046
9	13	0.508	0.087	0.873	0.057	0.662	0.049
9	15	0.504	0.079	0.870	0.054	0.659	0.044
11	5	0.438	0.066	0.870	0.049	0.615	0.039
11	7	0.436	0.075	0.875	0.046	0.615	0.046
11	9	0.440	0.072	0.878	0.048	0.618	0.040
11	11	0.454	0.074	0.877	0.051	0.627	0.041
11	13	0.460	0.075	0.877	0.051	0.632	0.042
11	15	0.462	0.076	0.876	0.051	0.632	0.042

Table A.38: GridSearch: Adaboost(DT) *2Categs15Less*. Selected parameters are: max_depth=7 and n_estimators=13. Further description can be found at the beginning of A.2.2

AdaBoost(DT) 2Categs15MFE							
Parameters		SE		SP		Gm	
max_depth	n_estimators	mean	std	mean	std	mean	std
None	5	0.509	0.054	0.785	0.048	0.630	0.022
None	7	0.509	0.054	0.785	0.048	0.630	0.022
None	9	0.509	0.054	0.785	0.048	0.630	0.022
None	11	0.509	0.054	0.785	0.048	0.630	0.022
None	13	0.509	0.054	0.785	0.048	0.630	0.022
None	15	0.509	0.054	0.785	0.048	0.630	0.022
5	5	0.516	0.097	0.934	0.051	0.690	0.056
5	7	0.561	0.090	0.925	0.059	0.716	0.047
5	9	0.585	0.085	0.919	0.069	0.730	0.044
5	11	0.595	0.084	0.915	0.069	0.734	0.040
5	13	0.602	0.087	0.912	0.070	0.737	0.040
5	15	0.605	0.090	0.910	0.070	0.737	0.040
7	5	0.579	0.077	0.889	0.067	0.714	0.033
7	7	0.603	0.077	0.885	0.073	0.727	0.033
7	9	0.605	0.071	0.881	0.074	0.727	0.029
7	11	0.610	0.076	0.877	0.078	0.728	0.031
7	13	0.612	0.069	0.872	0.077	0.727	0.029
7	15	0.616	0.066	0.876	0.074	0.732	0.029
9	5	0.597	0.065	0.840	0.065	0.705	0.026
9	7	0.590	0.072	0.841	0.067	0.701	0.031
9	9	0.582	0.069	0.839	0.065	0.696	0.026
9	11	0.576	0.070	0.841	0.066	0.693	0.028
9	13	0.562	0.072	0.846	0.068	0.686	0.030
9	15	0.556	0.076	0.846	0.070	0.682	0.030
11	5	0.513	0.071	0.830	0.054	0.650	0.032
11	7	0.513	0.065	0.831	0.056	0.650	0.027
11	9	0.520	0.068	0.836	0.051	0.656	0.030
11	11	0.535	0.077	0.831	0.055	0.664	0.035
11	13	0.537	0.077	0.836	0.055	0.667	0.036
11	15	0.537	0.071	0.832	0.055	0.666	0.032

Table A.39: GridSearch: Adaboost(DT) 2Categs15MFE. Selected parameters are: max_depth=7 and n_estimators=15. Further description can be found at the beginning of A.2.2

AdaBoost(DT) <i>2Categs15More</i>							
Parameters		SE		SP		Gm	
max_depth	n_estimators	mean	std	mean	std	mean	std
None	5	0.468	0.073	0.798	0.045	0.608	0.031
None	7	0.468	0.073	0.798	0.045	0.608	0.031
None	9	0.468	0.073	0.798	0.045	0.608	0.031
None	11	0.468	0.073	0.798	0.045	0.608	0.031
None	13	0.468	0.073	0.798	0.045	0.608	0.031
None	15	0.468	0.073	0.798	0.045	0.608	0.031
5	5	0.409	0.116	0.950	0.040	0.616	0.080
5	7	0.463	0.118	0.940	0.053	0.653	0.075
5	9	0.502	0.108	0.931	0.058	0.678	0.064
5	11	0.510	0.112	0.928	0.059	0.682	0.066
5	13	0.520	0.116	0.925	0.063	0.687	0.070
5	15	0.524	0.114	0.920	0.067	0.688	0.065
7	5	0.489	0.110	0.911	0.057	0.661	0.061
7	7	0.516	0.102	0.907	0.062	0.679	0.054
7	9	0.531	0.099	0.901	0.067	0.687	0.051
7	11	0.544	0.097	0.897	0.067	0.694	0.050
7	13	0.550	0.106	0.893	0.069	0.695	0.054
7	15	0.546	0.106	0.891	0.067	0.692	0.054
9	5	0.509	0.101	0.859	0.060	0.656	0.047
9	7	0.513	0.101	0.857	0.059	0.658	0.047
9	9	0.507	0.102	0.861	0.055	0.655	0.050
9	11	0.500	0.097	0.860	0.055	0.650	0.047
9	13	0.502	0.104	0.858	0.057	0.651	0.053
9	15	0.503	0.095	0.857	0.054	0.652	0.047
11	5	0.444	0.094	0.859	0.042	0.613	0.053
11	7	0.444	0.102	0.859	0.043	0.612	0.057
11	9	0.449	0.100	0.853	0.049	0.613	0.054
11	11	0.464	0.094	0.859	0.048	0.626	0.051
11	13	0.467	0.094	0.857	0.048	0.628	0.048
11	15	0.467	0.097	0.857	0.052	0.628	0.050

Table A.40: GridSearch: Adaboost(DT) *2Categs15More*. Selected parameters are: max_depth=7 and n_estimators=13. Further description can be found at the beginning of A.2.2

<i>DT 2Categs15Centroids</i>						
Parameters	SE		SP		Gm	
max_depth	mean	std	mean	std	mean	std
13	0.397	0.057	0.798	0.032	0.563	0.026
15	0.461	0.053	0.766	0.040	0.593	0.028
17	0.459	0.060	0.769	0.035	0.596	0.029
19	0.463	0.063	0.765	0.040	0.589	0.032
21	0.464	0.062	0.764	0.041	0.591	0.035
23	0.461	0.058	0.763	0.041	0.593	0.036
25	0.457	0.062	0.769	0.040	0.591	0.028
27	0.461	0.060	0.768	0.038	0.595	0.032
29	0.455	0.068	0.764	0.038	0.591	0.032

Table A.41: GridSearch: Decision Tree *2Categs15Centroids*. Selected parameters are: max_depth=21. Further description can be found at the beginning of A.2.2

<i>DT 2Categs15Less</i>						
Parameters	SE		SP		Gm	
max_depth	mean	std	mean	std	mean	std
13	0.357	0.055	0.842	0.038	0.551	0.035
15	0.439	0.063	0.801	0.044	0.587	0.025
17	0.435	0.064	0.803	0.045	0.592	0.028
19	0.431	0.059	0.805	0.042	0.590	0.030
21	0.436	0.059	0.803	0.044	0.588	0.026
23	0.437	0.059	0.804	0.043	0.588	0.027
25	0.437	0.055	0.803	0.044	0.589	0.029
27	0.440	0.064	0.802	0.044	0.589	0.025
29	0.434	0.056	0.803	0.042	0.588	0.028

Table A.42: GridSearch: Decision Tree *2Categs15Less*. Selected parameters are: max_depth=27. Further description can be found at the beginning of A.2.2

DT *2Categs15MFE*

Parameters	SE		SP		Gm	
	mean	std	mean	std	mean	std
max_depth						
13	0.452	0.048	0.816	0.036	0.601	0.025
15	0.513	0.052	0.781	0.045	0.634	0.020
17	0.509	0.055	0.786	0.045	0.628	0.018
19	0.515	0.051	0.782	0.047	0.633	0.021
21	0.516	0.045	0.787	0.045	0.635	0.023
23	0.508	0.042	0.782	0.047	0.632	0.024
25	0.511	0.052	0.783	0.046	0.629	0.022
27	0.514	0.052	0.788	0.044	0.633	0.021
29	0.517	0.047	0.785	0.050	0.632	0.020

Table A.43: GridSearch: Decision Tree *2Categs15MFE*. Selected parameters are: max_depth=21. Further description can be found at the beginning of A.2.2

DT *2Categs15More*

Parameters	SE		SP		Gm	
	mean	std	mean	std	mean	std
max_depth						
13	0.412	0.073	0.833	0.036	0.583	0.043
15	0.469	0.079	0.800	0.045	0.611	0.030
17	0.475	0.076	0.798	0.045	0.612	0.034
19	0.471	0.086	0.797	0.046	0.610	0.034
21	0.472	0.074	0.797	0.046	0.611	0.031
23	0.470	0.075	0.797	0.044	0.610	0.031
25	0.469	0.073	0.798	0.045	0.610	0.030
27	0.472	0.075	0.797	0.046	0.609	0.026
29	0.470	0.074	0.800	0.045	0.612	0.030

Table A.44: GridSearch: Decision Tree *2Categs15More*. Selected parameters are: max_depth=17. Further description can be found at the beginning of A.2.2

k-NN *2Categs15Centroids*

Parameters		SE		SP		Gm	
n_neighbors	weights	mean	std	mean	std	mean	std
2	distance	0.414	0.052	0.820	0.038	0.581	0.025
3	distance	0.585	0.062	0.716	0.055	0.645	0.024
4	distance	0.404	0.054	0.882	0.036	0.595	0.027
5	distance	0.509	0.060	0.829	0.047	0.647	0.025

Table A.45: GridSearch: k-NN *2Categs15Centroids*. Selected parameters are: n_neighbors=3. Further description can be found at the beginning of A.2.2

Parameters		SE		SP		Gm	
n_neighbors	weights	mean	std	mean	std	mean	std
2	distance	0.377	0.050	0.860	0.039	0.568	0.027
3	distance	0.531	0.052	0.779	0.057	0.641	0.017
4	distance	0.390	0.058	0.897	0.042	0.589	0.035
5	distance	0.486	0.055	0.863	0.057	0.646	0.024

Table A.46: GridSearch: k-NN *2Categs15Less*. Selected parameters are: n_neighbors=3. Further description can be found at the beginning of A.2.2

Parameters		SE		SP		Gm	
n_neighbors	weights	mean	std	mean	std	mean	std
2	distance	0.382	0.038	0.837	0.037	0.564	0.022
3	distance	0.575	0.048	0.720	0.058	0.641	0.013
4	distance	0.412	0.055	0.880	0.039	0.600	0.030
5	distance	0.524	0.049	0.820	0.059	0.654	0.016

Table A.47: GridSearch: k-NN *2Categs15MFE*. Selected parameters are: n_neighbors=3. Further description can be found at the beginning of A.2.2

Parameters		SE		SP		Gm	
n_neighbors	weights	mean	std	mean	std	mean	std
2	distance	0.347	0.055	0.875	0.030	0.549	0.036
3	distance	0.517	0.058	0.778	0.047	0.632	0.024
4	distance	0.330	0.059	0.909	0.032	0.545	0.039
5	distance	0.439	0.071	0.862	0.048	0.612	0.035

Table A.48: GridSearch: k-NN *2Categs15More*. Selected parameters are: n_neighbors=3. Further description can be found at the beginning of A.2.2

RandomForest <i>2Categs15Centroids</i>							
Parameters		SE		SP		Gm	
max_depth	n_estimators	mean	std	mean	std	mean	std
10	1	0.246	0.037	0.874	0.024	0.461	0.029
10	3	0.183	0.053	0.947	0.016	0.410	0.062
10	5	0.147	0.049	0.972	0.009	0.373	0.062
10	7	0.128	0.040	0.981	0.009	0.350	0.056
12	1	0.341	0.045	0.809	0.034	0.524	0.028
12	3	0.322	0.058	0.879	0.029	0.529	0.040
12	5	0.307	0.066	0.909	0.027	0.524	0.048
12	7	0.292	0.067	0.929	0.023	0.516	0.056
14	1	0.413	0.045	0.776	0.033	0.565	0.019
14	3	0.406	0.051	0.842	0.032	0.583	0.027
14	5	0.397	0.062	0.875	0.036	0.586	0.034
14	7	0.385	0.065	0.896	0.036	0.584	0.039
16	1	0.431	0.045	0.774	0.029	0.576	0.020
16	3	0.413	0.068	0.836	0.038	0.584	0.037
16	5	0.396	0.065	0.873	0.036	0.585	0.038
16	7	0.394	0.069	0.891	0.033	0.589	0.042
18	1	0.431	0.045	0.774	0.029	0.576	0.020
18	3	0.413	0.068	0.836	0.038	0.584	0.037
18	5	0.396	0.065	0.873	0.036	0.585	0.038
18	7	0.394	0.069	0.891	0.033	0.589	0.042

Table A.49: GridSearch: Random Forest *2Categs15Centroids*. Selected parameters are: max_depth=16 and n_estimators=1. Further description can be found at the beginning of A.2.2

RandomForest <i>2Categs15Less</i>							
Parameters		SE		SP		Gm	
max_depth	n_estimators	mean	std	mean	std	mean	std
10	1	0.235	0.045	0.916	0.030	0.461	0.042
10	3	0.142	0.053	0.977	0.017	0.366	0.064
10	5	0.126	0.057	0.988	0.011	0.344	0.076
10	7	0.108	0.048	0.992	0.008	0.318	0.075
12	1	0.306	0.067	0.864	0.036	0.510	0.053
12	3	0.276	0.065	0.931	0.030	0.503	0.055
12	5	0.266	0.072	0.951	0.028	0.497	0.063
12	7	0.253	0.080	0.966	0.022	0.487	0.075
14	1	0.397	0.040	0.824	0.037	0.570	0.021
14	3	0.368	0.052	0.886	0.041	0.569	0.033
14	5	0.357	0.065	0.911	0.042	0.567	0.046
14	7	0.344	0.071	0.927	0.038	0.561	0.051
16	1	0.405	0.052	0.818	0.039	0.573	0.029
16	3	0.379	0.062	0.882	0.039	0.576	0.038
16	5	0.352	0.074	0.907	0.036	0.561	0.054
16	7	0.351	0.076	0.924	0.036	0.565	0.057
18	1	0.405	0.052	0.818	0.039	0.573	0.029
18	3	0.379	0.062	0.882	0.039	0.576	0.038
18	5	0.352	0.074	0.907	0.036	0.561	0.054
18	7	0.351	0.076	0.924	0.036	0.565	0.057

Table A.50: GridSearch: Random Forest *2Categs15Less*. Selected parameters are: max_depth=16 and n_estimators=1. Further description can be found at the beginning of A.2.2

RandomForest <i>2Categs15MFE</i>							
Parameters		SE		SP		Gm	
max_depth	n_estimators	mean	std	mean	std	mean	std
10	1	0.292	0.024	0.861	0.024	0.501	0.021
10	3	0.221	0.039	0.951	0.015	0.456	0.037
10	5	0.185	0.050	0.972	0.014	0.420	0.054
10	7	0.173	0.051	0.983	0.009	0.408	0.059
12	1	0.377	0.034	0.799	0.027	0.548	0.019
12	3	0.370	0.046	0.874	0.030	0.567	0.028
12	5	0.358	0.065	0.907	0.027	0.567	0.045
12	7	0.358	0.071	0.923	0.025	0.571	0.050
14	1	0.444	0.032	0.758	0.046	0.579	0.011
14	3	0.441	0.069	0.835	0.046	0.604	0.035
14	5	0.434	0.067	0.864	0.044	0.609	0.037
14	7	0.423	0.062	0.884	0.043	0.609	0.033
16	1	0.456	0.029	0.764	0.038	0.590	0.013
16	3	0.441	0.045	0.828	0.033	0.603	0.022
16	5	0.428	0.057	0.867	0.034	0.607	0.033
16	7	0.417	0.064	0.890	0.034	0.607	0.039
18	1	0.456	0.029	0.764	0.038	0.590	0.013
18	3	0.441	0.045	0.828	0.033	0.603	0.022
18	5	0.428	0.057	0.867	0.034	0.607	0.033
18	7	0.417	0.064	0.890	0.034	0.607	0.039

Table A.51: GridSearch: Random Forest *2Categs15MFE*. Selected parameters are: max_depth=14 and n_estimators=5. Further description can be found at the beginning of A.2.2

RandomForest <i>2Categs15More</i>							
Parameters		SE		SP		Gm	
max_depth	n_estimators	mean	std	mean	std	mean	std
10	1	0.264	0.050	0.897	0.020	0.484	0.042
10	3	0.204	0.076	0.966	0.015	0.436	0.080
10	5	0.183	0.070	0.980	0.012	0.415	0.079
10	7	0.166	0.075	0.988	0.009	0.394	0.089
12	1	0.345	0.060	0.830	0.030	0.533	0.037
12	3	0.315	0.074	0.904	0.029	0.529	0.054
12	5	0.302	0.094	0.930	0.028	0.522	0.074
12	7	0.282	0.094	0.947	0.026	0.509	0.079
14	1	0.437	0.064	0.793	0.035	0.586	0.033
14	3	0.403	0.073	0.864	0.038	0.587	0.042
14	5	0.382	0.081	0.893	0.040	0.580	0.051
14	7	0.381	0.083	0.909	0.040	0.584	0.053
16	1	0.416	0.068	0.795	0.032	0.572	0.037
16	3	0.386	0.084	0.862	0.040	0.572	0.049
16	5	0.371	0.089	0.888	0.038	0.568	0.055
16	7	0.368	0.091	0.909	0.031	0.573	0.060
18	1	0.416	0.068	0.795	0.032	0.572	0.037
18	3	0.386	0.084	0.862	0.040	0.572	0.049
18	5	0.371	0.089	0.888	0.038	0.568	0.055
18	7	0.368	0.091	0.909	0.031	0.573	0.060

Table A.52: GridSearch: Random Forest *2Categs15More*. Selected parameters are: max_depth=14 and n_estimators=1. Further description can be found at the beginning of A.2.2

A.3 Results for MLPClassifiers

Experiment: 1

Classifier	Confusion matrix				Precision		Recall		F1-Support	
	TN	FP	FN	TP	1	0	1	0	1	0
<i>MLP_1_1</i>	3015894	3265	5	3654	0.53	1.00	1.00	1.00	0.69	1.00
<i>MLP_1_12</i>	3015916	3243	24	3635	0.53	1.00	0.99	1.00	0.69	1.00
<i>MLP_3_12</i>	3015897	3262	10	3649	0.53	1.00	1.00	1.00	0.69	1.00
<i>MLP_3_D</i>	3015921	3238	17	3642	0.53	1.00	1.00	1.00	0.69	1.00

Table A.53: Results of MLPs for *Experiment 1*. Further description can be found at the beginning of Section A.3

Experiment: 2

Classifier	Confusion matrix				Precision		Recall		F1-Support	
	TN	FP	FN	TP	1	0	1	0	1	0
<i>MLP_1_1</i>	3019159	0	3659	0	0.00	1.00	0.00	1.00	0.00	1.00
<i>MLP_1_12</i>	3015893	3266	3	3656	0.53	1.00	1.00	1.00	0.69	1.00
<i>MLP_3_12</i>	3015887	3272	0	3659	0.53	1.00	1.00	1.00	0.69	1.00
<i>MLP_3_D</i>	3015902	3257	9	3650	0.53	1.00	1.00	1.00	0.69	1.00

Table A.54: Results of MLPs for *Experiment 2*. Further description can be found at the beginning of Section A.3

Experiment: 3

Classifier	Confusion matrix				Precision		Recall		F1-Support	
	TN	FP	FN	TP	1	0	1	0	1	0
<i>MLP_1_1</i>	3019159	0	3659	0	0.00	1.00	0.00	1.00	0.00	1.00
<i>MLP_1_12</i>	3015897	3262	9	3650	0.53	1.00	1.00	1.00	0.69	1.00
<i>MLP_3_12</i>	3015887	3272	0	3659	0.53	1.00	1.00	1.00	0.69	1.00
<i>MLP_3_D</i>	3015896	3263	9	3650	0.53	1.00	1.00	1.00	0.69	1.00

Table A.55: Results of MLPs for *Experiment 3*. Further description can be found at the beginning of Section A.3

Experiment: 4

Classifier	Confusion matrix				Precision		Recall		F1-Support	
	TN	FP	FN	TP	1	0	1	0	1	0
<i>MLP_1_1</i>	3019159	0	3659	0	0.00	1.00	0.00	1.00	0.00	1.00
<i>MLP_1_12</i>	3015914	3245	17	3642	0.53	1.00	1.00	1.00	0.69	1.00
<i>MLP_3_12</i>	3015921	3238	23	3636	0.53	1.00	0.99	1.00	0.69	1.00
<i>MLP_3_D</i>	3015935	3224	29	3630	0.53	1.00	0.99	1.00	0.69	1.00

Table A.56: Results of MLPs for *Experiment 4*. Further description can be found at the beginning of Section A.3

Appendix B

B.1 CD contents

circrna-squence-homology	Source codes of the implementation
├─ README	Project contents described in detail
├─ graph_tables	Directory containing generated graphs and tables
├─ example.sh	Example script for running the implementation
└─ Sequential_homology_of_circular_RNA.pdf	Master Thesis in pdf format

B.2 Used Graphical Programs

- Draw_io [69]
- Inkscape [70]
- Notability [71]
- Matplotlib_venn [72]
- Matplotlib.pyplot [72]

Bibliography

- [1] Rnafold 2.4.18 documentation. [Online]. Available: <https://www.tbi.univie.ac.at/RNA/RNALfold.1.html>
- [2] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, “MicroRNA targets in drosophila,” *Genome Biology*, vol. 5, no. 1, p. R1, 2003.
- [3] Mlp classifier scikit-learn implementation manual. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
- [4] A. Mitra, K. Pfeifer, and K.-S. Park, “Circular rnas and competing endogenous rna (cerna) networks,” *Translational Cancer Research*, vol. 7, no. 5, 2018.
- [5] M. Maiti, K. Nauwelaerts, E. Lescrinier, F. C. Schuit, and P. Herdewijn, “Self-complementary sequence context in mature miRNAs,” *Biochemical and Biophysical Research Communications*, vol. 392, no. 4, pp. 572 – 576, 2010.
- [6] A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel, “MicroRNA targeting specificity in mammals: Determinants beyond seed pairing,” *Molecular Cell*, vol. 27, no. 1, pp. 91–105, jul 2007.
- [7] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel, “Predicting effective microRNA target sites in mammalian mRNAs,” *eLife*, vol. 4, aug 2015.
- [8] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, “mirbase: microRNA sequences, targets and gene nomenclature,” *Nucleic Acids Research*, vol. 34, pp. D140–D144, 01 2006.
- [9] J. Starega-Roslan and W. J. Krzyzosiak, “Analysis of MicroRNA length variety generated by recombinant human dicer,” in *MicroRNA Protocols*. Humana Press, aug 2012, pp. 21–34.
- [10] A. L. Leichter, M. J. Sullivan, M. R. Eccles, and A. Chatterjee, “MicroRNA expression patterns and signalling pathways in the development and progression of

- childhood solid tumours,” *Molecular Cancer*, vol. 16, no. 1, 2017. [Online]. Available: <http://molecular-cancer.biomedcentral.com/articles/10.1186/s12943-017-0584-0>
- [11] Z. Huang, J. Shi, Y. Gao, C. Cui, S. Zhang, J. Li, Y. Zhou, and Q. Cui, “HMDD v3.0: a database for experimentally supported human microRNA disease associations,” *Nucleic Acids Research*, vol. 47, no. D1, pp. D1013–D1017, 10 2018.
- [12] T. Jakobi and C. Dieterich, “Computational approaches for circular rna analysis,” *WIREs RNA*, vol. 10, no. 3, p. e1528, 2019.
- [13] M.-S. Xiao and J. E. Wilusz, “An improved method for circular rna purification using rnase r that efficiently removes linear rnas containing g-quadruplexes or structured 3 ends,” *Nucleic Acids Research*, vol. 47, no. 16, pp. 8755–8769, Sep. 2019. [Online]. Available: <https://academic.oup.com/nar/article/47/16/8755/5527976>
- [14] J. R. Brown and A. M. Chinnaiyan, “The potential of circular rnas as cancer biomarkers,” *American Association for Cancer Research (AACR)*, vol. 29, no. 12, pp. 2541–2555, Dec. 2020. [Online]. Available: <http://cebp.aacrjournals.org/lookup/doi/10.1158/1055-9965.EPI-20-0796>
- [15] N. R. Pamudurti, O. Bartok, M. Jens, R. Ashwal-Fluss, C. Stottmeister, L. Ruhe, M. Hanan, E. Wyler, D. Perez-Hernandez, E. Ramberger, S. Shenzenis, M. Samson, G. Dittmar, M. Landthaler, M. Chekulaeva, N. Rajewsky, and S. Kadener, “Translation of CircRNAs,” *Molecular Cell*, vol. 66, no. 1, pp. 9–21.e7, 2017.
- [16] D. Siede, K. Rapti, A. Gorska, H. Katus, J. Altmüller, J. Boeckel, B. Meder, C. Maack, M. Völkers, O. Müller, J. Backs, and C. Dieterich, “Identification of circular RNAs with host gene-independent expression in human model systems for cardiac differentiation and disease,” *Journal of Molecular and Cellular Cardiology*, vol. 109, pp. 48–56, 2017.
- [17] L. Statello, C.-J. Guo, L.-L. Chen, and M. Huarte, “Gene regulation by long non-coding rnas and its biological functions,” *Nature Reviews Molecular Cell Biology*, vol. 22, no. 2, pp. 96–118, 2021. [Online]. Available: <http://www.nature.com/articles/s41580-020-00315-9>
- [18] M. S. Ebert and P. A. Sharp, “MicroRNA sponges: Progress and possibilities,” *RNA*, vol. 16, no. 11, pp. 2043–2050, sep 2010.
- [19] N. Chen, G. Zhao, X. Yan, Z. Lv, H. Yin, S. Zhang, W. Song, X. Li, L. Li, Z. Du, L. Jia, L. Zhou, W. Li, A. R. Hoffman, J.-F. Hu, and J. Cui, “A novel FLL1 exonic

- circular RNA promotes metastasis in breast cancer by coordinately regulating TET1 and DNMT1,” *Genome Biology*, vol. 19, no. 1, dec 2018.
- [20] M. Zhang, N. Huang, X. Yang, J. Luo, S. Yan, F. Xiao, W. Chen, X. Gao, K. Zhao, H. Zhou, Z. Li, L. Ming, B. Xie, and N. Zhang, “A novel protein encoded by the circular form of the SHPRH gene suppresses glioma tumorigenesis,” *Oncogene*, vol. 37, no. 13, pp. 1805–1814, 2018.
- [21] M. Su, Y. Xiao, J. Ma, Y. Tang, B. Tian, Y. Zhang, X. Li, Z. Wu, D. Yang, Y. Zhou, H. Wang, Q. Liao, and W. Wang, “Circular RNAs in cancer: emerging functions in hallmarks, stemness, resistance and roles as potential biomarkers,” *Molecular Cancer*, vol. 18, no. 1, 2019.
- [22] X. Zhao, Y. Cai, and J. Xu, “Circular RNAs: Biogenesis, mechanism, and function in human cancers,” *International Journal of Molecular Sciences*, vol. 20, no. 16, p. 3926, aug 2019.
- [23] H. Robins, Y. Li, and R. W. Padgett, “Incorporating structure to predict microRNA targets,” *PNAS; Proceedings of the National Academy of Sciences*, vol. 102, no. 11, pp. 4006–4009, 2005.
- [24] A. Belter, D. Gudanis, K. Rolle, M. Piwecka, Z. Gdaniec, M. Z. Naskret-Barciszewska, and J. Barciszewski, “Mature miRNAs form secondary structure, which suggests their function beyond RISC,” *PLoS ONE*, vol. 9, no. 11, p. e113848, nov 2014.
- [25] J. Sheu-Gruttadauria and I. J. MacRae, “Structural foundations of RNA silencing by argonaute,” *Journal of Molecular Biology*, vol. 429, no. 17, pp. 2619–2639, 2017.
- [26] F. Moretti, R. Thermann, and M. W. Hentze, “Mechanism of translational regulation by miR-2 from sites in the 5' untranslated region or the open reading frame,” *RNA*, vol. 16, no. 12, pp. 2493–2502, oct 2010.
- [27] E. Lopez-Jimenez, A. M. Rojas, and E. Andres-Leon, “RNA sequencing and prediction tools for circular RNAs analysis,” in *Advances in Experimental Medicine and Biology*. Springer Singapore, 2018, vol. 1087, pp. 17–33.
- [28] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, “miRBase: from microRNA sequences to function,” *Nucleic Acids Research*, vol. 47, no. D1, pp. D155–D162, nov 2018.

- [29] X. Chen, P. Han, T. Zhou, X. Guo, X. Song, and Y. Li, “circRNADb: A comprehensive database for human circular RNAs with protein-coding annotations,” *Scientific Reports*, vol. 6, no. 1, oct 2016.
- [30] W. R. Jeck, J. A. Sorrentino, K. Wang, M. K. Slevin, C. E. Burd, J. Liu, W. F. Marzluff, and N. E. Sharpless, “Circular RNAs are abundant, conserved, and associated with ALU repeats,” *RNA*, vol. 19, no. 2, pp. 141–157, dec 2012.
- [31] S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler, C. Kocks, F. le Noble, and N. Rajewsky, “Circular RNAs are a large class of animal RNAs with regulatory potency,” *Nature*, vol. 495, no. 7441, pp. 333–338, feb 2013.
- [32] J. Salzman, C. Gawad, P. L. Wang, N. Lacayo, and P. O. Brown, “Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types,” *PLoS ONE*, vol. 7, no. 2, p. e30733, feb 2012.
- [33] X.-O. Zhang, H.-B. Wang, Y. Zhang, X. Lu, L.-L. Chen, and L. Yang, “Complementary sequence-mediated exon circularization,” *Cell*, vol. 159, no. 1, pp. 134–147, sep 2014.
- [34] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, and J.-H. Yang, “starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-seq data,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D92–D97, dec 2013. [Online]. Available: <https://doi.org/10.1093/nar/nfn1248>
- [35] K. Zhou, S. Liu, L. Cai, and B. L. Encori:the encyclopedia of rna interactomes. [Online]. Available: <http://starbase.sysu.edu.cn/index.php>
- [36] B. P. Lewis, C. B. Burge, and D. P. Bartel, “Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets,” *Cell*, vol. 120, no. 1, pp. 15 – 20, 2005.
- [37] G. Anders, S. D. Mackowiak, M. Jens, J. Maaskola, A. Kuntzagk, N. Rajewsky, M. Landthaler, and C. Dieterich, “doRiNA: a database of RNA interactions in post-transcriptional regulation,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D180–D186, nov 2011.
- [38] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, “The role of site accessibility in microRNA target recognition,” *Nature Genetics*, vol. 39, no. 10, pp. 1278–1284, sep 2007. [Online]. Available: s

- [39] P. Loher and I. Rigoutsos, “Interactive exploration of RNA22 microRNA target predictions,” *Bioinformatics*, vol. 28, no. 24, pp. 3322–3323, oct 2012. [Online]. Available: <https://doi.org/10.1093/bioinformatics/bts615>
- [40] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, and J.-H. Yang, “starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D92–D97, 11 2013.
- [41] D. B. Dudekula, A. C. Panda, I. Grammatikakis, S. De, K. Abdelmohsen, and M. Gorospe, “CircInteractome: A web tool for exploring circular RNAs and their interacting proteins and microRNAs,” *RNA Biology*, vol. 13, no. 1, pp. 34–42, dec 2015.
- [42] S. Xia, J. Feng, K. Chen, Y. Ma, J. Gong, F. Cai, Y. Jin, Y. Gao, L. Xia, H. Chang, L. Wei, L. Han, and C. He, “CSCD: a database for cancer-specific circular RNAs,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D925–D929, 2017.
- [43] R. Dong, X.-K. Ma, G.-W. Li, and L. Yang, “CIRCpedia v2: An updated database for comprehensive circular RNA annotation and expression comparison,” *Genomics, Proteomics & Bioinformatics*, vol. 16, no. 4, pp. 226–233, aug 2018.
- [44] X. Fan and L. Kurgan, “Comprehensive overview and assessment of computational prediction of microRNA targets in animals,” *Briefings in Bioinformatics*, vol. 16, no. 5, pp. 780–794, dec 2014.
- [45] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel, “Most mammalian mRNAs are conserved targets of microRNAs,” *Genome Research*, vol. 19, no. 1, pp. 92–105, 2009.
- [46] D. M. Garcia, D. Baek, C. Shin, G. W. Bell, A. Grimson, and D. P. Bartel, “Weak seed-pairing stability and high target-site abundance decrease the proficiency of *lscy-6* and other microRNAs,” *Nature Structural & Molecular Biology*, vol. 18, no. 10, pp. 1139–1146, sep 2011.
- [47] T. Smith and M. Waterman, “Identification of common molecular subsequences,” *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.
- [48] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster, “Complete suboptimal folding of RNA and the stability of secondary structures,” *Biopolymers*, vol. 49, no. 2, pp. 145–165, 1999.

- [49] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, “Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure1edited by i. tinoco,” *Journal of Molecular Biology*, vol. 288, no. 5, pp. 911 – 940, 1999.
- [50] M. Zuker and P. Stiegler, “Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information,” *Nucleic Acids Research*, vol. 9, no. 1, pp. 133–148, 1981.
- [51] J. Kruger and M. Rehmsmeier, “RNAhybrid: microRNA target prediction easy, fast and flexible,” *Nucleic Acids Research*, vol. 34, no. Web Server, pp. W451–W454, jul 2006.
- [52] M. Rehmsmeier, “Fast and effective prediction of microRNA-target duplexes,” *RNA*, vol. 10, no. 10, pp. 1507–1517, oct 2004.
- [53] M. Dori and S. Bicciato, “Integration of bioinformatic predictions and experimental data to identify circrna-mirna associations,” *Genes*, vol. 10, no. 9, 2019.
- [54] R. Lorenz, S. H. Bernhart, C. H. zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, “ViennaRNA package 2.0,” *Algorithms for Molecular Biology*, vol. 6, no. 1, nov 2011.
- [55] Y. Ding, “RNA secondary structure prediction by centroids in a boltzmann weighted ensemble,” *RNA*, vol. 11, no. 8, pp. 1157–1166, aug 2005.
- [56] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski *et al.*, “Biopython: freely available python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.
- [57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: <https://scikit-learn.org/stable/>
- [58] L. Breiman, J. Friedman, S. Charles J, and R. A. Olshen, *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.
- [59] T. H. nad Robert Tibshirani and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

- [60] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: <https://doi.org/10.1023%2Fa%3A1010933404324>
- [61] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S002200009791504X>
- [62] J. Zhu, H. Zou, T. Hastie, and S. Rosset, “Multi-class adaboost,” *Statistics and Its Interface*, vol. 2, no. 3, pp. 349–360, 2009.
- [63] C.-C. Chang and C.-J. Lin, “LIBSVM,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, apr 2011. [Online]. Available: <https://doi.org/10.1145%2F1961189.1961199>
- [64] J. C. Platt and J. C. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advance in Large Margin Classifiers*, pp. 61–74, 1999. [Online]. Available: <http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639>
- [65] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger, “Tackling the poor assumptions of naive bayes text classifiers,” in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 616–623.
- [66] G. E. Hinton and G. E. Hinton, “Connectionist learning procedures,” *Artificial Intelligence*, 1989. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.216.5594>
- [67] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterton, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256. [Online]. Available: <http://proceedings.mlr.press/v9/glorot10a.html>
- [68] R. Batuwita and V. Palade, “Adjusted geometric-mean: A novel performance measure for imbalanced bioinformatics datasets learning,” *Journal of Bioinformatics and Computational Biology*, vol. 10, no. 04, p. 1250003, jul 2012.
- [69] D. B. Gaudenz Alder, “draw.io,” c2005-2021, [cit. 2021-05-21]. [Online]. Available: <https://www.draw.io/>

- [70] Inkscape Project, “Inkscape,” [cit. 2021-05-21]. [Online]. Available: <https://inkscape.org>
- [71] Ginger Labs, “Notability,” c2020, [cit. 2021-05-21]. [Online]. Available: <https://www.gingerlabs.com/>
- [72] J. D. Hunter, “Matplotlib: A 2d graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. [Online]. Available: <https://ieeexplore.ieee.org/document/4160265>