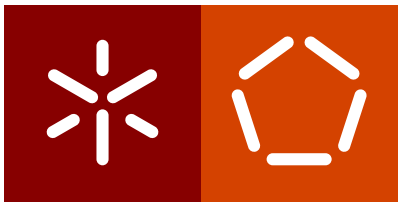
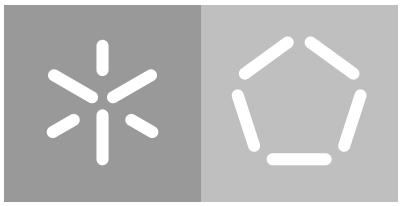


**Universidade do Minho**
Escola de Engenharia
Departamento de Informática

André Miguel Portugal Abrantes Cruzeiro Santiago

# A text mining based approach for biomarker discovery

December 2018



**Universidade do Minho**
Escola de Engenharia
Departamento de Informática

André Miguel Portugal Abrantes Cruzeiro Santiago

# A text mining based approach for biomarker discovery

Master Dissertation
Master Degree in Bioinformatics

Dissertation supervised by
**Miguel P. Rocha**
**Joel P. Arrais**

December 2018

# ACKNOWLEDGEMENTS

For all future readers, only this paragraph in this small section will be in English. This is because I can better present my thanks in my native language. Adding to this, if you are one of the recipients, continue on reading as I have some special words directed towards you.

Em primeiro lugar, tenho que agradecer a ambos os meus pais. Eles apoiaram-me incondicionalmente nesta caminhada, nunca deixando de acreditar em mim. Eles são das principais, senão mesmo a principal razão pela qual eu consegui terminar este projeto. Muito, mesmo muito obrigado a ambos.

As minhas irmãs também tiveram um papel importante nisto, nem que seja porque nunca pararam de me chatear para terminar a tese! Elas estiveram lá quando os meus pais não puderam, e estavam sempre mais que dispostas a dar-me um empurrãozinho quando achavam que eu estava a precisar.

Não posso, obviamente, deixar de agradecer a ambos os meus orientadores, as pessoas que mais me acompanharam neste projeto. Miguel e Joel, obrigado por terem sempre acreditado nas minhas capacidades e por estarem sempre lá disponíveis quando eu precisava que me apontassem na direção certa. Espero conseguir atingir e suplantar todas as vossas espetativas em qualquer aventura que venhamos ter juntos no futuro!

Quando a família não conseguia ajudar, sei que podia sempre contar com os meus melhores amigos. Jorge, mantiveste-me na linha nos meus dias mais difíceis e a tua companhia era o suficiente para me dar todas as forças que eu precisava. Marco, estiveste lá quando eu precisava de alguém com quem conversar e distraíste-me quando eu precisava de relaxar um bocado. Miguel, a tua confiança contagiante ajudou-me sempre a nunca deixar de acreditar em mim mesmo. Obrigado amigos, não teria chegado tão longe se vocês não estivessem lá.

E, finalmente, tenho de agradecer à Katia. Foi graças a ti que eu tive sequer forças suficientes para começar esta aventura. Foi bastante mais longa do que ambos alguma vez podíamos ter achado que iria ser, mas tinha que terminar algum dia, n'é? De qualquer das formas, obrigado por teres sempre acreditado em mim, mesmo quando eu próprio não acreditava. Se cheguei até onde cheguei e sou quem sou, é graças a ti.

Estou eternamente grato a todas as pessoas a quem me dirigi acima, e muitas outras mais. Podem acreditar que terão sempre um lugar especial no meu coração.

## ABSTRACT

Biomarkers have long been heralded as potential motivators for the emergence of new treatment and diagnostic procedures for disease conditions. However, for many years, the biomarker discovery process could only be achieved through experimental means, serving as a deterrent for their increase in popularity as the usually large number of candidates resulted in a costly and time-consuming discovery process. The increase in computational capabilities has led to a change in the paradigm of biomarker discovery, migrating from the clinical laboratory to *in silico* environments.

Furthermore, text mining, the act of automatically extracting information from text through computational means, has seen a rise in popularity in the biomedical fields. The number of studies and clinical trials in these fields has greatly increased in the past years, making the task of manually examining and annotating these, at the very least, incredibly cumbersome. Adding to this, even though the development of efficient and thorough natural language processing is still an on-going process, the potential for the discovery of common reported and hidden behaviours in the scientific literature is too high to be ignored. Several tools, technologies, pipelines and frameworks already exist capable of, at least, giving a glimpse on how the analysis of the available pile of scientific literature can pave the way for the development of novel medical techniques that might help in the prevention, diagnostic and treatment of diseases.

As such, a novel approach is presented in this work for achieving biomarker discovery, one that integrates both gene-disease associations extracted from current biomedical literature and RNA-Seq gene expression data in an $L_1$-regularization mixed-integer linear programming model for identifying potential biomarkers, potentially providing an optimal and robust genetic signature for disease diagnostic and helping identify novel biomarker candidates. This analysis was carried out on five publicly available RNA-Seq datasets obtained from the Genomic Data Commons Data Portal, related to breast, colon, lung and prostate cancer, and head and neck squamous cell carcinoma. Hyperparameter optimization was also performed for this approach, and the performance of the optimal set of parameters was compared against other machine learning methods.

# RESUMO

Os biomarcadores há muito que são considerados como os motivadores principais para o desenvolvimento de novos procedimentos de diagnóstico e tratamento de doenças. No entanto, até há relativamente pouco tempo, o processo de descoberta de biomarcadores estava dependente de métodos experimentais, sendo este um elemento dissuasor da sua aplicação e estudo em massa dado que o número elevado de candidatos implicava um processo de averiguação extremamente dispendioso e demorado. O grande aumento do poder computacional nas últimas décadas veio contrariar esta tendência, levando à migração do processo de descoberta de biomarcadores do laboratório para o ambiente *in silico*.

Para além disso, a aplicação de processos de mineração de textos, que consistem na extração de informação de documentos através de meios computacionais, tem visto um aumento da sua popularidade na comunidade biomédica devido ao aumento exponencial do número de estudos e ensaios clínicos nesta área, tornando todo o processo de análise e anotação manual destes bastante laborioso. A adicionar a isto, apesar do desenvolvimento de métodos eficientes capazes de processar linguagem natural na sua plenitude seja um processo que ainda esteja a decorrer, o potencial para a descoberta de comportamentos reportados e escondidos na literatura é demasiado elevado para ser ignorado. Já existem diversas ferramentas e tecnologias capazes de, pelo menos, dar uma indicação de como a análise da literatura científica disponível pode abrir o caminho para o desenvolvimento de novas técnicas e procedimentos médicos que poderão auxiliar na prevenção, diagnóstico e tratamento de doenças.

Como tal, é apresentado neste trabalho um novo método para realizar a descoberta de biomarcadores, que considera simultaneamente associações entre genes e doenças, já extraídas da literatura biomédica e dados de expressão de genes RNA-Seq num modelo de otimização linear com regularização $L_1$ com variáveis contínuas e inteiras (MILP) para identificar possíveis biomarcadores, sendo capaz potencialmente de providenciar assinaturas genéticas ótimas e robustas para o diagnóstico de doenças e ajudar a identificar novos candidatos a biomarcador. Esta análise foi levada a cabo em cinco conjuntos de dados RNA-Seq obtidos através do Portal de Dados do Genomic Data Commons (GDC) relacionados com os cancros da mama, cólon, pulmão, próstata, e carcinoma escamoso da cabeça e pescoço. Realizou-se também uma otimização dos hiperparâmetros deste método, e o desempenho do conjunto ideal de parâmetros foi comparado com o de outros métodos de aprendizagem máquina.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 1

# INTRODUCTION

This work relates to the development and validation of a pipeline for performing biomarker discovery, one that uses a mixed-integer linear programming model for building biomarker profiles capable of achieving accurate disease diagnostic.

For this purpose and to describe how this was achieved, this document is divided into five chapters. This chapter, Chapter 1, introduces the research problem, the motivation behind it and the reasoning behind this work, as well as the aims for the project as a whole. Chapter 2 presents the state-of-the-art for biomarker discovery, focusing on available techniques and their current challenges. Furthermore, chapter 3 explains the methodology developed and used in this work, while chapter 4 presents the obtained results and how they can be interpreted. Finally, in chapter 5 the conclusions that have been reached are presented and further analysed.

## 1.1 MOTIVATION

Large omics data, such as genomics, transcriptomics, proteomics, metagenomics, and metabolomics have radically altered how the search for knowledge is performed in biology and related sciences. Not only that, but the exponential increase in computational power, as well as new technological advancements in the last few years have facilitated the development of new and more powerful tools than what would have been possible even a decade ago. One of the fields that has seen a rise in focus by the biomedical community is biomarker discovery.

Biomarkers, also known as biological markers, which can include everything from the pulse rate and blood pressure of a patient to the expression level of genes or proteins, and present significant alterations between control and disease states [114]. The use of biomarkers in medicine can potentially provide tremendous aid in predictive, preventive and personalized medicine, allowing for the development of safer and more effective drugs for disease treatment, as well as better and faster diagnostic procedures.

Adding to this, the scientific community has understood their potential and everyday new methodologies emerge that provide means for the identification of potential biomark-

ers. Biomedical text mining is no foreigner to the discovery process of biomarkers and may provide crucial assistance in doing so as it enables the discovery of potential hidden relationships reported in the scientific literature which may lead to new, effective disease biomarkers.

As such, while a great variety of methods and technologies exist capable of identifying potential biomarkers and biomarker profiles for many diseases, these either provide low performance measures, or the reasoning behind the chosen genetic signature themselves is not easily understood, due in part to the rise in popularity of black-box methods. These have shown great promise in providing models capable of differentiating binary sample status with great precision, but this has come at the cost of easily understanding the reasons behind the effective status separation, becoming a barrier for their application in the biomedical sciences.

This provides us with the motivation to develop a method that is both capable of identifying performant biomarker profiles and provide signatures whose mode of operation can be easily understood by biomedical personnel. A mixed-integer linear model with a ternary status output system has the potential to allow both, due to its simpler linear structure and solutions, even with a highly dimensional feature space. Specifically, the performance of every selected potential biomarker will be evaluated linearly, with each profile constituting nothing more than the sum of the individual constituents discriminating capabilities. These models are simpler in nature as the role of each potential biomarker in the genetic signature is easily understood. Adding to this, the ternary classification system provides an inherent solution flexibility that evades the classic binary system of positive and negative, allowing the models to state any sample as ambiguous and requiring further manual analysis from biomedical personnel.

## 1.2 OBJECTIVES

The main aim of this project involved the development of an $L_1$-regularization mixed-integer linear programming (MILP) model, based on the work of Santiago *et al.* [100], that incorporates gene-disease associations obtained from the National Institute of Health Indexing Initiative extracted from the literature and RNA-Seq gene expression data. As such, in the approach presented in this work, the quantification of a given gene's importance for a specific disease is achieved by estimating a predictive model when taking into account the number of citations found for that particular association. This will naturally drive genes with a larger number of citations to be selected by the feature selection process and as such, ending up in the final selection of biomarkers. This, however, does not exclude less important genes as these can still be added to the final signature if they add predictive

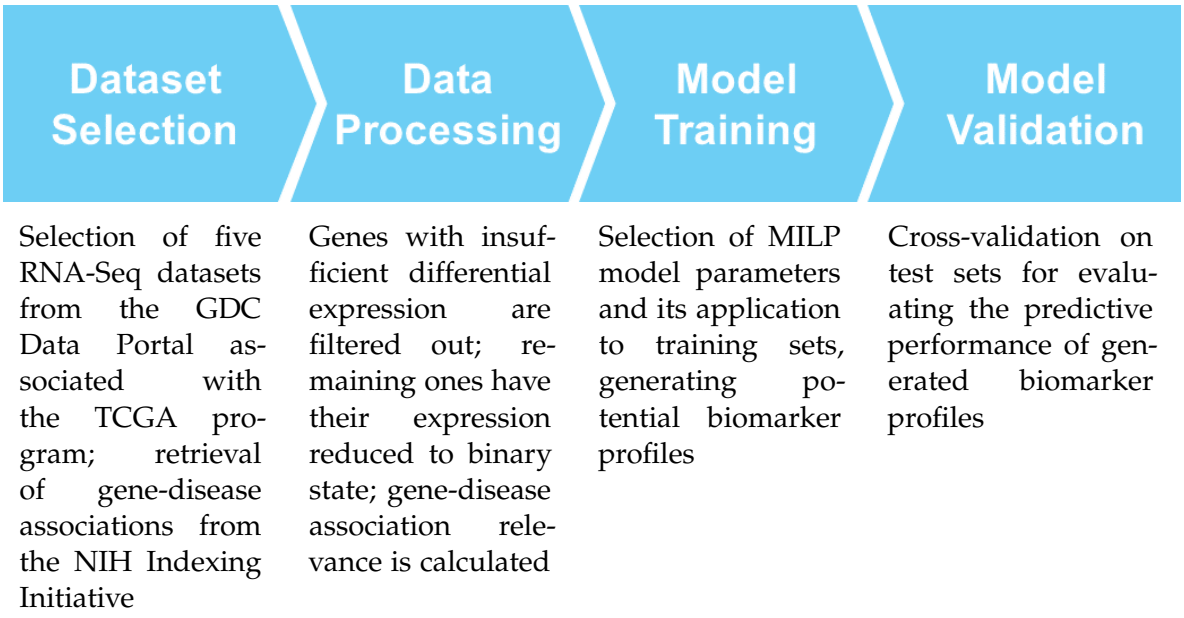



Figure 1.: Generic machine learning-based biomarker discovery workflow, with specific descriptions directed towards the methodology developed in this work.

value. Figure 1 summarizes the workflow that will be followed when generating biomarker profiles.

Adding to this, there are also various goals that were achieved during this project, which include:

- identifying potential biomarker profiles for five public RNA-Seq gene expression datasets for breast, colon, lung and prostate cancer, and head and neck squamous cell carcinoma;
- evaluating the predictive performance of the genetic signatures on one of the datasets, the prostate adenocarcinoma dataset;
- comparing the predictive performance of the models generated by this method with other known machine learning methods;
- analysing the biological significance of predicted potential biomarkers.

# 2

# STATE OF THE ART

This chapter contains a short review of the biomarkers discovery process and the most relevant methods used for their identification.

## 2.1 BIOMARKERS AND THEIR RELATION WITH DISEASE.

The term biological marker, or biomarker for short, was introduced in 1989 as a MeSH (Medical Subject Heading) term, which can be summarized as a biological parameter, measurable and quantifiable, which is representative of a specific health or disease state. Adding to this, an NIH group went further and standardized the definition of a biomarker as a characteristic that can be measured in an objective manner and can be considered an indicator of a series of different processes, which can be pathogenic as a result of clinical intervention, while also defining several existing types of biomarkers [17].

A biomarker may be measured on a biosample, as is the case of blood, urine or saliva, it may be a recording obtained from an individual, such as blood pressure or an electrocardiogram, or an imaging test. Biomarkers may also be generally classified according to their different possible applications such as risk (evaluates the risk of developing a certain disease condition), diagnostic (recognizing overt disease presence) and prognostic biomarkers (predicting future disease course) [17, 61, 73]. Considering this, it is important to note that the desirable properties of biomarkers vary according to their intended use [63]. However, regardless of their purpose, we are able to assess a specific group of features that a biomarker has to have, namely high sensitivity and specificity, it can be reproducibly obtained through standardized methods, acceptable to the patient in question and can be easily interpreted by clinical staff [71].

As a general rule, early detection of a disease condition plays a crucial role in successful therapy, where, in most cases, the earlier a disease condition is diagnosed, the more likely it can be successfully cured or well maintained. A dramatically reduced severity of the impact of the disease on the patients life results from this early management of the disease, allowing for the prevention or delay of subsequent complications. However, the majority of systemic disease states are not diagnosed until morbid symptoms emerge at a late

stage [64]. Molecular disease biomarkers, be they DNA, RNA or protein molecules, may prove to be key in overcoming this challenge, being that they act as indicators of particular physiological states and may reveal hidden lethal threats before the disease reaches a state where treatment becomes difficult. To this end, the impact and effectiveness that biomarkers may have for diagnostic use have been demonstrated [104], being able to detect genetic alterations through molecular diagnostics [106], as well as reaching the point of detecting abnormal nucleic acids and proteins in bodily fluids [127]. However, several constraints still limit our capability to be able to effectively recognize the full potential of disease detection, which come down to mainly three [64]: lack of definitive molecular biomarkers for a variety of different diseases, lack of an easy and inexpensive method for sampling and lack of a platform that is portable, accurate and easy-to-use, aiding in early disease detection.

## 2.2 Methods for Biomarker Discovery

While at first the search for biomarkers was essentially hypothesis-based, the advent of high throughput omic technologies in the past two decades has shifted the focus to a more discovery-based approach, thanks to the readily availability of large, quantitative datasets of differentially expressed mRNAs and proteins from case control studies generated by these technologies. Over the years, a multitude of sophisticated methods have been developed or adapted for the prediction and identification of potential molecular biomarkers, with most of these performing at first a more statistical analysis of these large datasets, many with relative accurate prediction capabilities. A few of these include:

- **Linear and Logistic Regression:** a common method for identifying a small, highly predictive set of biomarkers. While the focus was initially on univariate regression, selecting potential biomarkers taking solely into account their isolated predictive power, it was quickly shown that multivariate regression, considering potential influences between the selected markers, resulted in better models for effectively classifying new samples [145]. One recent example used multivariate logistic regression analysis for determining how gene promoter methylation varied between healthy control samples and samples with histopathologically confirmed adenoma of colorectal cancer, with several potential epigenetic markers being signalled for further investigation [5]. Another study also includes the use of multivariate regression in multidimensional analysis, essentially integrating miRNA, gene and protein data into the analysis, in which seven miRNAs, four genes and one protein were identified as prognostic biomarkers as being able to distinguish between aggressive and non-aggressive forms of colorectal cancer [72].

- **Support Vector Machines:** generally used for binary classification, a kernel function is used to map training data onto multidimensional space, with a dividing vector in high dimensional space, a hyperplane, being chosen capable of separating the two classes. In it, the margins between the hyperplane and the correct data points on each side which are the closest are maximized, while the distance to misclassified elements is minimized. With a SVM, separation under non-linear circumstances is achievable, which can prove to be a major benefit when dealing with omics data [84]. Examples of methodologies which use SVM-driven classification include utilizing seven lncRNAs as a candidate biomarker panel to improve early diagnosis of human pancreatic cancer [144] and improving early stage breast cancer detection by using miRNA differential expression profiles for SVM training [103].

- **Decision trees and random forests:** this approach consists of constructing a tree of attribute tests in such a manner that by reaching the terminal leaves, a correct classification has been achieved for any test sample. In the training phase, the tree is built as each branching imply a reduction in the resulting subgroup heterogeneity. Random forests emerge by compiling an ensemble of related trees which have been built each with a random sampling of the training data and data features, potentially conferring additional accuracy [60]. Various recent applications for this technology exist, namely involving the establishment of a 42-gene expression signature for distinguishing proteasome inhibitor treatment response in multiple myeloma [75] and identifying a total of 36 miRNAs for accurate classification of Alzheimer's disease [70].

- **Artificial Neural Networks:** inspired by biological neural networks, these consist, in essence, of a set of nodes akin to neurons that have associated functions or rules on how to process input and links between these with dynamic weights. An ANN is trained on data and, by comparing its output with the expected one, an error arises which is fed back to the neural network to adjust the weights. While less commonly used in biomarker prediction due to their tendency to generate models with minimal interpretability, they have shown promise when used to identify metabolic profiles for lung cancer screening [85].

However, while scientists praised the influx of high-throughput experimental data from omics technologies, namely genomics and proteomics, the bottleneck for the life science studies quickly went from generating data to interpreting results so as to infer the mechanisms behind the biological processes being studied.

## 2.3 TEXT MINING IN BIOMARKER DISCOVERY

Text mining has seen increasing popularity in applications in biomedical literature during the last decade [29], even if its first attempts were only made at the beginning of the millennium [115]. Biomedical text mining, which is how it is often referred to, has the main goal of extracting useful knowledge from the large repository of published scientific literature in the biomedical field available today.

Scientific literature provides itself as a rich wealth of information, be it for assessing the state-of-the-art of a particular field, or for providing information for constructing hypotheses for posterior experimental validation, or even for helping in providing explanations for experimental results [33]. There are many different databases which provide access to scientific literature in the life sciences domain [53], with PubMed being the most popular.

Another thing to consider is that the number of published articles in the biomedical field has been growing at an increasing rate. This situation may be related with the emergence of high throughput gene and protein profiling techniques, which allow simultaneous analysis of thousands of genes and proteins in multiple different conditions. This obviously leads to an overflow of data and information, which no researcher can ever hope to be able to manually curate. At most, the information retrieval process can consist of queries to databases of specific, pre-selected keywords, which ultimately result in incomplete search results, compromising the complete potential of these bibliographic databases [56].

This is where the automated analysis of text, which is usually shortened to text mining, can come in, helping researchers in assessing current and past scientific literature. A more complete definition of text mining has been provided by Marti Hearst, where it involves the discovery through *in silico* methods of new concepts and information through the automatic extraction of information and establishment of relations across multiple written sources, allowing the discovery of previously unknown and hidden meanings [46]. Text mining has many use cases in the biomedical field, ranging from drug target and biomarker discovery, creating a full overview of the state-of-the-art of a specific condition, to even setting up databases for specific domains [2, 34, 43, 52, 54, 55, 59, 92]. However, automated extraction of relevant knowledge and information from written resources is not a trivial task due to their heterogeneous nature, which explains why it developed into a fully fledged field in the biomedical sciences. Text mining in general employs several text processing and machine learning approaches, allowing the extraction of information from databases related to biological pathways, and genomic and proteomic expression profiles.

Text mining has several distinct steps associated, which have been previously reviewed [146, 7, 49] and are illustrated in figure 2. These consist of four main stages: information retrieval, named entity recognition, relation extraction, and general knowledge discovery. All text mining applications, to which biomedical ones are no stranger, implement one or

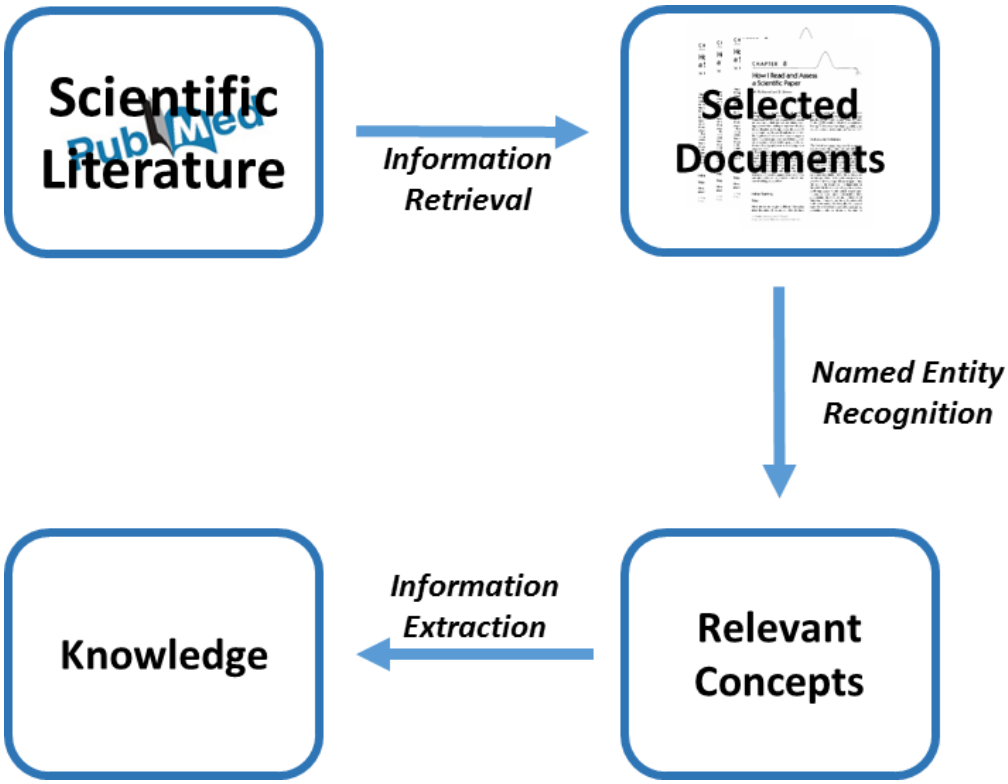



Figure 2.: General overview of a text mining workflow.

more of these tasks, while usually either resorting to providing their own, unique solutions for achieving these or implementing ones already described in the literature.

### 2.3.1 *Information Retrieval*

Before any analysis can be performed, a specific subject of interest needs to be selected to be able to retrieve textual sources relevant to that same topic. The information retrieval step is responsible for exactly this, which is usually achieved by performing queries with a set of specific keywords to bibliographic databases, the most used of which is PubMed for abstracts [34]. Articles are not, however, the only relevant sources of information, which may include medical records, patents, websites and more [53, 31, 35, 32].

As an alternative, several text mining applications extend the idea of querying databases by also resorting to keywords similar to the ones provided by the user, which include synonyms and alternative names, forming concepts from controlled vocabulary sources which include all of these. Queries which include this method may provide more comprehensive results. Adding to this, text mining tools may also classify the retrieved documents according to their content, performing sentence extraction for highlighting relevant parts of the documents for subsequent stages.

### 2.3.2 *Named Entity Recognition*

After all relevant documents have been retrieved, these are usually analysed for identifying specific keywords and how these relate to one another through an essential step, named entity recognition. Put quite simply, a named entity is a keyword or group of keywords which can be clearly identified as a specific concept, which here implies a biological entity. During the recognition process, any such keywords found in the text sources are linked to the specific concepts that are being referred in the text. An example of this may involve recognizing a gene not only by its name but also by its symbol, synonyms and potential past names.

The whole process of recognizing genes and proteins is one of the most complex in named entity recognition, mostly due to the different symbols and names that these can use but also due to the fact that these may be shared between completely different biological entities [77].

Several techniques exist for trying to solve this issue. The most common involves the use of controlled vocabularies, in which keywords have been previously grouped together in accordance with their specific category for facilitating matching in text [96].

Ontologies present themselves as an alternative for named entity recognition, defining more formally these concepts and their associated keywords, while also including relationships and specific rules for determining how these concepts relate to one another, as well as domain information related to biological pathways and diseases [140].

Machine-learning is also of popular use in named entity recognition [116]. A text mining tool might resort to conditional random fields [107, 66], hidden [141, 139] and maximum entropy Markov Models [23], and support vector machines [88, 42] for achieving this task. However, these methods require training on annotated data which is representative of what the method might encounter on real world documents before the named entity recognition task can be performed.

### 2.3.3 *Relationship Extraction*

Extraction of relationships from text can be performed after all relevant entities have been identified or even during this process, as its purpose is to detect explicit and implicit connections between these concepts. There are currently two ways through which relationship extraction can be performed:

- **Natural Language Processing.** Natural language processing based methods work by understanding the flow and structure of natural languages. Information of how the language is structured usually has to be previously supplied to these methods, together with information regarding how biological concepts can occur in text. These

provide information regarding the type of relationships between concepts, as well as being able to identify relationships between more than two concepts. They are however more computationally intensive and are trained to detect pre-defined relationships.

- **Co-Occurrence.** Co-occurrence-based methods consider one simple assumption, that two concepts which occur frequently together in the same text are functionally related. To prevent false positives, as many terms co-occur in text without being necessarily functionally related, co-occurrence-based methods try to assign a degree of confidence to the relation through scoring methods [4]. The nature of these methods implies that they are easy to implement but generally result in lower precision when compared with natural language processing based methods and do not provide any information respective to the relation that the concepts have towards one another. These also tend to be of a probabilistic nature, with probabilistic models seeing large growth in their use for this exact purpose in the literature.

Both types of methods are used together, with co-occurrence-based methods being used for detecting relationships and natural language processing based methods being used to ascertain the type of relationship described.

In any case, relationship extraction bares numerous issues when analysing common English texts and even more so for biomedical literature. These include resolution of coordination by linking structures connected by terms such as *and* and *or*, references to entities that have been previously mentioned by terms such as *it* and *they*, and correctly interpreting negation. However, these tend to be ignored by most approaches for biomedical relationship extraction, especially the latter two issues.

## 2.4 Gene Expression Data Analysis for Biomarker Discovery

High-throughput sequencing technology is quickly turning into the standard method for performing RNA expression levels analysis, or RNA-Seq [78]. These rapid sequencing technologies changed the landscape of practically every field in the life sciences, bringing with them lower sequencing costs together with higher level of detail for RNA-Seq profiles [14]. Adding to this, RNA-Seq technology also allows for the detailed identification of gene isoforms, nucleotide variations, post-transcriptional base modifications and translocation events [130]. The main purpose of these technologies involves identifying differentially expressed genes, up- or down-regulated, in relation to two or more specific clinical conditions or functional pathways, with one usually being a control state. Through the type and volume of data that these technologies have made possible, a large variety of biomarker

models have since been designed for diagnostic and prognostic purposes in many different types of diseases.

Biomarker discovery studies in this field started by simply comparing two classes of samples to see if gene expression differences were detectable by statistical methods and, if so, how they could be used to help predict how a disease emerges or progresses [94]. However, it was van't Veer *et al.* [125] that truly showed the possibilities of using gene expression profiles for predicting clinical outcomes. In this study, they were able to discover a set of 70 genes whose expression profiles allowed classifying the patients in accordance to how they were responding to therapy. Today, most RNA-Seq analysis approaches are based on widely-investigated tools and methods, integrating technical knowledge that ranges from statistical analysis to machine learning and data mining.

Typically, a RNA-Seq dataset can be represented by a matrix where the rows and columns can represent samples and genes, respectively. In this scenario, a row would be representative of a vector of expression values of the genes of a specific sample while a column would be representative of a vector of expression values of a specific gene in all the samples of the dataset. Such a dataset is subject to standardization and pre-processing procedures, after which the most differentially expressed genes will be determined. Several hypothesis-testing approaches are available to achieve this goal by using differential expression or fold change thresholds, which can be supported by unsupervised classification methods, such as clustering, and visualization. Filtering approaches can be used as well for further removing genes which have been proven to be either uninformative, noisy or redundant for posterior analysis. Supervised classification methods usually follow this step, which includes techniques capable of performing 'wrapper' and 'embedded' feature selection. There have been many attempts [67, 111, 112, 132] at trying to evaluate which supervised classification method or set of methods might be best at dealing with RNA-Seq data, but so far no definitive conclusion has been reached from such undertakings.

After models are obtained, their predictive capabilities need to be evaluated, which can be achieved through cross-validation, and combined with other types of validation methods. However, in the end, effective analysis of the obtained results can only be properly performed when multiple information sources are also considered, which may range from literature to curated functional annotations, as is the case of Gene Ontology terms, and disease-related pathways stored in available databases.

Returning to one of the first steps of the general RNA-Seq analysis, data pre-processing is essential when dealing with this type of data as it helps to tackle problems such as digital re-formatting, data encoding, missing data, initial filtering and data transformation, with the latter usually involving processes for performing data re-scaling, normalization and standardization. All of these techniques are essential for assuring data integrity and quality before proper integrative analysis and modelling can be performed [9].

# 3

# METHODOLOGY

This chapter will present the methods developed and applied as part of this work, namely the mixed-integer linear programming (MILP) model. Contrary to the increasing trend for the application of black-box machine learning models in problems that warrant this type of approach, the presented model is a white-box model, with a clear goal and restrictions. Adding to this, the produced solutions are simple to understand by biomedical personnel, facilitating its potential application for the prediction of biomarker profiles for any present and future disease. The chosen datasets will also be described, as well as how feature extraction and the retrieval of gene-disease associations from the literature was performed. Figure 3 illustrates and summarizes the whole pipeline that will be described in the following sections.

## 3.1 DATASETS

The Genomic Data Commons (GDC) is a research program from the National Cancer Institute (NCI) whose goal is to establish a single, standardized group of clinical and genomic cancer datasets. The GDC is also a data sharing platform as it provides access to high-quality NCI-generated data present in datasets such as the The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Therapies (TARGET).

A total of five gene expression quantification datasets were retrieved from the GDC data portal as part of The Cancer Genome Atlas (TCGA) program, related to the specific projects TCGA-BRCA, TCGA-COAD, TCGA-HNSC, TCGA-LUAD and TCGA-PRAD which deal with breast invasive carcinoma, colon adenocarcinoma, head and neck squamous cell carcinoma, lung adenocarcinoma and prostate adenocarcinoma, respectively. Each of these contain the calculated expression signal for each and all 60,483 considered genes, per sample. A summary of the contents of each dataset is provided in Table 1.

As a reference for future sections, it can be assumed that any dataset can be represented as $D = \{x_j, y_j\}$, with $j \in \{1 \dots N\}$, then $x_j = \{x_{j1} \dots x_{ji}\}$, with $i \in \{1 \dots P\}$, representing

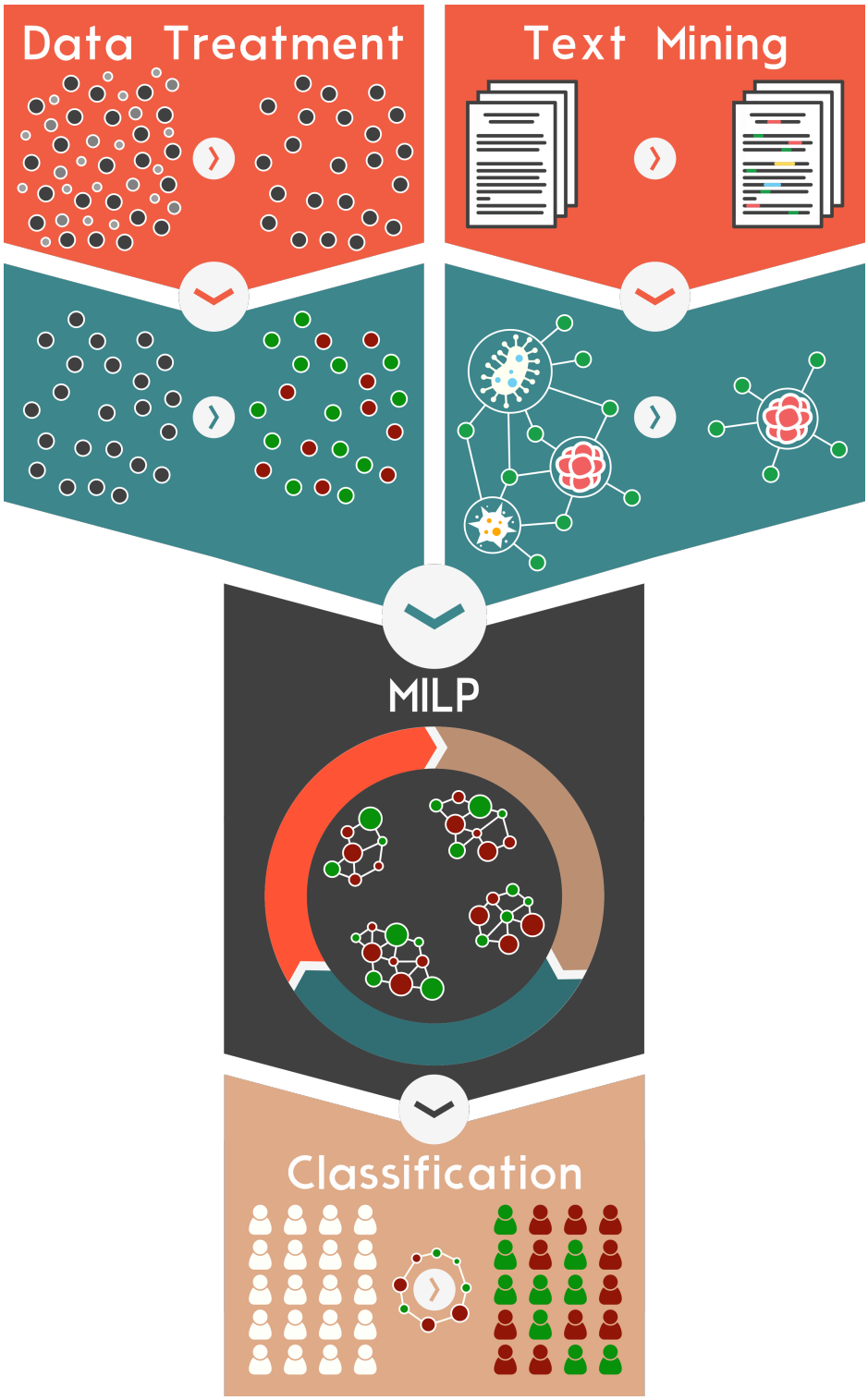



Figure 3.: Pipeline visually describing the methods applied and developed as part of this work. *Data Treatment* shows how the most relevant features were selected and grouped, while *Text Mining* shows the retrieval of the gene-disease associations. *MILP* illustrates the data fitting process, while *Classification* shows status prediction on new samples.

Table 1.: A comparison of the different datasets built from the data available at the GDC portal.

| Datasets | Cases | Samples | Features | Controls | Diseased |
|---|---|---|---|---|---|
| TCGA-BRCA | 1,097 | 1,222 | 60,483 | 113 | 1,109 |
| TCGA-COAD | 459 | 521 | | 41 | 480 |
| TCGA-HNSC | 528 | 546 | | 44 | 502 |
| TCGA-LUAD | 519 | 594 | | 59 | 535 |
| TCGA-PRAD | 498 | 551 | | 52 | 499 |

the vector of $P$ features describing the $j$-th sample, while $y_j$ is a binary variable representing the class label, 0 for control and 1 for diseased.

## 3.2 Feature Selection and Data Pre-processing

Feature selection and data pre-processing were achieved through two steps in the pipeline, a filtration step and a binarization step. After standardizing the gene expression values within each sample, a feature filtration step is taken where features whose overall discriminating value is insufficient are not considered in posterior steps. This is done to each feature by first evaluating its expression values' standard deviation and then comparing it to a set threshold. If a gene's expression value does not show enough variance between the samples, it is discarded as it can be assumed that it would not help in trying to achieve a clear distinction between the established classes, controls and diseased. This step allows discarding most of the features in the data, reducing its complexity and allowing for a more timely solution convergence.

The selected thresholds $\tau$ were calculated through equation 1, where $\tilde{\sigma}$ represents the median of all the genes expression values' standard deviation and $\psi$ acts as a weight, with higher values requiring genes to present a higher variance between samples. The values chosen for $\psi$ were 3 and 3.5, as these reduced the number of considered features to approximately 5% of the total number of 60,483 features. Adding to this, equation 2 indicates how all potential non-relevant genes were discarded, by comparing the standard deviation of each gene $i$ expression values across all samples $j$ with the calculated threshold $\tau$.

$$\tau = \psi \cdot \tilde{\sigma}, \qquad \psi \in \mathbb{R}^+ \quad (1)$$

$$x_{ji} = 0, \quad \text{if} \quad \sigma_i < \tau, \qquad \forall i, j \quad (2)$$

The binarization step comes after the filtration step, where the genes expression values were reduced to a Boolean variable. This step is essential for the optimization step represented by the equations 6-15, as it allows representing the expression of a gene by merely indicating if

it is present or absent in a sample. Equation 3 demonstrates how binarization was achieved, where $x_{ji}^0$ denotes the baseline expression value present in the raw dataset for gene $i$ in sample $j$ and $\tau$ is also used as the minimum a gene has to be under- or overexpressed to be considered as present in the same sample.

$$x_{ji} = \begin{cases} 1, & \text{if } \left|x_{ji}^0\right| > \tau \\ 0, & \text{if } \left|x_{ji}^0\right| \leq \tau \end{cases} \qquad \forall i, j \tag{3}$$

## 3.3 ANNOTATIONS

The annotations and associations between genes and diseases retrieved from existing scientific literature represent an integral part of the optimization model proposed in this work. The US National Library of Medicine (NLM) Indexing Initiative provides an in-depth database of biomedical information where a fully automated system named the NLM Medical Text Indexer (MTI) uses the title and abstract of articles available in the PubMed/MEDLINE database to suggest MeSH terms which are then reviewed by selected indexers. For the purposes of this work, only the co-occurrence data is utilized, which is based on validated index terms.

Through these annotations, it was possible to retrieve the citation frequency $\phi$ for each relevant gene-disease association, from which equation 4 shows how to obtain the relevance $\phi^*$ of each association, which is integrated in the optimization model. Adding to this, the smoothing factor $\epsilon$ helps establishing the relative importance of the actual number of citations, with higher values providing a smoother regularization effect for a specific gene, while lower values present a more aggressive approach. The best value for $\epsilon$ is empirically chosen from a set of sensible values used in the analysis.

$$\phi^* = \begin{cases} \left(\frac{1}{\phi}\right)^{\epsilon}, & \text{if } \phi > 0 \\ 2^{\epsilon}, & \text{if } \phi = 0 \end{cases} \qquad \epsilon \in \left]0, 1\right], \phi \in \mathbb{N}_0^+ \tag{4}$$

## 3.4 OPTIMIZATION MODEL

In this work, an alternative approach for performing biomarker discovery is proposed, one that integrates prior biological knowledge extracted from current public databases in the form of an adaptive-LASSO regularization in a mixed-integer linear programming (MILP) model based on the one previously developed by Santiago *et al.* [100].

Having established the relevant gene-disease associations and selected the most potentially significant set of genes, this information was integrated in an adaptive $L_1$-regularization

MILP model. LASSO-regularization in optimization models has been shown to produce good results when used in large P, small N settings, such as the biomarker discovery problem, by being able to overcome overfitting issues. Additionally, as the datasets used are very imbalanced, with the diseased class greatly outnumbering the controls around 10:1, a penalty is introduced for misclassifying the minority class as show in equation 5. In contrast, correctly classifying the minority class provides greater benefits to the model than the majority class. $\rho_m$ represents the penalty for the minority class, while $\rho_M$ represents the penalty for the majority class. In the same way, $f_m$ indicates the number of samples in the minority class, while $f_M$ indicates the number of samples in the majority class. This introduces a bias to the model to pay equal attention to both classes.

$$\rho_m = \frac{f_M}{f_m}, \quad \rho_M = 1 \tag{5}$$

Adding to this, features are also grouped into two sets, a disease-oriented set $P_D$ and a control-oriented set $P_H$. A feature's absolute and relative frequency is calculated for each of the class labels in the whole dataset, after which they will be inserted into each respective set. The chosen minimum thresholds were 7 for the absolute frequency and 0.5 for the relative frequency, meaning that, to be grouped into any set of one of the class labels, any feature would have to be present in at least 7 samples of that class, which represents approximately 1% of the total number of samples in the considered datasets, and these needed to represent at least 50% of the total number of samples in which the feature was present.

Now, the goal will be to discover the minimal number of features that are able to explain the different class labels between each sample. Equation 6 shows the minimization problem's objective function that this approach tries to optimize for a classification scenario. The objective function is composed of three independent components:

- $\lambda_1 \sum_j^N \rho_j \left(1 + (2 \cdot y_j - 1) \sum_i^P w_i \cdot x_{ji}\right)$, where the goal is to maximise the individual score for each sample in a way that allows for correct classification of its disease status;
- $\lambda_2 \sum_i^P s_i$, which favours the minimisation of the number of selected biomarkers;
- $\lambda_3 \sum_i^P \phi_i^* \cdot s_i$, which integrates the gene-disease associations extracted from the literature and favours the selection of genes as biomarkers that are known to interact with the considered disease

In these, $\lambda_{1-3}$ dictate the weight of each component in the objective function, supplying it with the needed flexibility to potentially find the solution that provides the most optimal disease class separation.

$$\underset{w,s}{\text{minimize}} \quad \lambda_1 \sum_j^N \rho_j \left(1 + (2 \cdot y_j - 1) \sum_i^P w_i \cdot x_{ji}\right) + \lambda_2 \sum_i^P s_i + \lambda_3 \sum_i^P \phi_i^* \cdot s_i \tag{6}$$

$$\text{subject to} \quad s_i \cdot w_i^L \leq |w_i| \leq s_i \cdot w_i^U, \qquad \forall i \in \{1 \dots P\} \tag{7}$$

$$\sum_i^{P_H} w_i \leq \sum_i^{P_H} s_i, \tag{8}$$

$$\sum_i^{P_D} w_i \geq -\sum_i^{P_D} s_i, \tag{9}$$

$$\sum_i^{P_D} s_i \geq 1, \tag{10}$$

$$\sum_i^{P_H} s_i \geq 1, \tag{11}$$

$$\sum_i w_i \cdot x_{ji} > -\left(1 + y_j \cdot U\right), \qquad \forall j \in \{1 \dots N\} \tag{12}$$

$$\sum_i w_i \cdot x_{ji} < \left(1 - y_j\right) U + 1, \qquad \forall j \in \{1 \dots N\} \tag{13}$$

$$w_i \in \mathbb{R}, \qquad \forall i \in \{1 \dots P\} \tag{14}$$

$$s_i \in \{0, 1\}, \qquad \forall i \in \{1 \dots P\} \tag{15}$$

In it, $w_i$ and $s_i$ represent the $i$-th feature's weight and if it has been selected by the solver, respectively, while $x_{ji}$ contains $i$-th feature behaviour for the $j$-th sample. Adding to this, $\phi_i^*$ helps steer the model towards features which have shown to be most relevant considering the biological knowledge previously obtained from the literature. Finally, several combinations of values for $\lambda_{1-3}$ were tested to determine the best weight of each component in the objective function.

These variables are, however, subject to several constraints. Equation 7 binds the absolute weight of a feature $i$ to $s_i$, with $w_i^L$ and $w_i^U$ equal to 0.5 and 2, respectively. This allows each potential biomarker to, by itself, reach and go beyond the required score for classifying any sample, indirectly helping to minimize the number of required features for correctly classifying each sample in the training process, while also forcing selected genes to have a minimum relevance to be considered as such.

Additionally, equations 8 and 9 ensure that a feature's weight reflects its importance in the resulting score, while equations 10 and 11 guarantee that at least one feature is selected from each of the sets $P_D$ and $P_H$.

Furthermore, equations 12 and 13 force the sum of the feature's weights in a sample to be at least -1 for healthy samples and at most 1 for diseased samples, not inclusive. In these, as the only interest is in favouring a correct, or, at least, an ambiguous classification, $U = 50$,

being high enough to turn the constraint inconsequential when it is being considered. This guides the model towards predicting control and diseased samples as healthy and diseased, respectively, while also providing some flexibility, allowing the model to predict a sample as ambiguous.

As such, equations 6-15 formulate a complete MILP model that is able to be solved to global optimality using CPLEX (ILOG 2013) through the PULP package in Python 3.6 for identifying the best values for the $w_i$ parameter. It is then used in the equation 16, allowing us to determine the status of a new sample, whether it is healthy ($S_j \geq 1$) or diseased ($S_j \leq -1$). It is also possible for a sample's status to be ambiguous if $-1 < S_j < 1$, requiring further investigation.

After the training step, what is obtained is a set of biomarkers, each with their own specific weight calculated by the model. For all samples to be tested, equation 16 allows the determination of their class label and, as such, determine its health status.

$$S_j = \sum_i w_i \cdot x_{ji} \tag{16}$$

In the following sections, the results of this study will be presented and discussed, as well as the model's viability for the potential diagnostic of any specific disease for any group of samples.

## 3.5 Model Parameter Optimization

It is important to the determine what are the combinations of parameters that provide the most performant biomarker profiles. To achieve this, a validation step is performed, where several combinations of the sample space are evaluated for each of the pipeline's parameters and weights, while also performing a cross-validation analysis on the trained models. The parameters in question include the optimization weights $\lambda_{1-3}$, the filtering threshold factor $\psi$, and the smoothing factor $\epsilon$. It is also important to note that training the models and, by consequence, performing the cross-validation analysis is very consuming in terms of time and processing power. As such, this analysis could only be performed on one of the selected datasets, whose best set of parameters will then be applied to the remaining datasets when performing the contextualization analysis. The chosen dataset was the prostate adenocarcinoma (PRAD) one, given that it possessed an intermediate number of samples which helps performing this analysis in a timely manner.

The selected values for the optimization weights $\lambda_{1-3}$ were $\{0.01, 0.10, 1.00\}$, while the selected values for the filtering threshold factor $\psi$ were $\{3.0, 3.5\}$, and for the smoothing factor $\epsilon$ were $\{0.1, 0.5, 1.0\}$. Adding to this, the quality of the generated models was also evalu-

ated for when the number of selected biomarkers was not relevant ($\lambda_2 = 0.0$) and when gene-disease associations extracted from the literature were not considered ($\lambda_3 = 0.0$). The resulting combined sample space included 288 unique combinations whose performance was evaluated through stratified 20-fold cross-validation.

## 3.6 OTHER CLASSIFIERS FOR PERFORMANCE COMPARISON

Another important way of evaluating the quality of the trained models through the method developed in this project is by comparing it to existing machine learning algorithms. Three classifiers were chosen as a base for this comparison analysis and implemented through the Scikit-learn Python package, K-Nearest Neighbors (KNN), Support Vector Machines (SVM) and Random Decision Forests (RDF). It is important to consider that this analysis was also performed on the TCGA-PRAD dataset, where the number of filtered features $f_n$ was 5,730 when $\psi = 3.00$ and 5,140 when $\psi = 3.50$.

Several different sets of parameters were tested for each algorithm. For the KNN classifier, these include the desired number of neighbors $k$ when querying the tree, the weight function used in predicting new points, the algorithm used to compute the nearest neighbors, the leaf size of the chosen algorithm, the distance metric used and, when using the Minkowski metric, its power parameter $p$.

The number of neighbors $k$ evaluated ranged from 1 to 7. The weight function can be uniform, where all points in the neighborhood are weighted equally, or distance based, where points are weighted by the inverse of their distance to the queried point. The available algorithms include two binary search tree ones, the Ball tree and $k$-d tree algorithms, with the first partitioning the data points into nested sets of hyperspheres and the second generating hyperplanes that split the data space into two parts at each non-leaf node. The tree leaf size is $2^i$, with $i \in [0, 8]$, specifying roughly the number of points at which the tree switches to brute-force, with lower values decreasing the cost of accessing any specific node but increasing the cost of computing the distance function.

Finally, there is also the distance metric used, which includes the Manhattan, Euclidean, Chebyshev and Minkowski metrics. The Minkowski metric is equivalent to the Manhattan and Euclidean metrics when $p = 1$ and $p = 2$, respectively. Adding to this, $p \in \{1.5, 2.5, 3.0, 3.5, 4.0\}$ when testing the Minkowski metric. Equations 17a and 17b describe how these distances are calculated for any two points $x$ and $y$.

$$D_{\text{Chebyshev}} = \max\left(|x - y|\right) \tag{17a}$$

$$D_{\text{Minkowski}} = \left(\sum |x - y|^p\right)^{\frac{1}{p}} \tag{17b}$$

For the SVM classifier, the evaluated parameters include the penalty of the error term $C$, for which higher values generate more complex models with more selected features, with $C \in \{0.025, 0.10, 0.25, 0.50, 0.80, 1.00, 1.50, 2.00\}$, and the type of kernel to be used. The available kernels include a linear, polynomial, gaussian and sigmoid kernel function, as described by equations 18a to 18d. Adding to this, several degrees $d$ of the polynomial kernel were evaluated, with $d \in [2, 5]$, as well as the kernel coefficient $\gamma$ for the polynomial, radial-based and sigmoid kernel functions, with $\gamma \in \left\{\frac{1}{n}, 0.10, 0.25, 0.50, 1.00, 2.00\right\}$.

$$\text{linear:} \quad \langle x, x' \rangle \tag{18a}$$

$$\text{polynomial:} \quad \left(\gamma \langle x, x' \rangle + 0\right)^d \tag{18b}$$

$$\text{gaussian:} \quad e^{-\gamma \|x - x'\|^2} \tag{18c}$$

$$\text{sigmoid:} \quad \tanh\left(\gamma \langle x, x' \rangle + 0\right) \tag{18d}$$

For the RDF, the evaluated parameters include the total number of estimators/trees in the forest, ranging from 1 to 19, the criterion function for measuring the quality of a split, the maximum depth of each tree, and the maximum number of features to consider when searching for the best split. The criterion functions available include the Gini impurity, which measures the chance of a randomly chosen sample to be incorrectly labelled if it was randomly labelled following the label distribution in a specific subset, and information gain, which at each step chooses the split that results in the least amount of class entropy. Finally, the maximum depth of each tree, when considered, is either 5 or 8, and the chosen maximum number of features is $\left\{1, \log_2 n, \sqrt{n}, n\right\}$.

To properly evaluate the quality of the models produced by each classifier, 10-fold stratified cross-validation with 20 repeats was performed for each set of parameters.

## 3.7 METRICS FOR ERROR ESTIMATION

Several metrics are used for determining the performance of each predicted genetic signature. These include accuracy, sensitivity, precision and specificity, together with the F-score and the Matthews Correlation Coefficient (MCC), as described by equations 19a to 19f. These are calculated from the confusion matrix values representing the four possible out-

comes when performing binary classification, namely the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

$$\text{Accuracy:} \quad ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{19a}$$

$$\text{Sensitivity:} \quad TPR = \frac{TP}{TP + FN} \tag{19b}$$

$$\text{Precision:} \quad PPV = \frac{TP}{TP + FP} \tag{19c}$$

$$\text{Specificity:} \quad TNR = \frac{TN}{TN + FP} \tag{19d}$$

$$\text{F}_1\text{-score:} \quad F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{19e}$$

$$\text{MCC:} \quad MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{19f}$$

The $F_1$-score represents the harmonic mean between the precision and sensitivity, and ranges from 0 to 1, where the latter indicates perfect precision and sensitivity. Generally, the F-score, while being widely used in machine learning, does not take into account true negatives, potentially introducing a bias when comparing the performance of any two methods.

As such, while a single number can not perfectly describe all the outcomes, the MCC proves preferable as it takes into account all performance outcomes simultaneously, while also compensating for class imbalances [93]. MCC returns a value between $-1$ and $+1$, with a coefficient of $+1$ representing perfect prediction capabilities, 0 being equivalent to random prediction and $-1$ indicating complete disagreement between the predictions and the observations.

Both metrics will be considered when evaluating the performance of the model developed in this project as well as all other methods or techniques used for performance comparison, but a special focus will be given to the MCC as it can potentially better represent the actual performance of any model.

Adding to this, two types of evaluations were performed. This problem is treated as a ternary classification one, in which any sample can be labelled as control/healthy, diseased, or ambiguous. As such, results can be evaluated by either taking into consideration or ignoring samples classified as ambiguous when calculating the performance outcomes. For achieving these different scores, positive class samples classified as ambiguous are considered as Type II errors (false negatives) while negative class samples classified as ambiguous are considered Type I errors (false positives).

# 4

# RESULTS AND DISCUSSION

Several analysis steps were performed, including a search for the best combination or combinations of a subset of the pipeline's parameters' sample space, another in which the model's performance is evaluated against other typical machine learning algorithms on the same datasets, and, finally, after training for each of the considered datasets, the selected features biological significance will be examined and evaluated. All of these steps will be described in further detail in the following sections, together with the obtained results and their discussion.

## 4.1 RESULTS FOR PARAMETER OPTIMIZATION

The method developed in this project has several parameters that can significantly influence the generated biomarker profiles. The parameters in question include the optimization weights $\lambda_{1-3}$, the filtering threshold factor $\psi$, and the smoothing factor $\epsilon$. Several values for these parameters were evaluated in an attempt to discover the best possible combination, within the analysed sample space. Adding to this, due to the ternary classification system inherent to the method, in which any sample can be labelled as control/healthy, diseased, or ambiguous, the latter can be included or excluded from typical performance evaluations usually performed for methods with binary classification systems. Tables S1 and S2 provide the results for each evaluation, respectively. Tables 2 and 3 also show the best 10 parameter combinations for each evaluation.

Both Table 2 and 3 present us with useful information. Table 2 shows that, even though the models created present a high $F_1$-score, the Matthews Correlation Coefficient (MCC) is very close to 0, indicating that the models do not present an improvement when compared to simple random prediction due to the high rate of false positives.

However, looking at Table 3, if ambiguous samples are not considered when calculating the performance metrics, at least two patterns emerge.

First, the best fitted models placed less weight on the actual data extracted from the GDC datasets than the co-occurrence annotation data, with one exception. Furthermore, the results were also better when the filtering threshold factor was also higher, which naturally

Table 2.: MCC and $F_1$-score performance measures for the top 10 parameter combinations, when considering ambiguous samples. **%Ambiguous** indicates the percentage of samples that were classified as ambiguous.

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\epsilon$ | $\psi$ | $F_1$-score | MCC | %Ambiguous |
|---|---|---|---|---|---|---|---|
| 0.10 | 1.00 | 1.00 | 0.1 | 3.0 | 0.881 | -0.011 | 1.27 |
| 0.01 | 0.10 | 0.10 | 0.1 | 3.0 | 0.881 | -0.011 | 1.27 |
| 0.10 | 0.01 | 1.00 | 0.1 | 3.5 | 0.864 | -0.025 | 3.99 |
| 0.10 | 0.00 | 1.00 | 0.1 | 3.5 | 0.863 | -0.025 | 3.81 |
| 0.01 | 0.00 | 0.10 | 0.1 | 3.5 | 0.863 | -0.028 | 3.81 |
| 0.10 | 1.00 | 1.00 | 0.5 | 3.5 | 0.870 | -0.028 | 3.99 |
| 0.01 | 0.10 | 0.10 | 0.5 | 3.5 | 0.870 | -0.036 | 3.99 |
| 0.10 | 1.00 | 1.00 | 1.0 | 3.0 | 0.881 | -0.036 | 1.09 |
| 0.01 | 0.10 | 0.10 | 1.0 | 3.0 | 0.881 | -0.037 | 1.09 |
| 0.10 | 0.01 | 1.00 | 1.0 | 3.0 | 0.880 | -0.037 | 1.45 |

Table 3.: MCC and $F_1$-score performance measures for the top 10 parameter combinations, when ignoring ambiguous samples. **%Ambiguous** indicates the percentage of samples that were classified as ambiguous.

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\epsilon$ | $\psi$ | $F_1$-score | MCC | %Ambiguous |
|---|---|---|---|---|---|---|---|
| 0.01 | 1.00 | 1.00 | 1.0 | 3.5 | 0.880 | 0.702 | 69.51 |
| 0.01 | 1.00 | 1.00 | 0.5 | 3.5 | 0.845 | 0.652 | 64.43 |
| 0.01 | 1.00 | 0.00 | - | 3.5 | 0.736 | 0.333 | 40.11 |
| 0.01 | 1.00 | 0.01 | 0.1 | 3.5 | 0.736 | 0.333 | 40.11 |
| 0.01 | 1.00 | 0.01 | 0.5 | 3.5 | 0.736 | 0.333 | 40.11 |
| 0.01 | 1.00 | 0.01 | 1.0 | 3.5 | 0.736 | 0.333 | 40.11 |
| 0.01 | 1.00 | 0.10 | 0.1 | 3.5 | 0.736 | 0.333 | 40.11 |
| 0.01 | 1.00 | 1.00 | 0.1 | 3.5 | 0.710 | 0.282 | 40.11 |
| 0.01 | 0.10 | 1.00 | 0.1 | 3.5 | 0.691 | 0.240 | 40.11 |
| 0.01 | 0.00 | 1.00 | 0.1 | 3.5 | 0.685 | 0.232 | 38.66 |

results in fewer extracted features. This presents an interesting hypothesis, that the data used is too noisy to allow for effective class separation with a linear model and/or the method used to filter and extract the most relevant features was not the appropriate one. Both of these are entirely possible, especially the first, as the MILP algorithm developed in this project had to undergo particularly large changes to be able to be applied to the considered datasets when compared to the original developed by Santiago *et al* [100]. Adding to this, the average number of selected features for the top 2 models in table 3 was close to 2 and the number of samples classified as ambiguous was very high as demonstrated by the low performance scores when considering ambiguous samples, providing even more confirmation to this hypothesis. In any case, the parameters' values for the top 2 models in table 3 will be the ones considered when evaluating the chosen biomarkers for each of the GDC datasets in the contextualization analysis further down.

Table 4.: MCC and $F_1$-score performance measures for the top 10 parameter combinations for the KNN classifier.

| $\psi$ | $k$ | Weight Function | Algorithm | Leaf Size | $p$ | Metric | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|
| 3.0 | 7 | Distance | Ball | 64 | - | Manhattan | 0.954 | 0.295 |
| 3.0 | 7 | Uniform | $k$-d | 8 | 4.0 | Minkowski | 0.954 | 0.292 |
| 3.0 | 7 | Distance | Ball | 4 | 1.5 | Minkowski | 0.954 | 0.288 |
| 3.0 | 7 | Uniform | Ball | 32 | 2.0 | Minkowski | 0.954 | 0.288 |
| 3.0 | 7 | Uniform | $k$-d | 32 | 3.0 | Minkowski | 0.954 | 0.288 |
| 3.0 | 7 | Distance | $k$-d | 64 | - | Euclidean | 0.954 | 0.285 |
| 3.0 | 7 | Distance | $k$-d | 8 | 3.0 | Minkowski | 0.954 | 0.285 |
| 3.0 | 7 | Uniform | $k$-d | 4 | 4.0 | Minkowski | 0.954 | 0.285 |
| 3.0 | 7 | Distance | $k$-d | 32 | - | Euclidean | 0.954 | 0.282 |
| 3.0 | 7 | Distance | Ball | 16 | - | Manhattan | 0.954 | 0.282 |

Table 5.: MCC and $F_1$-score performance measures for the top 10 parameter combinations for the SVM classifier.

| $\psi$ | C | Kernel | Degree | Gamma | $F_1$-score | MCC |
|---|---|---|---|---|---|---|
| 3.0 | 1.500 | Polynomial | 2 | 2 | 0.946 | 0.201 |
| 3.5 | 1.500 | Polynomial | 5 | 1 | 0.950 | 0.199 |
| 3.0 | 0.100 | Polynomial | 4 | 1 | 0.946 | 0.198 |
| 3.5 | 0.025 | Polynomial | 2 | 0.1 | 0.946 | 0.196 |
| 3.5 | 0.025 | Polynomial | 3 | 0.5 | 0.947 | 0.194 |
| 3.0 | 0.800 | Polynomial | 2 | 0.1 | 0.944 | 0.194 |
| 3.5 | 2.000 | Polynomial | 2 | 2 | 0.945 | 0.194 |
| 3.0 | 1.000 | Polynomial | 2 | 2 | 0.944 | 0.193 |
| 3.5 | 0.025 | Linear | - | - | 0.943 | 0.189 |
| 3.0 | 2.000 | Polynomial | 2 | 2 | 0.942 | 0.187 |

## 4.2 CLASSIFIER COMPARISON ANALYSIS

The predictive performance of the models generated by the method developed in this work needs to be compared to existing methods typically used in machine learning. The chosen methods to which it will be compared are K-Nearest Neighbors (KNN), Support Vector Machines (SVM) and Random Decision Forests (RDF). Tables S3, S4 and S5 illustrate the complete results with a positive MCC while tables 4, 5 and 6 show the top 10 combinations of parameters, for each classifier.

The models generated by these classifiers present, at best, mediocre results for the TCGA-PRAD dataset. Of these, the KNN models managed to attain the highest MCC scores, with the highest score being 0.295, while the SVM models did not show as much success. Even so, all of the best models for each of these classifiers are able to better discriminate the samples' labels than the best MILP models, when considering samples classified as ambiguous in the performance metrics, when we compare the results in table 2 to tables 4-6. The same

Table 6.: MCC and $F_1$-score performance measures for the top 10 parameter combinations for the RDF classifier. IG stands for information gain.

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | $F_1$-score | MCC |
|---|---|---|---|---|---|---|
| 3.5 | 13 | Gini | $\sqrt{n}$ | 8 | 0.954 | 0.287 |
| 3.0 | 7 | IG | $n$ | $\infty$ | 0.953 | 0.285 |
| 3.0 | 11 | IG | $n$ | $\infty$ | 0.953 | 0.281 |
| 3.0 | 15 | Gini | 1 | $\infty$ | 0.954 | 0.280 |
| 3.0 | 15 | IG | $n$ | 8 | 0.953 | 0.279 |
| 3.5 | 18 | IG | $\log_2 n$ | $\infty$ | 0.954 | 0.278 |
| 3.0 | 13 | IG | $n$ | 8 | 0.953 | 0.272 |
| 3.0 | 18 | Gini | $\log_2 n$ | $\infty$ | 0.953 | 0.268 |
| 3.0 | 19 | IG | $\sqrt{n}$ | 8 | 0.954 | 0.267 |
| 3.0 | 18 | Gini | 1 | $\infty$ | 0.954 | 0.266 |

cannot be said when we ignore ambiguous samples, as the MCC scores for the top 2 MILP models far surpass any of the models generated by the KNN, SVM and RDF classifiers, as shown by table 3. However, table 3 also shows that the percentage of samples classified as ambiguous is very high, decreasing the usefulness of such models.

The results obtained in tables 4-6 are very useful in any case, as they provide a potential indication of the non-optimal performance shown by the MILP models. The SVM models present the worst MCC scores of the three classifiers, even in a high dimension space. Given that for a high dimensional space, the distance between the points becomes less meaningful, which in turn makes the choice of a kernel become less significative, and creating a linear decision boundary should also become easier, it is interesting to observe that the linear kernel shows, in general, worse results that non-linear polynomial kernel. Adding to this, the gaussian and sigmoid kernels show a complete inability for properly separating the classes, for any of the evaluated values of C and $\gamma$, probably indicating that the models generated with these kernels are overfitting to the data when compared to the KNN and RDF generated models.

On another note, the models generated by the KNN and RDF classifiers show equivalent performance, with top KNN models showing slightly better performance than the RDF ones. This might be an indication that the data distribution over the sample space might approach a slightly hyperspherical form, but does not explain why the usage of the gaussian kernel for the SVM models is unable to train proper models, beyond what was already stated regarding the potential overfitting problem. Beyond this, the fact the best RDF models show equivalent performance to the best KNN models and improved performance when compared to the SVM models clearly demonstrates the non-linear data distribution in the TCGA-PRAD dataset, which explains the worse performance inherent to the MILP models.

Table 7.: Selected features for each TCGA dataset when training on all samples. With the exception of the $\epsilon$ parameter, the values for the remaining parameters are $\psi = 3.5$, $\lambda_1 = 0.01$, $\lambda_2 = 1.00$, and $\lambda_3 = 1.00$.

| Dataset | $\epsilon$ | Weight | Relevance | Feature |
|---|---|---|---|---|
| TCGA-BRCA | 0.5 | 2.00 | 0.117 | ENSG00000105974 |
| | | -1.00 | 0.121 | ENSG00000198712 |
| | 1.0 | 2.00 | 0.014 | ENSG00000105974 |
| | | -1.00 | 0.015 | ENSG00000198712 |
| TCGA-COAD | 0.5 | 1.00 | 0.177 | ENSG00000007306 |
| | | -2.00 | 0.236 | ENSG00000196611 |
| | 1.0 | -1.00 | 0.100 | ENSG00000102265 |
| | | 1.00 | 0.333 | ENSG00000197273 |
| TCGA-HNSC | 0.5 | -2.00 | 0.277 | ENSG00000122861 |
| | | 1.00 | 0.208 | ENSG00000198712 |
| | 1.0 | -2.00 | 0.077 | ENSG00000122861 |
| | | 1.00 | 0.043 | ENSG00000198712 |
| TCGA-LUAD | 0.5 | -2.00 | 0.267 | ENSG00000119888 |
| | | 1.00 | 0.204 | ENSG00000198712 |
| | 1.0 | -2.00 | 0.071 | ENSG00000119888 |
| | | 1.00 | 0.042 | ENSG00000198712 |
| TCGA-PRAD | 0.5 | -2.00 | 0.500 | ENSG00000088986 |
| | | 1.00 | 0.147 | ENSG00000120885 |
| | 1.0 | -2.00 | 0.250 | ENSG00000088986 |
| | | 1.00 | 0.022 | ENSG00000120885 |

## 4.3 FEATURE CONTEXTUALIZATION ANALYSIS

After the grid search performed for the MILP model developed in this work for the TCGA-PRAD dataset in the section above, two sets of parameters' values were discovered that provided the best performance, when ignoring samples classified as ambiguous. This occurs when $\psi = 3.5$, $\lambda_1 = 0.01$, $\lambda_2 = 1.00$, $\lambda_3 = 1.00$, and $\epsilon \in \{0.5, 1.0\}$. These parameters' values were then used when fitting the model to each dataset. The selected features as potential biomarkers are shown on table 7.

As it can be seen, the number of selected features as potential biomarkers is constant, with only two selected per dataset and per parameter combination. This is most likely a consequence of the optimal combination of $\lambda_{1-3}$ favouring one hundred times more the gene-disease associations and the minimization of the number of selected biomarker components in the objective function when compared to the maximization of the class separation component, while, at the same time, the restrictions represented by equations 10 and 11 enforce the selection of at least one feature associated to each class.

Adding to this, the smoothing factor $\epsilon$ does not appear to exert a lot of influence on the genes that are selected, as these only vary with it when training on the TCGA-COAD

Table 8.: Summary of the biological significance of each identified potential biomarker. MMP stands for matrix metalloproteinase.

| Feature | Gene Symbol | Biological Role |
|---|---|---|
| ENSG00000007306 | CEACAM7 | Regulates cellular differentiation |
| ENSG00000088986 | DYNLL1 | Associated with mitosis, organelle transport and nuclear migration |
| ENSG00000102265 | TIMP1 | Inhibitor of MMPs, helping regulate cellular migration and prevent apoptosis |
| ENSG00000105974 | CAV1 | May be responsible for preventing apoptosis and facilitating the development of drug resistance and metastasis |
| ENSG00000119888 | EPCAM | Associated with cellular proliferation, migration and mitosis |
| ENSG00000120885 | CLU | Stress-induced cytoprotective role |
| ENSG00000122861 | PLAU | Participates in the activation of MMPs, improving tissue regeneration and immune response |
| ENSG00000196611 | MMP1 | MMP responsible for tissue regeneration |
| ENSG00000197273 | GUCA2A | Activator of GUCY2C, a tumour suppressor |
| ENSG00000198712 | MT-CO2 | Component of Cytochrome C Oxidase, responsible for catalysing the reduction of $O_2$ to $H_2O$. |

dataset. In any case, it is important to analyse the biological significance of the chosen potential biomarkers. Table 8 summarizes the biological role of each individual predicted potential biomarker which will be analysed more in-depth below.

***Breast Adenocarcinoma***

The ENSEMBL identifier ENSG00000105974 stands for the CAV1 gene, which encodes the Caveolin-1 protein and was uniquely selected for the breast cancer dataset (TCGA-BRCA). There have been many studies over the years that try to identify CAV1's role in cancer. It has been found that in breast, colon, and lung cancer, as well as many others, CAV1 appears to be overexpressed, acting as a tumour promoter or suppressor depending on the type of tumour and development stage [41, 74]. On one hand, high expression of CAV1 has been reported to inhibit apoptosis, facilitating the development of drug resistance and metastasis [38, 50, 87, 119, 131]. On the other hand, low expression of CAV1 has shown to favour tumour development [90, 98, 138]. Its link to cancer is hard to deny, leading it to be considered a clinical biomarker for cancer [83, 110].

Adding to this, ENSG00000198712 has been identified by several of our models as a potential biomarker for some of the diseases. It represents MT-CO2 gene, which encodes the subunit 2 of the Cytochrome C Oxidase (CcO) protein complex, a component of the respiratory chain in the mitochondria responsible for catalysing the reduction of oxygen to water. This complex has been linked to the development of several types of cancer,

with gene silencing and general loss of activity resulting in a metabolic shift to glycolysis, invasiveness [6, 18, 39, 40], and a disruption in the regulation of the electron transport, oxidative phosphorylation and the production of reactive oxygen species (ROS) [1, 101, 126]. It makes sense that multiple of our models have selected it, as these characteristics are a staple of most cancer types.

### *Colon Adenocarcinoma*

Two sets of completely unique biomarkers were selected for the colon adenocarcinoma dataset (TCGA-COAD). The first set includes ENSG00000007306 and ENSG00000196611. ENSG00000007306, standing for the CEACAM7 gene, encodes the carcinoembryonic antigen-related cell adhesion molecule 7 protein. This protein is responsible for regulating normal cellular differentiation. High expression of this gene has been shown to be associated with the early onset of colorectal [102] and gastric carcinomas [142], while low expression has been related with adenomas and hyperplastic polyps [118, 142]. Furthermore, expression changes in the family of genes to which this gene belongs, the carcinoembryonic antigen (CEA) family, have been related to the emergence of neoplasms and metastases [45, 102].

Adding to this, ENSG00000196611 represents the MMP1 gene, which encodes the interstitial collagenase 1 or matrix metallopeptidase 1 protein. MMPs in general are responsible for remodelling the extracellular matrix in a multitude of physical processes, including the development and regeneration of tissue [109], and have been implicated in several pathological processes, one of which is cancer [27]. MMP1 in particular cleaves collagen types I, II, III, VII and X [3] and is expressed by metastatic tumor cells at the site of invasion, signifying its importance in this process [20]. It has also been identified as a prognostic marker for various types of cancer, be it breast, colorectal or esophageal cancer [58, 80, 81].

The other set of unique potential biomarkers identified for the colon adenocarcinoma dataset (TCGA-COAD) includes ENSG00000102265 and ENSG00000197273. The ENSEMBL identifier ENSG00000102265 relates to the TIMP1 gene, encoding the Metalloproteinase Inhibitor 1 protein. This protein is part of a family of Tissue Inhibitor of Metalloproteinases with four members, this one acting as an inhibitor of the proteolytic activity of MMPs [12], as well as possessing an important role regulating cellular proliferation and anti-apoptotic function [10, 82, 134, 135], resulting in an increase in drug resistance capabilities in tumour cells [36, 120]. Adding to this, links to several types of cancer have been discovered, such as breast cancer [135], colon cancer [108], and gastric cancer [129], leading it to be considered as a potential prognostic biomarker for these. ENSG00000197273 represents the GUCA2A gene, which encodes the Guanylin protein. This is a hormone acting as endogenous activator of GUCY2C, one of several transmembrane guanylyl cyclase receptors [11]. GUCY2C has been found to act as tumour suppressor, helping prevent the development of colorectal cancer [22, 30, 133]. As such, any event which disrupts normal GUCY2C signalling should

increase the risk for the development of colorectal cancer, such as the loss of expression of the Guanylin hormone.

### *Head and Neck Squamous Cell Carcinoma*

Specifically associated with the head and neck squamous cell carcinoma dataset (TCGA-HNSC), ENSG00000122861 has been identified as a potential unique biomarker for this disease. This gene, PLAU, encodes a serine protease named urokinase-type plasminogen activator (uPA) responsible for converting inactive plasminogen to active plasmin, which in turn helps in degrading the extracellular matrix and catalysing the activation of MMPs [15]. As such, it is implicated in a series of physiological processes which include tissue regeneration and immune response, but also pathological processes which involve tumour cell proliferation, adhesion and migration [8, 21, 24, 28, 91, 105]. Specifically, uPA appears to have an important role in promoting tumour angiogenesis [16, 19, 21, 24, 28], with high levels of expression being associated with increased neoplasm aggressiveness and poor clinical outcome in many different types of cancer [8, 13, 26, 21, 25, 28, 44, 121, 122, 128].

### *Lung Adenocarcinoma*

The lung adenocarcinoma dataset (TCGA-LUAD) also saw the selection of a potential unique biomarker, ENSG00000119888. This gene goes by the gene identifier EPCAM and encodes the epithelial cell adhesion molecule, a transmembrane glycoprotein normally expressed in epithelial cells and has been proven to be involved in biological processes such as cell proliferation and migration, and mitogenic signal transduction [68, 79]. It has also been verified that high expression levels of EPCAM are associated with epithelial cell cancers together with as breast, colon, and head and neck squamous cell carcinoma, as well as being implicated in angiogenesis [69, 86, 89, 117]. This led to it being postulated as a potential prognostic biomarker [57, 143].

### *Prostate Adenocarcinoma*

Finally, a set of unique potential biomarkers were selected for the prostate adenocarcinoma dataset (TCGA-PRAD), ENSG00000088986 and ENSG00000120885. ENSG00000088986 constitutes the DYNLL1 gene, encoding the dynein light chain 1 (DLC1), a component of the cytoplasmatic dynein motor complex [48] which has been implicated in cytoplasmatic organelle transport, mitosis and nuclear migration [51, 124]. It is a substrate of the signaling kinase p21-activated kinase 1 (PAK1), together with whom it has been discovered that they help promote anchorage-independent growth of breast cancer cells [123] when DLC1 is over-expressed. Abnormal expression of DLC1 has also been found to be implicated in increased sensitivity of tumour cells to estrogen [95]. ENSG00000120885 represents the CLU

gene, which encodes the clusterin protein. Clusterin is a chaperone with a cytoprotective role induced by stress conditions, usually upregulated in many types of cancer including bladder, breast, colon, lung and prostate cancer and conferring resistance to typical cancer treatments [37, 62, 65, 76, 97, 99, 113, 136, 137]. Adding to this, its overexpression has been reported to reliably correlate with tumour recurrence [76], and with tumour size and stage progression [47].

As the scientific community shows, the status as a potential biomarker for most selected genes is not unique to any single disease. Each of these have shown some type of association with several pathological processes, decreasing their individual desirability as biomarkers. The exception to these are ENSG00000197273 and ENSG00000088986, representing guanylin and DLC1, respectively. It is interesting to consider that these have not been directly implicated in the development of tumorous behaviour, with the focus being more directed toward their respective receptors, GUCY2C and PAK1, at least in the context of the diseases for which they were selected as potential biomarkers by this work. As such, a better clinical analysis of their behaviour in pathological conditions, such as cancer, may provide a greater understanding of their role in the development and progression of cancer.

# 5

# CONCLUSION

A general pipeline was proposed in this work for identifying novel biomarker profiles for the purpose of minimizing the burden of diagnosing disease conditions, with a focus on cancer states. The basis for the proposed method is a Mixed-Integer Linear Programming (MILP) model, based on the one initially developed by Santiago *et al.* [100], but integrating simultaneously gene expression data and gene-disease associations, where the goal is to discover an optimal marker combination capable of maximizing the separation of disease status in any group of considered samples, while providing a white box model where the importance of each selected biomarker can be easily analysed and understood. Additionally, the developed model provides a ternary (diseased, control or ambiguous) instead of a typical binary classification system, allowing for more flexibility and potential better generalization in any predicted biomarker profile.

However, some considerations have to be made. First, while the linear nature of the optimization model allows better understanding of the reasons behind the selected biomarker profiles, it also greatly limits the discovery of potential relationships between genes and the disease status with a non-linear nature, which are a staple of most biological processes. Second, the way the MILP model is structured also impedes finding the potential optimal set of genes for separating the class samples. Specifically, constraints 12 and 13 of the optimization model defined in section 3.4, while allowing the model to classify samples as ambiguous, does not technically enforce efficient class separation. This happens because any potential generated model will consider as sufficient classifying every sample as ambiguous, even if the objective function 6 tries to maximize the absolute score for each sample. This either results in low overall MCC scores, as shown in table 2, or a large percentage of samples classified as ambiguous, as shown in table 3.

In any case, for the sets of evaluated parameters that show the best performance in table 3, these perform better than any of the models generated through the K-Nearest Neighbours (KNN), Support Vector Machines (SVM) and Random Decision Forests (RDF) classifiers, at the cost of the majority of samples being classified as ambiguous. Adding to this, although most of the selected biomarkers for each considered disease in table 7 had already shown some relation to pathological processes, especially cancer, two of the genes,

ENSG00000197273 and ENSG00000088986, have emerged as potential novel biomarkers as their connection to cancer development does not appear to have been a focus of the scientific community.

Overall, the method developed in this work for performing biomarker discovery, while not providing the expected results of effective disease status classification, allowed uncovering two potential novel biomarkers. Furthermore, the integration of gene-disease associations from the literature appeared to improve the quality of the generated models as these, in general, presented the highest MCC scores. Further development of this approach would focus on improving and fine-tuning the model's constraints, especially in the context of minimizing the number of ambiguous samples, and improving how gene-disease associations are considered in the model, perhaps by also integrating information regarding other diseases beyond the one being considered. It should also be interesting to see the application of this method on other non-tumorous pathological conditions to better evaluate its generalizability.

Finally, it should not be forgotten that the task of developing successful computational prediction tools of any kind is, at the very least, a difficult one, and perhaps more so in a biological context. In any case, this approach stands as a stepping stone for the future development of better computational biomarker discovery tools, integrating both gene expression and gene-disease bibliographical association data.

# 6

## BIBLIOGRAPHY


[1] R. Acín-Pérez, P. Fernández-Silva, M. L. Peleato, A. Pérez-Martos, and J. A. Enriquez. Respiratory Active Mitochondrial Supercomplexes. *Molecular Cell*, 32(4):529–539, nov 2008. ISSN 10972765. doi: 10.1016/j.molcel.2008.10.021. URL http://www.ncbi.nlm.nih.gov/pubmed/19026783http://linkinghub.elsevier.com/retrieve/pii/S1097276508007624.

[2] F. Al-Shahrour, P. Minguez, J. Tárraga, I. Medina, E. Alloza, D. Montaner, and J. Dopazo. FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic acids research*, 35(Web Server issue):W91–6, jul 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm260. URL http://nar.oxfordjournals.org/content/35/suppl{_}2/W91.

[3] R. Ala-aho and V.-M. Kähäri. Collagenases in cancer. *Biochimie*, 87(3-4):273–286, mar 2005. ISSN 0300-9084. doi: 10.1016/J.BIOCHI.2004.12.009. URL https://www.sciencedirect.com/science/article/pii/S0300908404002664.

[4] B. T. F. Alako, A. Veldhoven, S. van Baal, R. Jelier, S. Verhoeven, T. Rullmann, J. Polman, and G. Jenster. CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC bioinformatics*, 6(1):51, jan 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-51. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-51.

[5] M. Alexander, J. B. Burch, S. E. Steck, C.-F. Chen, T. G. Hurley, P. Cavicchia, N. Shivappa, J. Guess, H. Zhang, S. D. Youngstedt, K. E. Creek, S. Lloyd, K. Jones, and J. R. H?bert. Case-control study of candidate gene methylation and adenomatous polyp formation. *International Journal of Colorectal Disease*, 32(2):183–192, feb 2017. ISSN 0179-1958. doi: 10.1007/s00384-016-2688-1. URL http://www.ncbi.nlm.nih.gov/pubmed/27771773http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5288296http://link.springer.com/10.1007/s00384-016-2688-1.

[6] G. Amuthan, G. Biswas, H. K. Ananadatheerthavarada, C. Vijayasarathy, H. M. Shephard, and N. G. Avadhani. Mitochondrial stress-induced calcium signaling, phenotypic changes and invasive behavior in human lung carcinoma A549 cells. *Oncogene*, 21(51):7839–7849, nov 2002. ISSN 0950-9232. doi: 10.1038/sj.onc.1205983. URL http://www.ncbi.nlm.nih.gov/pubmed/12420221http://www.nature.com/articles/1205983.

[7] S. Ananiadou, D. B. Kell, and J.-i. Tsujii. Text mining and its potential applications in systems biology. *Trends in biotechnology*, 24(12):571–9, dec 2006. ISSN 0167-7799. doi: 10.1016/j.tibtech.2006.10.002. URL http://www.cell.com/article/S0167779906002423/fulltext.

[8] P. A. Andreasen, L. Kjøller, L. Christensen, and M. J. Duffy. The urokinase-type plasminogen activator system in cancer metastasis: a review. *International journal of cancer*, 72(1):1–22, jul 1997. ISSN 0020-7136. URL http://www.ncbi.nlm.nih.gov/pubmed/9212216.

[9] F. Azuaje and J. Dopazo. *Data analysis and visualization in genomics and proteomics*. Wiley, 2005. ISBN 9780470094396.

[10] W. Bao, H.-J. Fu, L.-T. Jia, Y. Zhang, W. Li, B.-Q. Jin, L.-B. Yao, S.-Y. Chen, and A.-G. Yang. HER2-mediated upregulation of MMP-1 is involved in gastric cancer cell invasion. *Archives of Biochemistry and Biophysics*, 499(1-2):49–55, jul 2010. ISSN 00039861. doi: 10.1016/j.abb.2010.05.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0003986110001761.

[11] N. Basu, N. Arshad, and S. S. Visweswariah. Receptor guanylyl cyclase C (GC-C): regulation and signal transduction. *Molecular and Cellular Biochemistry*, 334(1-2):67–80, jan 2010. ISSN 0300-8177. doi: 10.1007/s11010-009-0324-x. URL http://www.ncbi.nlm.nih.gov/pubmed/19960363http://link.springer.com/10.1007/s11010-009-0324-x.

[12] J. Batra, J. Robinson, A. S. Soares, A. P. Fields, D. C. Radisky, and E. S. Radisky. Matrix Metalloproteinase-10 (MMP-10) Interaction with Tissue Inhibitors of Metalloproteinases TIMP-1 and TIMP-2. *Journal of Biological Chemistry*, 287(19):15935–15946, may 2012. ISSN 0021-9258. doi: 10.1074/jbc.M112.341156. URL http://www.ncbi.nlm.nih.gov/pubmed/22427646http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3346077http://www.jbc.org/lookup/doi/10.1074/jbc.M112.341156.

[13] M. C. Béné, G. Castoldi, W. Knapp, G. M. Rigolin, L. Escribano, P. Lemez, W.-D. Ludwig, E. Matutes, A. Orfao, F. Lanza, M. van't Veer, and EGIL, European

Group on Immunological Classification of Leukemias. CD87 (urokinase-type plasminogen activator receptor), function and pathology in hematological disorders: a review. *Leukemia*, 18(3):394–400, mar 2004. ISSN 0887-6924. doi: 10.1038/sj.leu.2403250. URL http://www.ncbi.nlm.nih.gov/pubmed/14671631http://www.nature.com/articles/2403250.

[14] M. F. Berger, J. Z. Levin, K. Vijayendran, A. Sivachenko, X. Adiconis, J. Maguire, L. A. Johnson, J. Robinson, R. G. Verhaak, C. Sougnez, R. C. Onofrio, L. Ziaugra, K. Cibulskis, E. Laine, J. Barretina, W. Winckler, D. E. Fisher, G. Getz, M. Meyerson, D. B. Jaffe, S. B. Gabriel, E. S. Lander, R. Dummer, A. Gnirke, C. Nusbaum, and L. A. Garraway. Integrative analysis of the melanoma transcriptome. *Genome Research*, 20(4):413–427, apr 2010. ISSN 1088-9051. doi: 10.1101/gr.103697.109. URL http://www.ncbi.nlm.nih.gov/pubmed/20179022http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2847744http://genome.cshlp.org/cgi/doi/10.1101/gr.103697.109.

[15] J. A. Bezerra, A. R. Currier, H. Melin-Aldana, G. Sabla, T. H. Bugge, K. W. Kombrinck, and J. L. Degen. Plasminogen Activators Direct Reorganization of the Liver Lobule after Acute Injury. *The American Journal of Pathology*, 158(3):921–929, mar 2001. ISSN 00029440. doi: 10.1016/S0002-9440(10)64039-4. URL http://www.ncbi.nlm.nih.gov/pubmed/11238040http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1850368http://linkinghub.elsevier.com/retrieve/pii/S0002944010640394.

[16] B. R. Binder, J. Mihaly, and G. W. Prager. uPAR-uPA-PAI-1 interactions and signaling: a vascular biologist's view. *Thrombosis and haemostasis*, 97(3):336–42, mar 2007. ISSN 0340-6245. URL http://www.ncbi.nlm.nih.gov/pubmed/17334498.

[17] Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology and therapeutics*, 69(3):89–95, mar 2001. ISSN 0009-9236. doi: 10.1067/mcp.2001.113989. URL http://www.ncbi.nlm.nih.gov/pubmed/11240971.

[18] G. Biswas, O. A. Adebanjo, B. D. Freedman, H. K. Anandatheerthavarada, C. Vijayasarathy, M. Zaidi, M. Kotlikoff, and N. G. Avadhani. Retrograde Ca2+ signaling in C2C12 skeletal myocytes in response to mitochondrial genetic and metabolic stress: a novel mode of inter-organelle crosstalk. *The EMBO Journal*, 18(3):522–533, feb 1999. ISSN 14602075. doi: 10.1093/emboj/18.3.522. URL http://www.ncbi.nlm.nih.gov/pubmed/9927412http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1171145http://emboj.embopress.org/cgi/doi/10.1093/emboj/18.3.522.

[19] F. Blasi and P. Carmeliet. uPAR: a versatile signalling orchestrator. *Nature Reviews Molecular Cell Biology*, 3(12):932–943, dec 2002. ISSN 1471-0072. doi: 10.1038/nrm977. URL http://www.ncbi.nlm.nih.gov/pubmed/12461559http://www.nature.com/articles/nrm977.

[20] C. E. Brinckerhoff, J. L. Rutter, and U. Benbow. Interstitial collagenases as markers of tumor progression. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 6(12):4823–30, dec 2000. ISSN 1078-0432. URL http://www.ncbi.nlm.nih.gov/pubmed/11156241.

[21] P. F. M. Choong and A. P. W. Nadesapillai. Urokinase plasminogen activator system: a multifunctional role in tumor progression and metastasis. *Clinical orthopaedics and related research*, (415 Suppl):S46–58, oct 2003. ISSN 0009-921X. doi: 10.1097/01.blo.0000093845.72468.bd. URL http://www.ncbi.nlm.nih.gov/pubmed/14600592.

[22] M. B. Cohen, J. A. Hawkins, and D. P. Witte. Guanylin mRNA expression in human intestine and colorectal adenocarcinoma. *Laboratory investigation; a journal of technical methods and pathology*, 78(1):101–8, jan 1998. ISSN 0023-6837. URL http://www.ncbi.nlm.nih.gov/pubmed/9461126.

[23] P. Corbett and A. Copestake. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC bioinformatics*, 9 Suppl 11(11):S4, jan 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-S11-S4. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-S11-S4.

[24] K. Danø, N. Behrendt, G. Høyer-Hansen, M. Johnsen, L. R. Lund, M. Ploug, and J. Rømer. Plasminogen activation and cancer. *Thrombosis and Haemostasis*, 93(4):676–81, mar 2005. ISSN 0340-6245. doi: 10.1160/TH05-01-0054. URL http://www.ncbi.nlm.nih.gov/pubmed/15841311http://www.schattauer.de/index.php?id=1214{&}doi=10.1160/TH05-01-0054{&}no{_}cache=1.

[25] K. Dass, A. Ahmad, A. S. Azmi, S. H. Sarkar, and F. H. Sarkar. Evolving role of uPA/uPAR system in human cancers. *Cancer Treatment Reviews*, 34(2):122–136, apr 2008. ISSN 03057372. doi: 10.1016/j.ctrv.2007.10.005. URL http://www.ncbi.nlm.nih.gov/pubmed/18162327http://linkinghub.elsevier.com/retrieve/pii/S0305737207001818.

[26] C. E. de Bock and Y. Wang. Clinical significance of urokinase-type plasminogen activator receptor (uPAR) expression in cancer. *Medicinal Research Reviews*, 24(1):13–39, jan 2004. ISSN 0198-6325. doi: 10.1002/med.10054. URL http://www.ncbi.nlm.nih.gov/pubmed/14595671http://doi.wiley.com/10.1002/med.10054.

[27] E. I. Deryugina and J. P. Quigley. Matrix metalloproteinases and tumor metastasis. *Cancer and Metastasis Reviews*, 25(1):9–34, mar 2006. ISSN 0167-7659. doi: 10.1007/s10555-006-7886-9. URL http://www.ncbi.nlm.nih.gov/pubmed/16680569http://link.springer.com/10.1007/s10555-006-7886-9.

[28] M. J. Duffy. The urokinase plasminogen activator system: role in malignancy. *Current pharmaceutical design*, 10(1):39–49, 2004. ISSN 1381-6128. URL http://www.ncbi.nlm.nih.gov/pubmed/14754404.

[29] J. A. Evans and A. Rzhetsky. Advancing science through mining libraries, ontologies, and communities. *The Journal of biological chemistry*, 286(27):23659–66, jul 2011. ISSN 1083-351X. doi: 10.1074/jbc.R110.176370. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3129146{&}tool=pmcentrez{&}rendertype=abstract.

[30] A. Fajardo, G. Piazza, and H. Tinsley. The Role of Cyclic Nucleotide Signaling Pathways in Cancer: Targets for Prevention and Treatment. *Cancers*, 6(1):436–458, feb 2014. ISSN 2072-6694. doi: 10.3390/cancers6010436. URL http://www.ncbi.nlm.nih.gov/pubmed/24577242http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3980602http://www.mdpi.com/2072-6694/6/1/436.

[31] M. E. Falagas, K. P. Giannopoulou, E. A. Issaris, and A. Spanos. World databases of summaries of articles in the biomedical fields. *Archives of internal medicine*, 167 (11):1204–6, jun 2007. ISSN 0003-9926. doi: 10.1001/archinte.167.11.1204. URL http://archinte.jamanetwork.com/article.aspx?articleid=412558.

[32] W. W. Fleuren, E. J. Toonen, S. Verhoeven, R. Frijters, T. Hulsen, T. Rullmann, R. van Schaik, J. de Vlieg, and W. Alkema. Identification of new biomarker candidates for glucocorticoid induced insulin resistance using literature mining. *BioData mining*, 6(1):2, jan 2013. ISSN 1756-0381. doi: 10.1186/1756-0381-6-2. URL http://biodatamining.biomedcentral.com/articles/10.1186/1756-0381-6-2.

[33] W. W. M. Fleuren and W. Alkema. Application of text mining in the biomedical domain. *Methods (San Diego, Calif.)*, 74:97–106, mar 2015. ISSN 1095-9130. doi: 10.1016/j.ymeth.2015.01.015. URL http://www.ncbi.nlm.nih.gov/pubmed/25641519.

[34] W. W. M. Fleuren, S. Verhoeven, R. Frijters, B. Heupers, J. Polman, R. van Schaik, J. de Vlieg, and W. Alkema. CoPub update: CoPub 5.0 a text mining system to answer biological questions. *Nucleic acids research*, 39(Web Server issue):W450–4, jul 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr310. URL http://nar.oxfordjournals.org/content/39/suppl{_}2/W450.

[35] R. Frijters, S. Verhoeven, W. Alkema, R. van Schaik, and J. Polman. Literature-based compound profiling: application to toxicogenomics. *Pharmacogenomics*, 8(11):1521–

34, nov 2007. ISSN 1744-8042. doi: 10.2217/14622416.8.11.1521. URL http://www.futuremedicine.com/doi/abs/10.2217/14622416.8.11.1521.

[36] Z. Fu, J. Lv, C. Ma, D. Yang, and T. Wang. Tissue inhibitor of metalloproteinase-1 decreased chemosensitivity of MDA-435 breast cancer cells to chemotherapeutic drugs through the PI3K/AKT/NF-B pathway. *Biomedicine & Pharmacotherapy*, 65(3):163–167, jun 2011. ISSN 07533322. doi: 10.1016/j.biopha.2011.02.004. URL http://www.ncbi.nlm.nih.gov/pubmed/21684102http://linkinghub.elsevier.com/retrieve/pii/S0753332211000229.

[37] M. E. Gleave, H. Miyake, T. Zellweger, K. Chi, L. July, C. Nelson, and P. Rennie. Use of antisense oligonucleotides targeting the antiapoptotic gene, clusterin/testosterone-repressed prostate message 2, to enhance androgen sensitivity and chemosensitivity in prostate cancer. *Urology*, 58(2 Suppl 1):39–49, aug 2001. ISSN 1527-9995. URL http://www.ncbi.nlm.nih.gov/pubmed/11502446.

[38] J. G. Goetz, P. Lajoie, S. M. Wiseman, and I. R. Nabi. Caveolin-1 in tumor progression: the good, the bad and the ugly. *Cancer and Metastasis Reviews*, 27(4):715–735, dec 2008. ISSN 0167-7659. doi: 10.1007/s10555-008-9160-9. URL http://www.ncbi.nlm.nih.gov/pubmed/18506396http://link.springer.com/10.1007/s10555-008-9160-9.

[39] M. Guha, H. Pan, J.-K. Fang, and N. G. Avadhani. Heterogeneous Nuclear Ribonucleoprotein A2 Is a Common Transcriptional Coactivator in the Nuclear Transcription Response to Mitochondrial Respiratory Stress. *Molecular Biology of the Cell*, 20 (18):4107–4119, sep 2009. ISSN 1059-1524. doi: 10.1091/mbc.e09-04-0296. URL http://www.ncbi.nlm.nih.gov/pubmed/19641020http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2743628http://www.molbiolcell.org/doi/10.1091/mbc.e09-04-0296.

[40] M. Guha, J.-K. Fang, R. Monks, M. J. Birnbaum, and N. G. Avadhani. Activation of Akt is essential for the propagation of mitochondrial respiratory stress signaling and activation of the transcriptional coactivator heterogeneous ribonucleoprotein A2. *Molecular biology of the cell*, 21(20):3578–89, oct 2010. ISSN 1939-4586. doi: 10.1091/mbc.E10-03-0192. URL http://www.ncbi.nlm.nih.gov/pubmed/20719961http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2954122.

[41] R. Gupta, C. Toufaily, and B. Annabi. Caveolin and cavin family members: Dual roles in cancer. *Biochimie*, 107:188–202, dec 2014. ISSN 03009084. doi: 10.1016/j.biochi.2014.09.010. URL http://www.ncbi.nlm.nih.gov/pubmed/25241255https://linkinghub.elsevier.com/retrieve/pii/S0300908414002594.

[42] M. S. Habib and J. Kalita. Scalable biomedical Named Entity Recognition: investigation of a database-supported SVM approach. *International journal of bioinformatics research and applications*, 6(2):191–208, jan 2010. ISSN 1744-5485. doi: 10.1504/IJBRA.2010.032121. URL http://www.ncbi.nlm.nih.gov/pubmed/20223740.

[43] A. S. Haqqani, J. Kelly, E. Baumann, R. F. Haseloff, I. E. Blasig, and D. B. Stanimirovic. Protein markers of ischemic insult in brain endothelial cells identified using 2D gel electrophoresis and ICAT-based quantitative proteomics. *Journal of proteome research*, 6(1):226–39, jan 2007. ISSN 1535-3893. doi: 10.1021/pr0603811. URL http://dx.doi.org/10.1021/pr0603811.

[44] N. Harbeck, M. Schmitt, S. Paepke, H. Allgayer, and R. E. Kates. Tumor-Associated Proteolytic Factors uPA and PAI-1: Critical Appraisal of Their Clinical Relevance in Breast Cancer and Their Integration into Decision-Support Algorithms. *Critical Reviews in Clinical Laboratory Sciences*, 44(2):179–201, jan 2007. ISSN 1040-8363. doi: 10.1080/10408360601040970. URL http://www.ncbi.nlm.nih.gov/pubmed/17364692http://www.tandfonline.com/doi/full/10.1080/10408360601040970.

[45] J. Hashino, Y. Fukuda, S. Oikawa, H. Nakazato, and T. Nakanishi. Metastatic potential of human colorectal carcinoma SW1222 cells transfected with cDNA encoding carcinoembryonic antigen. *Clinical & experimental metastasis*, 12(4):324–8, jul 1994. ISSN 0262-0898. URL http://www.ncbi.nlm.nih.gov/pubmed/8039306.

[46] M. A. Hearst. Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* -, pages 3–10, Morristown, NJ, USA, jun 1999. Association for Computational Linguistics. ISBN 1558606093. doi: 10.3115/1034678.1034679. URL http://dl.acm.org/citation.cfm?id=1034678.1034679.

[47] E. B. Henderson-Jackson, A. Nasir, D.-T. Chen, P. Nandyala, J. Djeu, J. Strosberg, L. Kvols, and D. Coppola. Cytoplasmic Clusterin Expression Correlates With Pancreatic Neuroendocrine Tumor Size and Pathological Stage. *Pancreas*, 42(6):967–970, aug 2013. ISSN 0885-3177. doi: 10.1097/MPA.0b013e318293734b. URL http://www.ncbi.nlm.nih.gov/pubmed/23770713http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4644941http://insights.ovid.com/crossref?an=00006676-201308000-00011.

[48] N. Hirokawa, Y. Noda, and Y. Okada. Kinesin and dynein superfamily proteins in organelle transport and cell division. *Current opinion in cell biology*, 10(1):60–73, feb 1998. ISSN 0955-0674. URL http://www.ncbi.nlm.nih.gov/pubmed/9484596.

[49] L. Hirschman, G. A. P. C. Burns, M. Krallinger, C. Arighi, K. B. Cohen, A. Valencia, C. H. Wu, A. Chatr-Aryamontri, K. G. Dowell, E. Huala, A. Lourenço, R. Nash, A.-L. Veuthey, T. Wiegers, and A. G. Winter. Text mining for the biocuration workflow. *Database : the journal of biological databases and curation*, 2012(0):bas020, jan 2012. ISSN 1758-0463. doi: 10.1093/database/bas020. URL http://database.oxfordjournals.org/content/2012/bas020.

[50] C.-C. Ho, S.-H. Kuo, P.-H. Huang, H.-Y. Huang, C.-H. Yang, and P.-C. Yang. Caveolin-1 expression is significantly associated with drug resistance and poor prognosis in advanced non-small cell lung cancer patients treated with gemcitabine-based chemotherapy. *Lung Cancer*, 59(1):105–110, jan 2008. ISSN 01695002. doi: 10.1016/j.lungcan.2007.07.024. URL http://www.ncbi.nlm.nih.gov/pubmed/17850918http://linkinghub.elsevier.com/retrieve/pii/S016950020700431X.

[51] E. L. F. Holzbaur and R. B. Vallee. Dyneins: Molecular Structure and Cellular Function. *Annual Review of Cell Biology*, 10(1):339–372, nov 1994. ISSN 0743-4634. doi: 10.1146/annurev.cb.10.110194.002011. URL http://www.ncbi.nlm.nih.gov/pubmed/7888180http://www.annualreviews.org/doi/10.1146/annurev.cb.10.110194.002011.

[52] Z.-X. Huang, H.-Y. Tian, Z.-F. Hu, Y.-B. Zhou, J. Zhao, and K.-T. Yao. GenCLiP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords. *BMC bioinformatics*, 9(1):308, jan 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-308. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-308.

[53] K. M. Izet Masic. On-line Biomedical Databasesthe Best Source for Quick Search of the Scientific Information in the Biomedicine. *Acta Informatica Medica*, 20(2):72–84, 2012. URL http://www.scopemed.org/?mno=20169.

[54] D. G. Jamieson, M. Gerner, F. Sarafraz, G. Nenadic, and D. L. Robertson. Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database. *Database : the journal of biological databases and curation*, 2012 (0):bas023, jan 2012. ISSN 1758-0463. doi: 10.1093/database/bas023. URL http://database.oxfordjournals.org/content/2012/bas023.

[55] K. Jensen, G. Panagiotou, and I. Kouskoumvekaki. Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level. *PLoS computational biology*, 10(1):e1003432, jan 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003432. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003432.

[56] L. J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews. Genetics*, 7(2):119–29, feb 2006. ISSN 1471-0056. doi: 10.1038/nrg1768. URL http://dx.doi.org/10.1038/nrg1768.

[57] J. Ji, T. Yamashita, A. Budhu, M. Forgues, H.-L. Jia, C. Li, C. Deng, E. Wauthier, L. M. Reid, Q.-H. Ye, L.-X. Qin, W. Yang, H.-Y. Wang, Z.-Y. Tang, C. M. Croce, and X. W. Wang. Identification of microRNA-181 by genome-wide screening as a critical player in EpCAM-positive hepatic cancer stem cells. *Hepatology*, 50(2):472–480, aug 2009. ISSN 02709139. doi: 10.1002/hep.22989. URL http://www.ncbi.nlm.nih.gov/pubmed/19585654http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2721019http://doi.wiley.com/10.1002/hep.22989.

[58] Y. Kang, P. M. Siegel, W. Shu, M. Drobnjak, S. M. Kakonen, C. Cordón-Cardo, T. A. Guise, and J. Massagué. A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell*, 3(6):537–549, jun 2003. ISSN 1535-6108. doi: 10.1016/S1535-6108(03)00132-6. URL https://www.sciencedirect.com/science/article/pii/S1535610803001326.

[59] A. Kentsis, F. Monigatti, K. Dorff, F. Campagne, R. Bachur, and H. Steen. Urine proteomics for profiling of human disease using high accuracy mass spectrometry. *Proteomics. Clinical applications*, 3(9):1052–1061, sep 2009. ISSN 1862-8354. doi: 10.1002/prca.200900008. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2994589{&}tool=pmcentrez{&}rendertype=abstract.

[60] C. Kingsford and S. L. Salzberg. What are decision trees? *Nature biotechnology*, 26(9):1011–3, sep 2008. ISSN 1546-1696. doi: 10.1038/nbt0908-1011. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2701298{&}tool=pmcentrez{&}rendertype=abstract.

[61] V. Kulasingam and E. P. Diamandis. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nature Clinical Practice Oncology*, 5(10):588–599, oct 2008. ISSN 1743-4254. doi: 10.1038/ncponc1187. URL http://www.ncbi.nlm.nih.gov/pubmed/18695711http://www.nature.com/doifinder/10.1038/ncponc1187.

[62] T. Kurahashi, M. Muramaki, K. Yamanaka, I. Hara, and H. Miyake. Expression of the secreted form of clusterin protein in renal cell carcinoma as a predictor of disease extension. *BJU International*, 96(6):895–899, oct 2005. ISSN 1464-4096. doi: 10.1111/j.1464-410X.2005.05733.x. URL http://www.ncbi.nlm.nih.gov/pubmed/16153225http://doi.wiley.com/10.1111/j.1464-410X.2005.05733.x.

[63] J. LaBaer. So, you want to look for biomarkers. *Journal of proteome research*, 4(4):1053–9, jan 2005. ISSN 1535-3893. doi: 10.1021/pr0501259. URL http://dx.doi.org/10.1021/pr0501259.

[64] Y.-H. Lee and D. T. Wong. Saliva: an emerging biofluid for early detection of diseases. *American journal of dentistry*, 22(4):241–8, aug 2009. ISSN 0894-8275. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2860957{&}tool=pmcentrez{&}rendertype=abstract.

[65] H. Li, S. Liu, X. Zhu, S. Yang, J. Xiang, and H. Chen. Clusterin Immunoexpression and its Clinical Significance in Patients with Non-Small Cell Lung Cancer. *Lung*, 188(5):423–431, oct 2010. ISSN 0341-2040. doi: 10.1007/s00408-010-9248-1. URL http://www.ncbi.nlm.nih.gov/pubmed/20614220http://link.springer.com/10.1007/s00408-010-9248-1.

[66] L. Li, R. Zhou, and D. Huang. Two-phase biomedical named entity recognition using CRFs. *Computational biology and chemistry*, 33(4):334–8, aug 2009. ISSN 1476-928X. doi: 10.1016/j.compbiolchem.2009.07.004. URL http://www.sciencedirect.com/science/article/pii/S1476927109000590.

[67] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–2437, oct 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth267. URL http://www.ncbi.nlm.nih.gov/pubmed/15087314https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bth267.

[68] S. V. Litvinov, M. P. Velders, H. A. Bakker, G. J. Fleuren, and S. O. Warnaar. Ep-CAM: a human epithelial antigen is a homophilic cell-cell adhesion molecule. *The Journal of cell biology*, 125(2):437–46, apr 1994. ISSN 0021-9525. URL http://www.ncbi.nlm.nih.gov/pubmed/8163559http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2120036.

[69] S. V. Litvinov, M. Balzar, M. J. Winter, H. A. Bakker, I. H. Briaire-de Bruijn, F. Prins, G. J. Fleuren, and S. O. Warnaar. Epithelial cell adhesion molecule (Ep-CAM) modulates cell-cell interactions mediated by classic cadherins. *The Journal of cell biology*, 139(5):1337–48, dec 1997. ISSN 0021-9525. URL http://www.ncbi.nlm.nih.gov/pubmed/9382878http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2140211.

[70] T. A. Lusardi, J. I. Phillips, J. T. Wiedrick, C. A. Harrington, B. Lind, J. A. Lapidus, J. F. Quinn, and J. A. Saugstad. MicroRNAs in Human Cerebrospinal Fluid as Biomarkers for Alzheimer's Disease. *Journal of*

*Alzheimer's Disease*, 55(3):1223–1233, dec 2016. ISSN 13872877. doi: 10.3233/JAD-160835. URL http://www.ncbi.nlm.nih.gov/pubmed/27814298http://www.medra.org/servlet/aliasResolver?alias=iospress{&}doi=10.3233/JAD-160835.

[71] T. Manolio. Novel risk markers and clinical practice. *The New England journal of medicine*, 349(17):1587–9, oct 2003. ISSN 1533-4406. doi: 10.1056/NEJMp038136. URL http://www.ncbi.nlm.nih.gov/pubmed/14573728.

[72] M. Mariani, S. He, M. McHugh, M. Andreoli, D. Pandya, S. Sieber, Z. Wu, P. Fiedler, S. Shahabi, and C. Ferlini. Integrated Multidimensional Analysis Is Required for Accurate Prognostic Biomarkers in Colorectal Cancer. *PLoS ONE*, 9(7):e101065, jul 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0101065. URL http://www.ncbi.nlm.nih.gov/pubmed/24988459http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4079703http://dx.plos.org/10.1371/journal.pone.0101065.

[73] J. E. McDermott, J. Wang, H. Mitchell, B.-J. Webb-Robertson, R. Hafen, J. Ramey, and K. D. Rodland. Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. *Expert opinion on medical diagnostics*, 7(1):37–51, jan 2013. ISSN 1753-0067. doi: 10.1517/17530059.2012.718329. URL http://www.ncbi.nlm.nih.gov/pubmed/23335946http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3548234.

[74] C. Meyer, Y. Liu, and S. Dooley. Caveolin and TGF-$\beta$ entanglements. *Journal of Cellular Physiology*, 228(11):2097–2102, nov 2013. ISSN 00219541. doi: 10.1002/jcp.24380. URL http://www.ncbi.nlm.nih.gov/pubmed/23559144http://doi.wiley.com/10.1002/jcp.24380.

[75] A. K. Mitra, T. Harding, U. K. Mukherjee, J. S. Jang, Y. Li, R. HongZheng, J. Jen, P. Sonneveld, S. Kumar, W. M. Kuehl, V. Rajkumar, and B. Van Ness. A gene expression signature distinguishes innate response and resistance to proteasome inhibitors in multiple myeloma. *Blood Cancer Journal*, 7(6):e581, jun 2017. ISSN 2044-5385. doi: 10.1038/bcj.2017.56. URL http://www.ncbi.nlm.nih.gov/pubmed/28665416http://www.nature.com/doifinder/10.1038/bcj.2017.56.

[76] H. Miyake, M. Gleave, S. Kamidono, and I. Hara. Overexpression of clusterin in transitional cell carcinoma of the bladder is related to disease progression and recurrence. *Urology*, 59(1):150–4, jan 2002. ISSN 1527-9995. URL http://www.ncbi.nlm.nih.gov/pubmed/11796313.

[77] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, C. Sun, H.-h. Liu, R. Torres, M. Krauthammer, W. W. Lau,

H. Liu, C.-N. Hsu, M. Schuemie, K. B. Cohen, and L. Hirschman. Overview of BioCreative II gene normalization. *Genome biology*, 9 Suppl 2(2):S3, jan 2008. ISSN 1474-760X. doi: 10.1186/gb-2008-9-s2-s3. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-s2-s3.

[78] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, jul 2008. ISSN 1548-7091. doi: 10.1038/nmeth.1226. URL http://www.ncbi.nlm.nih.gov/pubmed/18516045http://www.nature.com/doifinder/10.1038/nmeth.1226.

[79] M. Munz, P. A. Baeuerle, and O. Gires. The Emerging Role of EpCAM in Cancer and Stem Cell Signaling. *Cancer Research*, 69(14):5627–5629, jul 2009. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-09-0654. URL http://www.ncbi.nlm.nih.gov/pubmed/19584271http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-09-0654.

[80] G. I. Murray, M. E. Duncan, P. O'Neil, W. T. Melvin, and J. E. Fothergill. Matrix metalloproteinase-1 is associated with poor prognosis in colorectal cancer. *Nature medicine*, 2(4):461–2, apr 1996. ISSN 1078-8956. URL http://www.ncbi.nlm.nih.gov/pubmed/8597958.

[81] G. I. Murray, M. E. Duncan, P. O'Neil, J. A. McKay, W. T. Melvin, and J. E. Fothergill. Matrix metalloproteinase-1 is associated with poor prognosis in oesophageal cancer. *The Journal of Pathology*, 185(3):256–261, jul 1998. ISSN 0022-3417. doi: 10.1002/(SICI)1096-9896(199807)185:3⟨256::AID-PATH115⟩3.0.CO;2-A. URL http://www.ncbi.nlm.nih.gov/pubmed/9771478http://doi.wiley.com/10.1002/{%}28SICI{%}291096-9896{%}28199807{%}29185{%}3A3{%}3C256{%}3A{%}3AAID-PATH115{%}3E3.0.CO{%}3B2-A.

[82] S. Nalluri, S. Ghoshal-Gupta, A. Kutiyanawalla, S. Gayatri, B. R. Lee, S. Jiwani, A. M. Rojiani, and M. V. Rojiani. TIMP-1 Inhibits Apoptosis in Lung Adenocarcinoma Cells via Interaction with Bcl-2. *PLOS ONE*, 10(9):e0137673, sep 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0137673. URL http://www.ncbi.nlm.nih.gov/pubmed/26366732http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4569297http://dx.plos.org/10.1371/journal.pone.0137673.

[83] K. L. Ng, C. Morais, A. Bernard, N. Saunders, H. Samaratunga, G. Gobe, and S. Wood. A systematic review and meta-analysis of immunohistochemical biomarkers that differentiate chromophobe renal cell carcinoma from renal oncocytoma. *Journal of Clinical Pathology*, 69(8):661–671, aug 2016. ISSN 0021-9746. doi: 10.1136/jclinpath-2015-203585. URL http://www.ncbi.nlm.nih.gov/pubmed/26951082http://jcp.bmj.com/lookup/doi/10.1136/jclinpath-2015-203585.

[84] W. S. Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–7, dec 2006. ISSN 1087-0156. doi: 10.1038/nbt1206-1565. URL http://www.ncbi.nlm.nih.gov/pubmed/17160063.

[85] K. O'Shea, S. J. Cameron, K. E. Lewis, C. Lu, and L. A. Mur. Metabolomic-based biomarker discovery for non-invasive lung cancer screening: A case study. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1860(11): 2682–2687, nov 2016. ISSN 03044165. doi: 10.1016/j.bbagen.2016.07.007. URL http://www.ncbi.nlm.nih.gov/pubmed/27423423http://linkinghub.elsevier.com/retrieve/pii/S030441651630246X.

[86] W. A. Osta, Y. Chen, K. Mikhitarian, M. Mitas, M. Salem, Y. A. Hannun, D. J. Cole, and W. E. Gillanders. EpCAM Is Overexpressed in Breast Cancer and Is a Potential Target for Breast Cancer Gene Therapy. *Cancer Research*, 64(16):5818–5824, aug 2004. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-04-0754. URL http://www.ncbi.nlm.nih.gov/pubmed/15313925http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-04-0754.

[87] N. Patani, L.-A. Martin, J. S. Reis-Filho, and M. Dowsett. The role of caveolin-1 in human breast cancer. *Breast Cancer Research and Treatment*, 131(1):1–15, jan 2012. ISSN 0167-6806. doi: 10.1007/s10549-011-1751-4. URL http://www.ncbi.nlm.nih.gov/pubmed/21901387http://link.springer.com/10.1007/s10549-011-1751-4.

[88] R. Patra and S. K. Saha. A kernel-based approach for biomedical named entity recognition. *TheScientificWorldJournal*, 2013:950796, jan 2013. ISSN 1537-744X. doi: 10.1155/2013/950796. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3891429{&}tool=pmcentrez{&}rendertype=abstract.

[89] C. Pauli, M. Münz, C. Kieu, B. Mack, P. Breinl, B. Wollenberg, S. Lang, R. Zeidler, and O. Gires. Tumor-specific glycosylation of the carcinoma-associated epithelial cell adhesion molecule EpCAM in head and neck carcinomas. *Cancer letters*, 193(1):25–32, apr 2003. ISSN 0304-3835. URL http://www.ncbi.nlm.nih.gov/pubmed/12691820.

[90] S. Pavlides, D. Whitaker-Menezes, R. Castello-Cros, N. Flomenberg, A. K. Witkiewicz, P. G. Frank, M. C. Casimiro, C. Wang, P. Fortina, S. Addya, R. G. Pestell, U. E. Martinez-Outschoorn, F. Sotgia, and M. P. Lisanti. The reverse Warburg effect: Aerobic glycolysis in cancer associated fibroblasts and the tumor stroma. *Cell Cycle*, 8(23):3984–4001, dec 2009. ISSN 1538-4101. doi: 10.4161/cc.8.23.10238. URL http://www.ncbi.nlm.nih.gov/pubmed/19923890http://www.tandfonline.com/doi/abs/10.4161/cc.8.23.10238.

[91] M. S. Pepper. Role of the matrix metalloproteinase and plasminogen activator-plasmin systems in angiogenesis. *Arteriosclerosis, thrombosis, and vascular biology*, 21 (7):1104–17, jul 2001. ISSN 1524-4636. URL http://www.ncbi.nlm.nih.gov/pubmed/11451738.

[92] C. Plake, L. Royer, R. Winnenburg, J. Hakenberg, and M. Schroeder. GoGene: gene annotation in the fast lane. *Nucleic acids research*, 37(Web Server issue):W300–4, jul 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp429. URL http://nar.oxfordjournals.org/content/37/suppl{_}2/W300.

[93] D. M. W. Powers. Evaluation: from Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2(1): 37–63, 2011. URL http://www.bioinfo.in/contents.php?id=51.

[94] J. Quackenbush. Microarray Analysis and Tumor Classification. *New England Journal of Medicine*, 354(23):2463–2472, jun 2006. ISSN 0028-4793. doi: 10.1056/NEJMra042342. URL http://www.ncbi.nlm.nih.gov/pubmed/16760446http://www.nejm.org/doi/abs/10.1056/NEJMra042342.

[95] S. K. Rayala, P. den Hollander, S. Balasenthil, Z. Yang, R. R. Broaddus, and R. Kumar. Functional regulation of oestrogen receptor pathway by the dynein light chain 1. *EMBO reports*, 6(6):538–44, jun 2005. ISSN 1469-221X. doi: 10.1038/sj.embor.7400417. URL http://www.ncbi.nlm.nih.gov/pubmed/15891768http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1369089.

[96] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno. Text processing through Web services: calling Whatizit. *Bioinformatics (Oxford, England)*, 24(2):296–8, jan 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm557. URL http://bioinformatics.oxfordjournals.org/content/24/2/296.

[97] M. Redondo, E. Villar, J. Torres-Muñoz, T. Tellez, M. Morell, and C. K. Petito. Overexpression of Clusterin in Human Breast Carcinoma. *The American Journal of Pathology*, 157(2):393–399, aug 2000. ISSN 00029440. doi: 10.1016/S0002-9440(10)64552-X. URL http://www.ncbi.nlm.nih.gov/pubmed/10934144http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1850123http://linkinghub.elsevier.com/retrieve/pii/S000294401064552X.

[98] A. Rimessi, S. Marchi, S. Patergnani, and P. Pinton. H-Ras-driven tumoral maintenance is sustained through caveolin-1-dependent alterations in calcium signaling. *Oncogene*, 33(18):2329–2340, may 2014. ISSN 0950-9232. doi: 10.1038/onc.2013.192. URL http://www.ncbi.nlm.nih.gov/pubmed/23728347http://www.nature.com/articles/onc2013192.

[99] F. Rizzi and S. Bettuzzi. The clusterin paradigm in prostate and breast carcinogenesis. *Endocrine-Related Cancer*, 17(1):R1–R17, mar 2010. ISSN 1351-0088. doi: 10.1677/ERC-09-0140. URL http://www.ncbi.nlm.nih.gov/pubmed/19903745https://erc.bioscientifica.com/view/journals/erc/17/1/R1.xml.

[100] A. M. Santiago, M. Rocha, A. Dourado, and J. P. Arrais. Mixed-Integer Programming Model for Profiling Disease Biomarkers from Gene Expression Studies. In I. Rojas and F. Ortuño, editors, *Bioinformatics and Biomedical Engineering*, pages 50–61. Springer, Cham, Granada, apr 2017. doi: 10.1007/978-3-319-56154-7_6. URL https://link.springer.com/chapter/10.1007/978-3-319-56154-7{_}6.

[101] E. Schäfer, H. Seelert, N. H. Reifschneider, F. Krause, N. A. Dencher, and J. Vonck. Architecture of Active Mammalian Respiratory Chain Supercomplexes. *Journal of Biological Chemistry*, 281(22):15370–15375, jun 2006. ISSN 0021-9258. doi: 10.1074/jbc.M513525200. URL http://www.ncbi.nlm.nih.gov/pubmed/16551638http://www.jbc.org/lookup/doi/10.1074/jbc.M513525200.

[102] S. Schölzel, W. Zimmermann, G. Schwarzkopf, F. Grunert, B. Rogaczewski, and J. Thompson. Carcinoembryonic Antigen Family Members CEACAM6 and CEACAM7 Are Differentially Expressed in Normal Tissues and Oppositely Deregulated in Hyperplastic Colorectal Polyps and Early Adenomas. *The American Journal of Pathology*, 156(2):595–605, feb 2000. ISSN 00029440. doi: 10.1016/S0002-9440(10)64764-5. URL http://www.ncbi.nlm.nih.gov/pubmed/10666389http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1850034http://linkinghub.elsevier.com/retrieve/pii/S0002944010647645.

[103] M. G. Schrauder, R. Strick, R. Schulz-Wendtland, P. L. Strissel, L. Kahmann, C. R. Loehberg, M. P. Lux, S. M. Jud, A. Hartmann, A. Hein, C. M. Bayer, M. R. Bani, S. Richter, B. R. Adamietz, E. Wenkel, C. Rauh, M. W. Beckmann, and P. A. Fasching. Circulating Micro-RNAs as Potential Blood-Based Markers for Early Stage Breast Cancer Detection. *PLoS ONE*, 7(1):e29770, jan 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0029770. URL http://dx.plos.org/10.1371/journal.pone.0029770.

[104] A.-S. Schrohl, S. Würtz, E. Kohn, R. E. Banks, H. J. Nielsen, F. C. G. J. Sweep, and N. Brünner. Banking of biological fluids for studies of disease-associated protein biomarkers. *Molecular & cellular proteomics : MCP*, 7(10):2061–6, oct 2008. ISSN 1535-9484. doi: 10.1074/mcp.R800010-MCP200. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2559931{&}tool=pmcentrez{&}rendertype=abstract.

[105] N. Sidenius and F. Blasi. The urokinase plasminogen activator system in cancer: recent advances and implication for prognosis and therapy. *Cancer metastasis re-*

*views*, 22(2-3):205–22. ISSN 0167-7659. URL http://www.ncbi.nlm.nih.gov/pubmed/12784997.

[106] D. Sidransky. Nucleic acid-based methods for the detection of cancer. *Science (New York, N.Y.)*, 278(5340):1054–9, nov 1997. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/9353179.

[107] M. Skeppstedt, M. Kvist, G. H. Nilsson, and H. Dalianis. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics*, 49:148–158, jun 2014. ISSN 15320464. doi: 10.1016/j.jbi.2014.01.012. URL http://www.j-biomed-inform.com/article/S1532046414000148/fulltext.

[108] G. Song, S. Xu, H. Zhang, Y. Wang, C. Xiao, T. Jiang, L. Wu, T. Zhang, X. Sun, L. Zhong, C. Zhou, Z. Wang, Z. Peng, J. Chen, and X. Wang. TIMP1 is a prognostic marker for the progression and metastasis of colon cancer through FAK-PI3K/AKT and MAPK pathway. *Journal of experimental & clinical cancer research : CR*, 35(1):148, 2016. ISSN 1756-9966. doi: 10.1186/s13046-016-0427-7. URL http://www.ncbi.nlm.nih.gov/pubmed/27644693http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5028967.

[109] T. Sorsa, L. Tjäderhane, and T. Salo. Matrix metalloproteinases (MMPs) in oral diseases. *Oral Diseases*, 10(6):311–318, nov 2004. ISSN 1354523X. doi: 10.1111/j.1601-0825.2004.01038.x. URL http://www.ncbi.nlm.nih.gov/pubmed/15533204http://doi.wiley.com/10.1111/j.1601-0825.2004.01038.x.

[110] F. Sotgia, U. E. Martinez-Outschoorn, A. Howell, R. G. Pestell, S. Pavlides, and M. P. Lisanti. Caveolin-1 and Cancer Metabolism in the Tumor Microenvironment: Markers, Models, and Mechanisms. *Annual Review of Pathology: Mechanisms of Disease*, 7 (1):423–467, feb 2012. ISSN 1553-4006. doi: 10.1146/annurev-pathol-011811-120856. URL http://www.ncbi.nlm.nih.gov/pubmed/22077552http://www.annualreviews.org/doi/10.1146/annurev-pathol-011811-120856.

[111] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, mar 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti033. URL http://www.ncbi.nlm.nih.gov/pubmed/15374862https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti033.

[112] A. Statnikov, L. Wang, and C. F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification.

*BMC bioinformatics*, 9:319, jul 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-319. URL http://www.ncbi.nlm.nih.gov/pubmed/18647401http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2492881.

[113] J. Steinberg, R. Oyasu, S. Lang, S. Sintich, A. Rademaker, C. Lee, J. M. Kozlowski, and J. A. Sensibar. Intracellular levels of SGP-2 (Clusterin) correlate with tumor grade in prostate cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 3(10):1707–11, oct 1997. ISSN 1078-0432. URL http://www.ncbi.nlm.nih.gov/pubmed/9815554.

[114] K. Strimbu and J. A. Tavel. What are biomarkers? *Current opinion in HIV and AIDS*, 5(6):463–6, nov 2010. ISSN 1746-6318. doi: 10.1097/COH.0b013e32833ed177. URL http://www.ncbi.nlm.nih.gov/pubmed/20978388http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3078627.

[115] L. Tanabe, U. Scherf, L. H. Smith, J. K. Lee, L. Hunter, and J. N. Weinstein. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques*, 27(6):1210–4, 1216–7, dec 1999. ISSN 0736-6205. URL http://www.ncbi.nlm.nih.gov/pubmed/10631500.

[116] B. Tang, H. Cao, X. Wang, Q. Chen, and H. Xu. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international*, 2014:240403, jan 2014. ISSN 2314-6141. doi: 10.1155/2014/240403. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3963372{&}tool=pmcentrez{&}rendertype=abstract.

[117] F. Tas, S. Karabulut, M. Serilmez, R. Ciftci, and D. Duranyildiz. Clinical significance of serum epithelial cell adhesion molecule (EPCAM) and vascular cell adhesion molecule-1 (VCAM-1) levels in patients with epithelial ovarian cancer. *Tumor Biology*, 35(4):3095–3102, apr 2014. ISSN 1010-4283. doi: 10.1007/s13277-013-1401-z. URL http://www.ncbi.nlm.nih.gov/pubmed/24307621http://link.springer.com/10.1007/s13277-013-1401-z.

[118] J. Thompson, W. Zimmermann, P. Nollau, M. Neumaier, J. Weber-Arden, H. Schrewe, I. Craig, and T. Willcocks. CGM2, a member of the carcinoembryonic antigen gene family is down-regulated in colorectal carcinomas. *The Journal of biological chemistry*, 269(52):32924–31, dec 1994. ISSN 0021-9258. URL http://www.ncbi.nlm.nih.gov/pubmed/7806520.

[119] E. Y. Ting Tse, F. C. Fat Ko, E. K. Kwan Tung, L. K. Chan, T. K. Wah Lee, E. S. Wai Ngan, K. Man, A. S. Tsai Wong, I. O.-L. Ng, and J. W. Ping Yam. Caveolin-1 overexpression is associated with hepatocellular carcinoma tumourigenesis and

metastasis. *The Journal of Pathology*, 226(4):645–653, mar 2012. ISSN 00223417. doi: 10.1002/path.3957. URL http://www.ncbi.nlm.nih.gov/pubmed/22072235http://doi.wiley.com/10.1002/path.3957.

[120] M. Toricelli, F. H. Melo, G. B. Peres, D. C. Silva, and M. G. Jasiulionis. Timp1 interacts with beta-1 integrin and CD63 along melanoma genesis and confers anoikis resistance by activating PI3-K signaling pathway independently of Akt phosphorylation. *Molecular Cancer*, 12(1):1095, dec 2013. ISSN 1476-4598. doi: 10.1186/1476-4598-12-22. URL http://www.ncbi.nlm.nih.gov/pubmed/23522389http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3635912http://molecular-cancer.biomedcentral.com/articles/10.1186/1476-4598-12-22.

[121] S. Ulisse, E. Baldini, M. Toller, E. Marchioni, L. Giacomelli, E. De Antoni, E. Ferretti, A. Marzullo, F. Graziano, P. Trimboli, L. Biordi, F. Curcio, A. Gulino, F. Ambesi-Impiombato, and M. D'Armiento. Differential expression of the components of the plasminogen activating system in human thyroid tumour derived cell lines and papillary carcinomas. *European Journal of Cancer*, 42(15):2631–2638, oct 2006. ISSN 0959-8049. doi: 10.1016/J.EJCA.2006.04.017. URL https://www.sciencedirect.com/science/article/pii/S0959804906005570.

[122] S. Ulisse, E. Baldini, S. Sorrenti, and M. D'Armiento. The urokinase plasminogen activator system: a target for anti-cancer therapy. *Current cancer drug targets*, 9(1):32–71, feb 2009. ISSN 1873-5576. URL http://www.ncbi.nlm.nih.gov/pubmed/19200050.

[123] R. K. Vadlamudi, R. Bagheri-Yarmand, Z. Yang, S. Balasenthil, D. Nguyen, A. A. Sahin, P. den Hollander, and R. Kumar. Dynein light chain 1, a p21-activated kinase 1-interacting substrate, promotes cancerous phenotypes. *Cancer Cell*, 5(6):575–585, jun 2004. ISSN 15356108. doi: 10.1016/j.ccr.2004.05.022. URL http://www.ncbi.nlm.nih.gov/pubmed/15193260http://linkinghub.elsevier.com/retrieve/pii/S1535610804001473.

[124] E. A. Vaisberg, M. P. Koonce, and J. R. McIntosh. Cytoplasmic dynein plays a role in mammalian mitotic spindle formation. *The Journal of cell biology*, 123(4):849–58, nov 1993. ISSN 0021-9525. doi: 10.1083/JCB.123.4.849. URL http://www.ncbi.nlm.nih.gov/pubmed/8227145http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2200153.

[125] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, jan 2002. ISSN

00280836. doi: 10.1038/415530a. URL http://www.nature.com/doifinder/10.1038/415530a.

[126] J. Vonck and E. Schäfer. Supramolecular organization of protein complexes in the mitochondrial inner membrane. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1793(1):117–124, jan 2009. ISSN 01674889. doi: 10.1016/j.bbamcr.2008.05.019. URL http://www.ncbi.nlm.nih.gov/pubmed/18573282http://linkinghub.elsevier.com/retrieve/pii/S0167488908002012.

[127] Q. Wang, P. Gao, X. Wang, and Y. Duan. Investigation and identification of potential biomarkers in human saliva for the early diagnosis of oral squamous cell carcinoma. *Clinica chimica acta; international journal of clinical chemistry*, 427:79–85, jan 2014. ISSN 1873-3492. doi: 10.1016/j.cca.2013.10.004. URL http://www.ncbi.nlm.nih.gov/pubmed/24144867.

[128] Y. Wang. The role and regulation of urokinase-type plasminogen activator receptor gene expression in cancer invasion and metastasis. *Medicinal research reviews*, 21(2):146–70, mar 2001. ISSN 0198-6325. URL http://www.ncbi.nlm.nih.gov/pubmed/11223863.

[129] Y.-Y. Wang, L. Li, Z.-S. Zhao, and H.-J. Wang. Clinical utility of measuring expression levels of KAP1, TIMP1 and STC2 in peripheral blood of patients with gastric cancer. *World Journal of Surgical Oncology*, 11(1):81, apr 2013. ISSN 1477-7819. doi: 10.1186/1477-7819-11-81. URL http://www.ncbi.nlm.nih.gov/pubmed/23548070http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3620905http://wjso.biomedcentral.com/articles/10.1186/1477-7819-11-81.

[130] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, jan 2009. ISSN 1471-0056. doi: 10.1038/nrg2484. URL http://www.ncbi.nlm.nih.gov/pubmed/19015660http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2949280http://www.nature.com/doifinder/10.1038/nrg2484.

[131] Z. Wang, N. Wang, P. Liu, F. Peng, H. Tang, Q. Chen, R. Xu, Y. Dai, Y. Lin, X. Xie, C. Peng, and H. Situ. Caveolin-1, a stress-related oncotarget, in drug resistance. *Oncotarget*, 6(35):37135–50, nov 2015. ISSN 1949-2553. doi: 10.18632/oncotarget.5789. URL http://www.ncbi.nlm.nih.gov/pubmed/26431273http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4741920http://www.oncotarget.com/fulltext/5789.

[132] J. B. Welsh, P. P. Zarrinkar, L. M. Sapinoso, S. G. Kern, C. A. Behling, B. J. Monk, D. J. Lockhart, R. A. Burger, G. M. Hampton, A. Ben-Dor, and Z. X. Za. Analysis of

gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proceedings of the National Academy of Sciences*, 98(3):1176–1181, jan 2001. ISSN 0027-8424. doi: 10.1073/pnas.98.3.1176. URL http://www.pnas.org/cgi/doi/10.1073/pnas.98.3.1176.

[133] C. Wilson, J. E. Lin, P. Li, A. E. Snook, J. Gong, T. Sato, C. Liu, M. A. Girondo, H. Rui, T. Hyslop, and S. A. Waldman. The Paracrine Hormone for the GUCY2C Tumor Suppressor, Guanylin, Is Universally Lost in Colorectal Cancer. *Cancer Epidemiology Biomarkers & Prevention*, 23(11):2328–2337, nov 2014. ISSN 1055-9965. doi: 10.1158/1055-9965.EPI-14-0440. URL http://www.ncbi.nlm.nih.gov/pubmed/25304930http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4221461http://cebp.aacrjournals.org/cgi/doi/10.1158/1055-9965.EPI-14-0440.

[134] S. Ø. Würtz, S. Møller, H. Mouridsen, P. B. Hertel, E. Friis, and N. Brünner. Plasma and Serum Levels of Tissue Inhibitor of Metalloproteinases-1 Are Associated with Prognosis in Node-negative Breast Cancer. *Molecular & Cellular Proteomics*, 7(2):424–430, feb 2008. ISSN 1535-9476. doi: 10.1074/mcp.M700305-MCP200. URL http://www.mcponline.org/lookup/doi/10.1074/mcp.M700305-MCP200.

[135] S. Ø. Würtz, S. Ø. Würtz, A.-S. Schrohl, H. Mouridsen, and N. Brünner. TIMP-1 as a tumor marker in breast cancer An update. *Acta Oncologica*, 47(4):580–590, jan 2008. ISSN 0284-186X. doi: 10.1080/02841860802022976. URL http://www.ncbi.nlm.nih.gov/pubmed/18465326http://www.tandfonline.com/doi/full/10.1080/02841860802022976.

[136] P. Xiu, X.-F. Dong, X.-P. Li, and J. Li. Clusterin: Review of research progress and looking ahead to direction in hepatocellular carcinoma. *World journal of gastroenterology*, 21(27):8262–70, jul 2015. ISSN 2219-2840. doi: 10.3748/wjg.v21.i27.8262. URL http://www.ncbi.nlm.nih.gov/pubmed/26217078http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4507096.

[137] K. YAMANAKA, P. ROCCHI, H. MIYAKE, L. FAZLI, A. SO, U. ZANGEMEISTER-WITTKE, and M. E. GLEAVE. Induction of apoptosis and enhancement of chemosensitivity in human prostate cancer LNCaP cells using bispecific antisense oligonucleotide targeting Bcl-2 and Bcl-xL genes. *BJU International*, 97(6):1300–1308, jun 2006. ISSN 1464-4096. doi: 10.1111/j.1464-410X.2006.06147.x. URL http://www.ncbi.nlm.nih.gov/pubmed/16686729http://doi.wiley.com/10.1111/j.1464-410X.2006.06147.x.

[138] S.-f. Yang, J.-Y. Yang, C.-H. Huang, S.-N. Wang, C.-P. Lu, C.-J. Tsai, C.-Y. Chai, and Y.-T. Yeh. Increased caveolin-1 expression associated with prolonged overall survival

rate in hepatocellular carcinoma. *Pathology*, 42(5):438–445, aug 2010. ISSN 00313025. doi: 10.3109/00313025.2010.494293. URL http://www.ncbi.nlm.nih.gov/pubmed/20632820http://linkinghub.elsevier.com/retrieve/pii/S0031302516334055.

[139] L. Yeganova, L. Smith, and W. J. Wilbur. Identification of related gene/protein names based on an HMM of name variations. *Computational biology and chemistry*, 28(2): 97–107, apr 2004. ISSN 1476-9271. doi: 10.1016/j.compbiolchem.2003.12.003. URL http://www.sciencedirect.com/science/article/pii/S1476927103001014.

[140] E. Younesi, L. Toldo, B. Müller, C. M. Friedrich, N. Novac, A. Scheer, M. Hofmann-Apitius, and J. Fluck. Mining biomarker information in biomedical literature. *BMC medical informatics and decision making*, 12(1):148, jan 2012. ISSN 1472-6947. doi: 10.1186/1472-6947-12-148. URL http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-12-148.

[141] J. Zhang, D. Shen, G. Zhou, J. Su, and C.-L. Tan. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. *Journal of biomedical informatics*, 37(6):411–22, dec 2004. ISSN 1532-0464. doi: 10.1016/j.jbi.2004.08.005. URL http://www.j-biomed-inform.com/article/S1532046404000838/fulltext.

[142] J. Zhou, L. Zhang, Y. Gu, K. Li, Y. Nie, D. Fan, and Y. Feng. Dynamic expression of CEACAM7 in precursor lesions of gastric carcinoma and its prognostic value in combination with CEA. *World journal of surgical oncology*, 9:172, dec 2011. ISSN 1477-7819. doi: 10.1186/1477-7819-9-172. URL http://www.ncbi.nlm.nih.gov/pubmed/22195770http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3258196.

[143] L. Zhou and Y. Zhu. The EpCAM overexpression is associated with clinicopathological significance and prognosis in hepatocellular carcinoma patients: A systematic review and meta-analysis. *International Journal of Surgery*, 56:274–280, aug 2018. ISSN 1743-9191. doi: 10.1016/J.IJSU.2018.06.025. URL https://www.sciencedirect.com/science/article/pii/S1743919118315188?via{%}3Dihub{#}bib7.

[144] M. Zhou, Z. Diao, X. Yue, Y. Chen, H. Zhao, L. Cheng, and J. Sun. Construction and analysis of dysregulated lncRNA-associated ceRNA network identified novel lncRNA biomarkers for early diagnosis of human pancreatic cancer. *Oncotarget*, 7(35):56383–56394, aug 2016. ISSN 1949-2553. doi: 10.18632/oncotarget.10891. URL http://www.ncbi.nlm.nih.gov/pubmed/27487139http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5302921http://www.oncotarget.com/abstract/10891.

[145] M. Zucknick, S. Richardson, and E. A. Stronach. Comparing the Characteristics of Gene Expression Profiles Derived by Univariate and Multivariate Classification Methods. *Statistical Applications in Genetics and Molecular Biology*, 7 (1):Article7, jan 2008. ISSN 1544-6115. doi: 10.2202/1544-6115.1307. URL http://www.ncbi.nlm.nih.gov/pubmed/18312212http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2496885https://www.degruyter.com/view/j/sagmb.2008.7.1/sagmb.2008.7.1.1307/sagmb.2008.7.1.1307.xml.

[146] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen. Frontiers of biomedical text mining: current progress. *Briefings in bioinformatics*, 8(5):358–75, sep 2007. ISSN 1477-4054. doi: 10.1093/bib/bbm045. URL http://bib.oxfordjournals.org/content/8/5/358.

# A

## SUPPORT MATERIAL

This chapter contains the supplementary results tables whose length, if they were integrated into the main text of this document, could compromise its readability. Tables S1 and S2 describe the performance metrics results for all models with different parameter combinations generated through the method developed in this work greater or equal to a specified MCC, when considering or ignoring ambiguous samples, respectively. Adding to this, Tables S3, S4 and S5 describe the performance metrics results, greater or equal to a specified MCC, for the alternative methods used for comparison with the method developed in this work, namely K-Nearest Neighbours (KNN), Support Vector Machines (SVM) and Random Decision Forests (RDF).

Table S1.: MCC and $F_1$-score performance measures, together with accuracy, sensitivity, precision and specificity, for all parameter combinations with a MCC greater than $-0.1$, when considering ambiguous samples. **%Ambiguous** indicates the percentage of samples that were classified as ambiguous.

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\epsilon$ | $\psi$ | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC | %Ambiguous |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.10 | 1.00 | 1.00 | 0.1 | 3.0 | 78.95 | 0.858 | 0.905 | 0.135 | 0.881 | -0.011 | 1.27 |
| 0.01 | 0.10 | 0.10 | 0.1 | 3.0 | 78.95 | 0.858 | 0.905 | 0.135 | 0.881 | -0.011 | 1.27 |
| 0.10 | 0.01 | 1.00 | 0.1 | 3.5 | 76.41 | 0.828 | 0.904 | 0.154 | 0.864 | -0.025 | 3.99 |
| 0.10 | 0.00 | 1.00 | 0.1 | 3.5 | 76.23 | 0.826 | 0.904 | 0.154 | 0.863 | -0.028 | 3.81 |
| 0.01 | 0.00 | 0.10 | 0.1 | 3.5 | 76.23 | 0.826 | 0.904 | 0.154 | 0.863 | -0.028 | 3.81 |
| 0.10 | 1.00 | 1.00 | 0.5 | 3.5 | 77.31 | 0.840 | 0.903 | 0.135 | 0.870 | -0.036 | 3.99 |
| 0.01 | 0.10 | 0.10 | 0.5 | 3.5 | 77.31 | 0.840 | 0.903 | 0.135 | 0.870 | -0.036 | 3.99 |
| 0.10 | 1.00 | 1.00 | 1.0 | 3.0 | 78.95 | 0.860 | 0.903 | 0.115 | 0.881 | -0.037 | 1.09 |
| 0.01 | 0.10 | 0.10 | 1.0 | 3.0 | 78.95 | 0.860 | 0.903 | 0.115 | 0.881 | -0.037 | 1.09 |
| 0.10 | 0.01 | 1.00 | 1.0 | 3.0 | 78.77 | 0.858 | 0.903 | 0.115 | 0.880 | -0.040 | 1.45 |
| 0.10 | 0.10 | 1.00 | 1.0 | 3.0 | 78.77 | 0.858 | 0.903 | 0.115 | 0.880 | -0.040 | 1.27 |
| 0.01 | 0.01 | 0.10 | 1.0 | 3.0 | 78.77 | 0.858 | 0.903 | 0.115 | 0.880 | -0.040 | 1.27 |
| 0.10 | 1.00 | 0.00 | 0.1 | 3.5 | 76.59 | 0.832 | 0.902 | 0.135 | 0.865 | -0.047 | 3.99 |
| 0.10 | 1.00 | 0.00 | 0.5 | 3.5 | 76.59 | 0.832 | 0.902 | 0.135 | 0.865 | -0.047 | 3.99 |
| 0.10 | 1.00 | 0.00 | 1.0 | 3.5 | 76.59 | 0.832 | 0.902 | 0.135 | 0.865 | -0.047 | 3.99 |
| 0.10 | 1.00 | 0.01 | 0.1 | 3.5 | 76.59 | 0.832 | 0.902 | 0.135 | 0.865 | -0.047 | 4.17 |
| 0.10 | 1.00 | 0.01 | 0.5 | 3.5 | 76.59 | 0.832 | 0.902 | 0.135 | 0.865 | -0.047 | 4.17 |
| 0.10 | 1.00 | 0.01 | 1.0 | 3.5 | 76.59 | 0.832 | 0.902 | 0.135 | 0.865 | -0.047 | 4.17 |
| 0.01 | 0.10 | 0.00 | 0.1 | 3.5 | 76.59 | 0.832 | 0.902 | 0.135 | 0.865 | -0.047 | 3.99 |
| 0.01 | 0.10 | 0.00 | 0.5 | 3.5 | 76.59 | 0.832 | 0.902 | 0.135 | 0.865 | -0.047 | 3.99 |
| 0.01 | 0.10 | 0.00 | 1.0 | 3.5 | 76.59 | 0.832 | 0.902 | 0.135 | 0.865 | -0.047 | 3.99 |
| 0.10 | 1.00 | 0.10 | 0.5 | 3.5 | 76.41 | 0.830 | 0.902 | 0.135 | 0.864 | -0.050 | 4.36 |

*Continued on next page*

Table S1 – *Continued from previous page*

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\epsilon$ | $\psi$ | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC | %Ambiguous |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 0.10 | 0.01 | 0.5 | 3.5 | 76.41 | 0.830 | 0.902 | 0.135 | 0.864 | -0.050 | 4.36 |
| 0.10 | 0.10 | 1.00 | 0.1 | 3.5 | 76.23 | 0.828 | 0.902 | 0.135 | 0.863 | -0.052 | 4.36 |
| 0.10 | 1.00 | 0.10 | 1.0 | 3.5 | 76.23 | 0.828 | 0.902 | 0.135 | 0.863 | -0.052 | 4.17 |
| 0.01 | 0.01 | 0.10 | 0.1 | 3.5 | 76.23 | 0.828 | 0.902 | 0.135 | 0.863 | -0.052 | 4.36 |
| 0.01 | 0.10 | 0.01 | 1.0 | 3.5 | 76.23 | 0.828 | 0.902 | 0.135 | 0.863 | -0.052 | 4.17 |
| 0.10 | 1.00 | 1.00 | 0.5 | 3.0 | 79.49 | 0.868 | 0.902 | 0.096 | 0.885 | -0.057 | 1.27 |
| 0.01 | 0.10 | 0.10 | 0.5 | 3.0 | 79.49 | 0.868 | 0.902 | 0.096 | 0.885 | -0.057 | 1.27 |
| 0.10 | 1.00 | 0.10 | 0.5 | 3.0 | 77.50 | 0.844 | 0.901 | 0.115 | 0.872 | -0.060 | 1.63 |
| 0.10 | 1.00 | 1.00 | 1.0 | 3.5 | 77.50 | 0.844 | 0.901 | 0.115 | 0.872 | -0.060 | 3.27 |
| 0.01 | 0.10 | 0.01 | 0.5 | 3.0 | 77.50 | 0.844 | 0.901 | 0.115 | 0.872 | -0.060 | 1.63 |
| 0.01 | 0.10 | 0.10 | 1.0 | 3.5 | 77.50 | 0.844 | 0.901 | 0.115 | 0.872 | -0.060 | 3.27 |
| 0.10 | 1.00 | 0.10 | 0.1 | 3.5 | 75.68 | 0.822 | 0.901 | 0.135 | 0.860 | -0.060 | 4.54 |
| 0.01 | 0.10 | 0.01 | 0.1 | 3.5 | 75.68 | 0.822 | 0.901 | 0.135 | 0.860 | -0.060 | 4.54 |
| 0.10 | 1.00 | 0.10 | 1.0 | 3.0 | 77.31 | 0.842 | 0.901 | 0.115 | 0.870 | -0.062 | 1.63 |
| 0.01 | 0.10 | 0.01 | 1.0 | 3.0 | 77.31 | 0.842 | 0.901 | 0.115 | 0.870 | -0.062 | 1.63 |
| 0.10 | 1.00 | 0.00 | 0.1 | 3.0 | 77.13 | 0.840 | 0.901 | 0.115 | 0.869 | -0.065 | 2.00 |
| 0.10 | 1.00 | 0.00 | 0.5 | 3.0 | 77.13 | 0.840 | 0.901 | 0.115 | 0.869 | -0.065 | 2.00 |
| 0.10 | 1.00 | 0.00 | 1.0 | 3.0 | 77.13 | 0.840 | 0.901 | 0.115 | 0.869 | -0.065 | 2.00 |
| 0.10 | 1.00 | 0.01 | 0.1 | 3.0 | 77.13 | 0.840 | 0.901 | 0.115 | 0.869 | -0.065 | 2.00 |
| 0.10 | 1.00 | 0.01 | 0.5 | 3.0 | 77.13 | 0.840 | 0.901 | 0.115 | 0.869 | -0.065 | 2.00 |
| 0.10 | 1.00 | 0.01 | 1.0 | 3.0 | 77.13 | 0.840 | 0.901 | 0.115 | 0.869 | -0.065 | 2.00 |
| 0.01 | 0.10 | 0.00 | 0.1 | 3.0 | 77.13 | 0.840 | 0.901 | 0.115 | 0.869 | -0.065 | 2.00 |
| 0.01 | 0.10 | 0.00 | 0.5 | 3.0 | 77.13 | 0.840 | 0.901 | 0.115 | 0.869 | -0.065 | 2.00 |
| 0.01 | 0.10 | 0.00 | 1.0 | 3.0 | 77.13 | 0.840 | 0.901 | 0.115 | 0.869 | -0.065 | 2.00 |

*Continued on next page*

Table S1 – *Continued from previous page*

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\epsilon$ | $\psi$ | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC | %Ambiguous |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.10 | 0.00 | 1.00 | 1.0 | 3.5 | 76.77 | 0.836 | 0.901 | 0.115 | 0.867 | -0.071 | 4.17 |
| 0.10 | 0.10 | 1.00 | 0.1 | 3.0 | 76.77 | 0.836 | 0.901 | 0.115 | 0.867 | -0.071 | 1.45 |
| 0.01 | 0.00 | 0.10 | 1.0 | 3.5 | 76.77 | 0.836 | 0.901 | 0.115 | 0.867 | -0.071 | 4.17 |
| 0.01 | 0.01 | 0.10 | 0.1 | 3.0 | 76.77 | 0.836 | 0.901 | 0.115 | 0.867 | -0.071 | 1.45 |
| 0.10 | 0.01 | 1.00 | 1.0 | 3.5 | 76.59 | 0.834 | 0.900 | 0.115 | 0.866 | -0.073 | 4.36 |
| 0.10 | 0.01 | 1.00 | 0.5 | 3.0 | 78.04 | 0.852 | 0.900 | 0.096 | 0.875 | -0.080 | 1.81 |
| 0.10 | 0.01 | 1.00 | 0.1 | 3.0 | 77.50 | 0.846 | 0.900 | 0.096 | 0.872 | -0.088 | 1.27 |
| 0.10 | 0.00 | 1.00 | 0.5 | 3.5 | 75.50 | 0.822 | 0.899 | 0.115 | 0.859 | -0.089 | 4.90 |
| 0.10 | 0.10 | 1.00 | 0.5 | 3.5 | 75.50 | 0.822 | 0.899 | 0.115 | 0.859 | -0.089 | 4.36 |
| 0.01 | 0.00 | 0.10 | 0.5 | 3.5 | 75.50 | 0.822 | 0.899 | 0.115 | 0.859 | -0.089 | 4.90 |
| 0.01 | 0.01 | 0.10 | 0.5 | 3.5 | 75.50 | 0.822 | 0.899 | 0.115 | 0.859 | -0.089 | 4.36 |
| 1.00 | 1.00 | 1.00 | 0.1 | 3.5 | 73.50 | 0.798 | 0.898 | 0.135 | 0.845 | -0.091 | 6.35 |
| 0.10 | 0.10 | 0.10 | 0.1 | 3.5 | 73.50 | 0.798 | 0.898 | 0.135 | 0.845 | -0.091 | 6.35 |
| 0.01 | 0.01 | 0.01 | 0.1 | 3.5 | 73.50 | 0.798 | 0.898 | 0.135 | 0.845 | -0.091 | 6.35 |
| 0.10 | 0.01 | 1.00 | 0.5 | 3.5 | 75.32 | 0.820 | 0.899 | 0.115 | 0.857 | -0.092 | 4.90 |
| 0.10 | 0.00 | 1.00 | 0.1 | 3.0 | 77.13 | 0.842 | 0.899 | 0.096 | 0.870 | -0.093 | 1.81 |
| 0.01 | 0.00 | 0.10 | 0.1 | 3.0 | 77.13 | 0.842 | 0.899 | 0.096 | 0.870 | -0.093 | 1.81 |
| 0.10 | 1.00 | 0.10 | 0.1 | 3.0 | 76.95 | 0.840 | 0.899 | 0.096 | 0.868 | -0.096 | 1.45 |
| 0.01 | 0.10 | 0.01 | 0.1 | 3.0 | 76.95 | 0.840 | 0.899 | 0.096 | 0.868 | -0.096 | 1.45 |

Table S2.: MCC and $F_1$-score performance measures, together with accuracy, sensitivity, precision and specificity, for all parameter combinations with a positive MCC, when ignoring ambiguous samples. **%Ambiguous** indicates the percentage of samples that were classified as ambiguous.

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\epsilon$ | $\psi$ | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC | %Ambiguous |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 1.00 | 1.00 | 1.0 | 3.5 | 80.40 | 0.794 | 0.988 | 0.904 | 0.880 | 0.702 | 69.51 |
| 0.01 | 1.00 | 1.00 | 0.5 | 3.5 | 75.50 | 0.739 | 0.987 | 0.904 | 0.845 | 0.652 | 64.43 |
| 0.01 | 1.00 | 0.00 | 0.1 | 3.5 | 61.16 | 0.599 | 0.955 | 0.731 | 0.736 | 0.333 | 40.11 |
| 0.01 | 1.00 | 0.00 | 0.5 | 3.5 | 61.16 | 0.599 | 0.955 | 0.731 | 0.736 | 0.333 | 40.11 |
| 0.01 | 1.00 | 0.00 | 1.0 | 3.5 | 61.16 | 0.599 | 0.955 | 0.731 | 0.736 | 0.333 | 40.11 |
| 0.01 | 1.00 | 0.01 | 0.1 | 3.5 | 61.16 | 0.599 | 0.955 | 0.731 | 0.736 | 0.333 | 40.11 |
| 0.01 | 1.00 | 0.01 | 0.5 | 3.5 | 61.16 | 0.599 | 0.955 | 0.731 | 0.736 | 0.333 | 40.11 |
| 0.01 | 1.00 | 0.01 | 1.0 | 3.5 | 61.16 | 0.599 | 0.955 | 0.731 | 0.736 | 0.333 | 40.11 |
| 0.01 | 1.00 | 0.10 | 0.1 | 3.5 | 61.16 | 0.599 | 0.955 | 0.731 | 0.736 | 0.333 | 40.11 |
| 0.01 | 1.00 | 1.00 | 0.1 | 3.5 | 58.08 | 0.567 | 0.950 | 0.712 | 0.710 | 0.282 | 38.66 |
| 0.01 | 0.10 | 1.00 | 0.1 | 3.5 | 55.90 | 0.545 | 0.944 | 0.692 | 0.691 | 0.240 | 35.21 |
| 0.01 | 0.00 | 1.00 | 0.1 | 3.5 | 55.17 | 0.537 | 0.944 | 0.692 | 0.685 | 0.232 | 34.12 |
| 0.01 | 0.01 | 1.00 | 0.1 | 3.5 | 55.17 | 0.537 | 0.944 | 0.692 | 0.685 | 0.232 | 34.12 |
| 0.01 | 0.10 | 1.00 | 0.5 | 3.5 | 61.34 | 0.613 | 0.939 | 0.615 | 0.742 | 0.229 | 42.11 |
| 0.01 | 1.00 | 1.00 | 1.0 | 3.0 | 54.26 | 0.531 | 0.936 | 0.654 | 0.678 | 0.186 | 29.95 |
| 0.01 | 1.00 | 0.10 | 0.5 | 3.5 | 54.08 | 0.531 | 0.933 | 0.635 | 0.677 | 0.167 | 34.30 |
| 0.01 | 1.00 | 0.10 | 1.0 | 3.5 | 53.36 | 0.523 | 0.932 | 0.635 | 0.670 | 0.159 | 33.58 |
| 1.00 | 0.00 | 0.10 | 0.5 | 3.5 | 82.03 | 0.884 | 0.915 | 0.212 | 0.899 | 0.129 | 10.89 |
| 1.00 | 0.00 | 0.10 | 1.0 | 3.5 | 82.03 | 0.884 | 0.915 | 0.212 | 0.899 | 0.129 | 10.89 |
| 0.10 | 0.00 | 0.01 | 0.5 | 3.5 | 82.03 | 0.884 | 0.915 | 0.212 | 0.899 | 0.129 | 10.89 |
| 0.10 | 0.00 | 0.01 | 1.0 | 3.5 | 82.03 | 0.884 | 0.915 | 0.212 | 0.899 | 0.129 | 10.89 |
| 1.00 | 0.01 | 0.10 | 0.5 | 3.5 | 81.67 | 0.880 | 0.915 | 0.212 | 0.897 | 0.123 | 9.98 |

*Continued on next page*

Table S2 – *Continued from previous page*

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\epsilon$ | $\psi$ | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC | %Ambiguous |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.00 | 0.10 | 0.10 | 1.0 | 3.5 | 81.67 | 0.880 | 0.915 | 0.212 | 0.897 | 0.123 | 9.98 |
| 0.10 | 0.01 | 0.01 | 1.0 | 3.5 | 81.67 | 0.880 | 0.915 | 0.212 | 0.897 | 0.123 | 9.98 |
| 1.00 | 0.01 | 0.10 | 1.0 | 3.5 | 81.31 | 0.876 | 0.914 | 0.212 | 0.895 | 0.117 | 10.16 |
| 1.00 | 0.01 | 0.01 | 0.5 | 3.5 | 81.85 | 0.884 | 0.913 | 0.192 | 0.898 | 0.105 | 9.98 |
| 1.00 | 0.01 | 0.01 | 1.0 | 3.5 | 81.85 | 0.884 | 0.913 | 0.192 | 0.898 | 0.105 | 10.16 |
| 1.00 | 0.00 | 0.00 | 0.1 | 3.5 | 81.85 | 0.884 | 0.913 | 0.192 | 0.898 | 0.105 | 10.34 |
| 1.00 | 0.00 | 0.00 | 0.5 | 3.5 | 81.85 | 0.884 | 0.913 | 0.192 | 0.898 | 0.105 | 10.34 |
| 1.00 | 0.00 | 0.00 | 1.0 | 3.5 | 81.85 | 0.884 | 0.913 | 0.192 | 0.898 | 0.105 | 10.34 |
| 0.10 | 0.00 | 0.00 | 0.1 | 3.5 | 81.85 | 0.884 | 0.913 | 0.192 | 0.898 | 0.105 | 10.34 |
| 0.10 | 0.00 | 0.00 | 0.5 | 3.5 | 81.85 | 0.884 | 0.913 | 0.192 | 0.898 | 0.105 | 10.34 |
| 0.10 | 0.00 | 0.00 | 1.0 | 3.5 | 81.85 | 0.884 | 0.913 | 0.192 | 0.898 | 0.105 | 10.34 |
| 0.01 | 0.00 | 0.00 | 0.1 | 3.5 | 81.85 | 0.884 | 0.913 | 0.192 | 0.898 | 0.105 | 10.34 |
| 0.01 | 0.00 | 0.00 | 0.5 | 3.5 | 81.85 | 0.884 | 0.913 | 0.192 | 0.898 | 0.105 | 10.34 |
| 0.01 | 0.00 | 0.00 | 1.0 | 3.5 | 81.85 | 0.884 | 0.913 | 0.192 | 0.898 | 0.105 | 10.34 |
| 1.00 | 0.00 | 0.01 | 0.5 | 3.5 | 81.67 | 0.882 | 0.913 | 0.192 | 0.897 | 0.102 | 9.62 |
| 1.00 | 0.10 | 0.00 | 0.1 | 3.5 | 81.67 | 0.882 | 0.913 | 0.192 | 0.897 | 0.102 | 9.62 |
| 1.00 | 0.10 | 0.00 | 0.5 | 3.5 | 81.67 | 0.882 | 0.913 | 0.192 | 0.897 | 0.102 | 9.62 |
| 1.00 | 0.10 | 0.00 | 1.0 | 3.5 | 81.67 | 0.882 | 0.913 | 0.192 | 0.897 | 0.102 | 9.62 |
| 0.10 | 0.01 | 0.00 | 0.1 | 3.5 | 81.67 | 0.882 | 0.913 | 0.192 | 0.897 | 0.102 | 9.62 |
| 0.10 | 0.01 | 0.00 | 0.5 | 3.5 | 81.67 | 0.882 | 0.913 | 0.192 | 0.897 | 0.102 | 9.62 |
| 0.10 | 0.01 | 0.00 | 1.0 | 3.5 | 81.67 | 0.882 | 0.913 | 0.192 | 0.897 | 0.102 | 9.62 |
| 1.00 | 0.00 | 0.01 | 0.1 | 3.5 | 81.67 | 0.882 | 0.913 | 0.192 | 0.897 | 0.102 | 9.80 |
| 0.01 | 1.00 | 0.00 | 0.1 | 3.0 | 54.45 | 0.543 | 0.922 | 0.558 | 0.683 | 0.101 | 22.69 |
| 0.01 | 1.00 | 0.00 | 0.5 | 3.0 | 54.45 | 0.543 | 0.922 | 0.558 | 0.683 | 0.101 | 22.69 |

*Continued on next page*

Table S2 – *Continued from previous page*

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\epsilon$ | $\psi$ | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC | %Ambiguous |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 1.00 | 0.00 | 1.0 | 3.0 | 54.45 | 0.543 | 0.922 | 0.558 | 0.683 | 0.101 | 22.69 |
| 0.01 | 1.00 | 0.01 | 0.1 | 3.0 | 54.45 | 0.543 | 0.922 | 0.558 | 0.683 | 0.101 | 22.69 |
| 0.01 | 1.00 | 0.01 | 0.5 | 3.0 | 54.45 | 0.543 | 0.922 | 0.558 | 0.683 | 0.101 | 22.69 |
| 0.01 | 1.00 | 0.01 | 1.0 | 3.0 | 54.45 | 0.543 | 0.922 | 0.558 | 0.683 | 0.101 | 22.69 |
| 0.01 | 1.00 | 0.10 | 0.1 | 3.0 | 54.45 | 0.543 | 0.922 | 0.558 | 0.683 | 0.101 | 22.69 |
| 1.00 | 0.01 | 0.00 | 0.1 | 3.5 | 81.49 | 0.880 | 0.913 | 0.192 | 0.896 | 0.099 | 9.44 |
| 1.00 | 0.01 | 0.00 | 0.5 | 3.5 | 81.49 | 0.880 | 0.913 | 0.192 | 0.896 | 0.099 | 9.44 |
| 1.00 | 0.01 | 0.00 | 1.0 | 3.5 | 81.49 | 0.880 | 0.913 | 0.192 | 0.896 | 0.099 | 9.44 |
| 1.00 | 0.10 | 0.01 | 1.0 | 3.5 | 81.49 | 0.880 | 0.913 | 0.192 | 0.896 | 0.099 | 9.44 |
| 0.01 | 1.00 | 0.10 | 1.0 | 3.0 | 57.35 | 0.579 | 0.920 | 0.519 | 0.711 | 0.099 | 20.87 |
| 1.00 | 0.10 | 0.10 | 0.5 | 3.5 | 81.31 | 0.878 | 0.913 | 0.192 | 0.895 | 0.096 | 9.26 |
| 0.10 | 0.01 | 0.01 | 0.5 | 3.5 | 81.31 | 0.878 | 0.913 | 0.192 | 0.895 | 0.096 | 9.26 |
| 1.00 | 0.00 | 0.01 | 1.0 | 3.5 | 81.31 | 0.878 | 0.913 | 0.192 | 0.895 | 0.096 | 9.62 |
| 1.00 | 0.01 | 0.01 | 0.1 | 3.5 | 81.31 | 0.878 | 0.913 | 0.192 | 0.895 | 0.096 | 9.62 |
| 1.00 | 0.01 | 0.10 | 0.1 | 3.5 | 81.31 | 0.878 | 0.913 | 0.192 | 0.895 | 0.096 | 9.62 |
| 1.00 | 1.00 | 0.01 | 0.1 | 3.5 | 81.31 | 0.878 | 0.913 | 0.192 | 0.895 | 0.096 | 8.53 |
| 1.00 | 1.00 | 0.01 | 0.5 | 3.5 | 81.31 | 0.878 | 0.913 | 0.192 | 0.895 | 0.096 | 8.53 |
| 1.00 | 0.10 | 0.10 | 0.1 | 3.5 | 81.13 | 0.876 | 0.912 | 0.192 | 0.894 | 0.093 | 8.89 |
| 0.10 | 0.01 | 0.01 | 0.1 | 3.5 | 81.13 | 0.876 | 0.912 | 0.192 | 0.894 | 0.093 | 8.89 |

Table S3.: MCC and $F_1$-score performance measures, together with accuracy, sensitivity, precision and specificity, for all parameter combinations for the KNN classifier, with a MCC greater or equal to 0.25.

| $\psi$ | $k$ | Weights | Algorithm | Leaf Size | $p$ | Metric | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 7 | Distance | Ball | 64 | - | Manhattan | 91.41 | 0.995 | 0.917 | 0.135 | 0.955 | 0.295 |
| 3.0 | 7 | Uniform | *k*-d | 8 | 4.0 | Minkowski | 91.41 | 0.996 | 0.916 | 0.128 | 0.955 | 0.293 |
| 3.0 | 7 | Distance | Ball | 4 | 1.5 | Minkowski | 91.35 | 0.995 | 0.917 | 0.135 | 0.954 | 0.288 |
| 3.0 | 7 | Uniform | Ball | 32 | 2.0 | Minkowski | 91.35 | 0.995 | 0.917 | 0.135 | 0.954 | 0.288 |
| 3.0 | 7 | Uniform | *k*-d | 32 | 3.0 | Minkowski | 91.35 | 0.995 | 0.917 | 0.135 | 0.954 | 0.288 |
| 3.0 | 7 | Distance | *k*-d | 64 | - | Euclidean | 91.35 | 0.995 | 0.916 | 0.128 | 0.954 | 0.285 |
| 3.0 | 7 | Distance | *k*-d | 8 | 3.0 | Minkowski | 91.35 | 0.995 | 0.916 | 0.128 | 0.954 | 0.285 |
| 3.0 | 7 | Uniform | *k*-d | 4 | 4.0 | Minkowski | 91.35 | 0.995 | 0.916 | 0.128 | 0.954 | 0.285 |
| 3.0 | 7 | Distance | *k*-d | 32 | - | Euclidean | 91.29 | 0.994 | 0.917 | 0.135 | 0.954 | 0.282 |
| 3.0 | 7 | Distance | Ball | 16 | - | Manhattan | 91.29 | 0.994 | 0.917 | 0.135 | 0.954 | 0.282 |
| 3.0 | 7 | Uniform | *k*-d | 64 | - | Manhattan | 91.29 | 0.994 | 0.917 | 0.135 | 0.954 | 0.282 |
| 3.0 | 7 | Distance | Ball | 128 | 1.5 | Minkowski | 91.29 | 0.994 | 0.917 | 0.135 | 0.954 | 0.282 |
| 3.0 | 7 | Uniform | Ball | 256 | 2.0 | Minkowski | 91.29 | 0.994 | 0.917 | 0.135 | 0.954 | 0.282 |
| 3.0 | 7 | Distance | *k*-d | 1 | 4.0 | Minkowski | 91.29 | 0.994 | 0.917 | 0.135 | 0.954 | 0.282 |
| 3.0 | 7 | Distance | *k*-d | 1 | 2.0 | Minkowski | 91.29 | 0.994 | 0.917 | 0.135 | 0.954 | 0.282 |
| 3.0 | 7 | Uniform | *k*-d | 64 | 4.0 | Minkowski | 91.29 | 0.994 | 0.917 | 0.135 | 0.954 | 0.282 |
| 3.0 | 7 | Distance | Ball | 8 | 1.5 | Minkowski | 91.29 | 0.995 | 0.916 | 0.128 | 0.954 | 0.279 |
| 3.0 | 7 | Uniform | Ball | 8 | 2.5 | Minkowski | 91.29 | 0.995 | 0.916 | 0.128 | 0.954 | 0.279 |
| 3.0 | 7 | Distance | Ball | 128 | 2.0 | Minkowski | 91.29 | 0.995 | 0.916 | 0.128 | 0.954 | 0.278 |
| 3.0 | 7 | Distance | Ball | 256 | 3.0 | Minkowski | 91.29 | 0.995 | 0.916 | 0.128 | 0.954 | 0.278 |
| 3.0 | 7 | Uniform | *k*-d | 16 | 3.5 | Minkowski | 91.29 | 0.995 | 0.916 | 0.128 | 0.954 | 0.278 |
| 3.0 | 7 | Uniform | Ball | 128 | 4.0 | Minkowski | 91.29 | 0.995 | 0.916 | 0.128 | 0.954 | 0.278 |

*Continued on next page*

Table S3 – *Continued from previous page*

| $\psi$ | $k$ | Weights | Algorithm | Leaf Size | $p$ | Metric | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 7 | Distance | Ball | 128 | 4.0 | Minkowski | 91.29 | 0.995 | 0.916 | 0.128 | 0.954 | 0.278 |
| 3.0 | 5 | Distance | *k*-d | 128 | 1.5 | Minkowski | 90.99 | 0.987 | 0.919 | 0.167 | 0.952 | 0.277 |
| 3.0 | 7 | Uniform | Ball | 16 | 3.0 | Minkowski | 91.23 | 0.993 | 0.917 | 0.135 | 0.954 | 0.277 |
| 3.0 | 7 | Distance | *k*-d | 2 | 3.5 | Minkowski | 91.29 | 0.995 | 0.916 | 0.122 | 0.954 | 0.276 |
| 3.0 | 7 | Distance | *k*-d | 8 | - | Manhattan | 91.23 | 0.993 | 0.917 | 0.135 | 0.954 | 0.276 |
| 3.0 | 7 | Distance | *k*-d | 64 | - | Manhattan | 91.23 | 0.993 | 0.917 | 0.135 | 0.954 | 0.276 |
| 3.0 | 7 | Distance | Ball | 1 | 2.0 | Minkowski | 91.23 | 0.993 | 0.917 | 0.135 | 0.954 | 0.276 |
| 3.0 | 7 | Uniform | *k*-d | 16 | 2.0 | Minkowski | 91.23 | 0.993 | 0.917 | 0.135 | 0.954 | 0.276 |
| 3.0 | 7 | Uniform | *k*-d | 256 | 2.5 | Minkowski | 91.23 | 0.993 | 0.917 | 0.135 | 0.954 | 0.276 |
| 3.0 | 7 | Uniform | *k*-d | 8 | 3.0 | Minkowski | 91.23 | 0.993 | 0.917 | 0.135 | 0.954 | 0.276 |
| 3.0 | 7 | Uniform | Ball | 1 | 3.5 | Minkowski | 91.23 | 0.993 | 0.917 | 0.135 | 0.954 | 0.276 |
| 3.0 | 7 | Uniform | Ball | 16 | 3.5 | Minkowski | 91.23 | 0.993 | 0.917 | 0.135 | 0.954 | 0.276 |
| 3.0 | 7 | Uniform | *k*-d | 16 | 4.0 | Minkowski | 91.23 | 0.993 | 0.917 | 0.135 | 0.954 | 0.276 |
| 3.0 | 7 | Distance | *k*-d | 128 | 4.0 | Minkowski | 91.23 | 0.993 | 0.917 | 0.135 | 0.954 | 0.276 |
| 3.0 | 7 | Uniform | Ball | 32 | - | Euclidean | 91.29 | 0.995 | 0.916 | 0.122 | 0.954 | 0.275 |
| 3.0 | 7 | Distance | *k*-d | 16 | - | Manhattan | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.275 |
| 3.0 | 7 | Distance | Ball | 256 | 3.5 | Minkowski | 91.29 | 0.995 | 0.916 | 0.122 | 0.954 | 0.274 |
| 3.0 | 7 | Distance | *k*-d | 32 | 3.0 | Minkowski | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.273 |
| 3.0 | 7 | Uniform | *k*-d | 16 | - | Euclidean | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.273 |
| 3.0 | 7 | Uniform | Ball | 256 | 1.5 | Minkowski | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.273 |
| 3.0 | 7 | Distance | *k*-d | 64 | 3.0 | Minkowski | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.273 |
| 3.0 | 7 | Uniform | Ball | 128 | - | Euclidean | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.273 |
| 3.0 | 7 | Distance | Ball | 2 | 2.0 | Minkowski | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.273 |
| 3.0 | 7 | Uniform | *k*-d | 128 | 3.0 | Minkowski | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.273 |

*Continued on next page*

Table S3 – *Continued from previous page*

| $\psi$ | $k$ | Weights | Algorithm | Leaf Size | $p$ | Metric | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 7 | Distance | $k$-d | 32 | 4.0 | Minkowski | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.273 |
| 3.0 | 7 | Distance | $k$-d | 4 | - | Euclidean | 91.29 | 0.996 | 0.915 | 0.115 | 0.954 | 0.272 |
| 3.0 | 7 | Uniform | Ball | 128 | - | Manhattan | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.272 |
| 3.0 | 7 | Distance | Ball | 1 | 1.5 | Minkowski | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.272 |
| 3.0 | 7 | Distance | $k$-d | 8 | 1.5 | Minkowski | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.272 |
| 3.0 | 7 | Uniform | Ball | 8 | 3.0 | Minkowski | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.272 |
| 3.0 | 7 | Distance | Ball | 32 | 3.5 | Minkowski | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.272 |
| 3.0 | 7 | Distance | $k$-d | 8 | 3.5 | Minkowski | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.272 |
| 3.0 | 7 | Distance | Ball | 8 | 4.0 | Minkowski | 91.23 | 0.994 | 0.916 | 0.128 | 0.954 | 0.272 |
| 3.0 | 7 | Distance | $k$-d | 256 | 2.0 | Minkowski | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.272 |
| 3.0 | 7 | Distance | Ball | 2 | 1.5 | Minkowski | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.271 |
| 3.0 | 7 | Distance | Ball | 64 | 2.0 | Minkowski | 91.23 | 0.995 | 0.916 | 0.122 | 0.954 | 0.271 |
| 3.0 | 7 | Uniform | Ball | 8 | - | Euclidean | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |
| 3.0 | 7 | Distance | Ball | 8 | - | Manhattan | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |
| 3.0 | 7 | Uniform | $k$-d | 128 | - | Manhattan | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |
| 3.0 | 7 | Distance | $k$-d | 2 | - | Manhattan | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |
| 3.0 | 7 | Uniform | Ball | 2 | 2.0 | Minkowski | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |
| 3.0 | 7 | Uniform | Ball | 64 | 2.0 | Minkowski | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |
| 3.0 | 7 | Distance | Ball | 32 | 2.0 | Minkowski | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |
| 3.0 | 7 | Distance | $k$-d | 16 | 2.0 | Minkowski | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |
| 3.0 | 7 | Distance | Ball | 8 | 2.5 | Minkowski | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |
| 3.0 | 7 | Distance | Ball | 16 | 3.0 | Minkowski | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |
| 3.0 | 7 | Distance | Ball | 128 | 3.0 | Minkowski | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |
| 3.0 | 7 | Uniform | $k$-d | 256 | 3.5 | Minkowski | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |

*Continued on next page*

Table S3 – *Continued from previous page*

| $\psi$ | $k$ | Weights | Algorithm | Leaf Size | $p$ | Metric | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 7 | Distance | $k$-d | 64 | 3.5 | Minkowski | 91.17 | 0.993 | 0.917 | 0.135 | 0.953 | 0.270 |
| 3.0 | 5 | Uniform | Ball | 1 | 2.0 | Minkowski | 90.93 | 0.987 | 0.919 | 0.160 | 0.952 | 0.270 |
| 3.0 | 7 | Uniform | $k$-d | 1 | - | Euclidean | 91.23 | 0.995 | 0.916 | 0.122 | 0.954 | 0.269 |
| 3.0 | 7 | Uniform | Ball | 4 | 1.5 | Minkowski | 91.23 | 0.995 | 0.916 | 0.122 | 0.954 | 0.269 |
| 3.0 | 7 | Uniform | $k$-d | 2 | 2.5 | Minkowski | 91.23 | 0.995 | 0.916 | 0.122 | 0.954 | 0.269 |
| 3.0 | 7 | Uniform | $k$-d | 128 | 2.5 | Minkowski | 91.23 | 0.995 | 0.916 | 0.122 | 0.954 | 0.269 |
| 3.0 | 7 | Uniform | Ball | 16 | - | Euclidean | 91.23 | 0.995 | 0.916 | 0.122 | 0.954 | 0.269 |
| 3.5 | 7 | Distance | Ball | 1 | 3.5 | Minkowski | 91.23 | 0.995 | 0.916 | 0.122 | 0.954 | 0.269 |
| 3.5 | 7 | Distance | Ball | 16 | 1.5 | Minkowski | 91.23 | 0.995 | 0.916 | 0.122 | 0.954 | 0.269 |
| 3.0 | 7 | Uniform | $k$-d | 128 | - | Euclidean | 91.23 | 0.995 | 0.916 | 0.122 | 0.954 | 0.268 |
| 3.0 | 7 | Distance | Ball | 4 | 2.5 | Minkowski | 91.23 | 0.995 | 0.916 | 0.122 | 0.954 | 0.268 |
| 3.0 | 7 | Distance | $k$-d | 128 | - | Manhattan | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.267 |
| 3.0 | 7 | Uniform | Ball | 32 | 2.5 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.267 |
| 3.0 | 7 | Distance | $k$-d | 1 | - | Euclidean | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.267 |
| 3.0 | 7 | Distance | $k$-d | 128 | 2.0 | Minkowski | 91.11 | 0.992 | 0.917 | 0.135 | 0.953 | 0.266 |
| 3.0 | 7 | Distance | Ball | 1 | - | Euclidean | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Uniform | $k$-d | 256 | - | Euclidean | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Uniform | $k$-d | 4 | 1.5 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Distance | Ball | 2 | 2.5 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Distance | $k$-d | 1 | 2.5 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Distance | Ball | 2 | 3.0 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Uniform | Ball | 128 | 3.5 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Uniform | Ball | 4 | 4.0 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Distance | Ball | 16 | 4.0 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |

*Continued on next page*

Table S3 – *Continued from previous page*

| $\psi$ | $k$ | Weights | Algorithm | Leaf Size | $p$ | Metric | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 7 | Uniform | Ball | 4 | - | Euclidean | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Uniform | Ball | 8 | - | Manhattan | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Distance | Ball | 2 | - | Manhattan | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Distance | Ball | 4 | 2.0 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Uniform | $k$-d | 256 | 2.0 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Uniform | Ball | 1 | 2.5 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Uniform | Ball | 256 | 2.5 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Distance | Ball | 16 | 3.5 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Uniform | Ball | 2 | 4.0 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Uniform | $k$-d | 256 | 4.0 | Minkowski | 91.17 | 0.993 | 0.916 | 0.128 | 0.953 | 0.266 |
| 3.0 | 7 | Uniform | $k$-d | 32 | - | Euclidean | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.265 |
| 3.5 | 7 | Distance | $k$-d | 1 | 2.5 | Minkowski | 91.11 | 0.992 | 0.917 | 0.135 | 0.953 | 0.265 |
| 3.0 | 7 | Distance | $k$-d | 1 | - | Manhattan | 91.11 | 0.992 | 0.917 | 0.135 | 0.953 | 0.265 |
| 3.0 | 7 | Distance | Ball | 32 | 2.5 | Minkowski | 91.11 | 0.992 | 0.917 | 0.135 | 0.953 | 0.265 |
| 3.0 | 7 | Distance | Ball | 32 | 4.0 | Minkowski | 91.11 | 0.992 | 0.917 | 0.135 | 0.953 | 0.265 |
| 3.0 | 7 | Uniform | Ball | 128 | 2.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.263 |
| 3.0 | 7 | Distance | Ball | 1 | 2.5 | Minkowski | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.263 |
| 3.0 | 7 | Distance | $k$-d | 128 | - | Euclidean | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.263 |
| 3.0 | 7 | Distance | $k$-d | 4 | 3.0 | Minkowski | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.263 |
| 3.0 | 7 | Uniform | $k$-d | 8 | 1.5 | Minkowski | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.263 |
| 3.0 | 7 | Uniform | $k$-d | 128 | 2.0 | Minkowski | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.263 |
| 3.0 | 7 | Distance | Ball | 128 | 2.5 | Minkowski | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.263 |
| 3.0 | 7 | Uniform | $k$-d | 1 | 4.0 | Minkowski | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.263 |
| 3.0 | 7 | Uniform | $k$-d | 1 | 2.5 | Minkowski | 91.23 | 0.995 | 0.915 | 0.115 | 0.954 | 0.263 |

*Continued on next page*

Table S3 – *Continued from previous page*

| $\psi$ | $k$ | Weights | Algorithm | Leaf Size | $p$ | Metric | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 7 | Uniform | Ball | 16 | 2.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.262 |
| 3.0 | 7 | Distance | $k$-d | 16 | 4.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.262 |
| 3.0 | 7 | Distance | $k$-d | 2 | 1.5 | Minkowski | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.262 |
| 3.0 | 7 | Uniform | Ball | 128 | 2.5 | Minkowski | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.262 |
| 3.0 | 7 | Uniform | Ball | 8 | 3.5 | Minkowski | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.262 |
| 3.5 | 7 | Uniform | Ball | 64 | - | Manhattan | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.262 |
| 3.5 | 7 | Distance | Ball | 256 | 3.5 | Minkowski | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.262 |
| 3.5 | 7 | Uniform | $k$-d | 16 | 2.0 | Minkowski | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.262 |
| 3.0 | 7 | Distance | $k$-d | 256 | 2.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.262 |
| 3.0 | 7 | Uniform | $k$-d | 4 | - | Manhattan | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.261 |
| 3.0 | 7 | Uniform | $k$-d | 8 | - | Manhattan | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.261 |
| 3.0 | 7 | Uniform | $k$-d | 64 | 2.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.261 |
| 3.0 | 7 | Uniform | $k$-d | 8 | 3.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.261 |
| 3.0 | 7 | Uniform | Ball | 64 | 4.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.261 |
| 3.0 | 7 | Distance | Ball | 1 | - | Manhattan | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.261 |
| 3.0 | 5 | Uniform | Ball | 32 | 2.5 | Minkowski | 90.87 | 0.987 | 0.918 | 0.154 | 0.951 | 0.261 |
| 3.0 | 5 | Uniform | $k$-d | 128 | 1.5 | Minkowski | 90.99 | 0.990 | 0.917 | 0.141 | 0.952 | 0.261 |
| 3.0 | 7 | Uniform | Ball | 2 | - | Euclidean | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Distance | Ball | 32 | - | Manhattan | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Distance | $k$-d | 4 | 1.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Uniform | Ball | 8 | 2.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Distance | Ball | 64 | 2.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Uniform | $k$-d | 4 | 2.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Distance | $k$-d | 32 | 3.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |

*Continued on next page*

Table S3 – *Continued from previous page*

| $\psi$ | $k$ | Weights | Algorithm | Leaf Size | $p$ | Metric | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 7 | Distance | Ball | 128 | - | Euclidean | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Distance | *k*-d | 256 | - | Manhattan | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Uniform | *k*-d | 64 | 1.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Uniform | Ball | 4 | 2.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Uniform | Ball | 64 | 2.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Distance | *k*-d | 2 | 2.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Distance | *k*-d | 8 | 2.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Uniform | Ball | 256 | 3.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Uniform | Ball | 32 | 4.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 7 | Uniform | *k*-d | 128 | 4.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.5 | 7 | Distance | Ball | 32 | - | Manhattan | 91.11 | 0.993 | 0.916 | 0.128 | 0.953 | 0.260 |
| 3.0 | 5 | Distance | *k*-d | 32 | - | Manhattan | 90.80 | 0.986 | 0.918 | 0.160 | 0.951 | 0.260 |
| 3.5 | 7 | Uniform | *k*-d | 4 | 1.5 | Minkowski | 91.17 | 0.995 | 0.915 | 0.115 | 0.953 | 0.259 |
| 3.5 | 7 | Uniform | Ball | 128 | 3.5 | Minkowski | 91.17 | 0.995 | 0.915 | 0.115 | 0.953 | 0.259 |
| 3.5 | 7 | Uniform | *k*-d | 256 | 3.0 | Minkowski | 91.17 | 0.995 | 0.915 | 0.115 | 0.953 | 0.259 |
| 3.0 | 7 | Uniform | *k*-d | 8 | 2.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.258 |
| 3.5 | 7 | Uniform | *k*-d | 32 | 2.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.258 |
| 3.5 | 7 | Distance | Ball | 4 | 3.5 | Minkowski | 91.17 | 0.995 | 0.915 | 0.115 | 0.953 | 0.258 |
| 3.0 | 7 | Distance | *k*-d | 256 | 4.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.257 |
| 3.5 | 7 | Uniform | *k*-d | 256 | - | Euclidean | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.257 |
| 3.0 | 7 | Uniform | *k*-d | 4 | 3.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.257 |
| 3.0 | 7 | Distance | *k*-d | 2 | - | Euclidean | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.257 |
| 3.0 | 7 | Distance | Ball | 256 | 2.0 | Minkowski | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.257 |
| 3.0 | 7 | Uniform | *k*-d | 64 | 3.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |

*Continued on next page*

Table S3 – *Continued from previous page*

| $\psi$ | $k$ | Weights | Algorithm | Leaf Size | $p$ | Metric | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 7 | Uniform | $k$-d | 64 | 3.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.0 | 7 | Distance | $k$-d | 256 | 3.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.5 | 7 | Uniform | Ball | 1 | 2.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.5 | 7 | Distance | Ball | 16 | 2.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.5 | 7 | Distance | $k$-d | 256 | 2.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.0 | 7 | Uniform | Ball | 4 | 2.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.0 | 7 | Uniform | Ball | 4 | 3.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.0 | 7 | Distance | Ball | 2 | 4.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.5 | 7 | Distance | Ball | 128 | - | Manhattan | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.5 | 7 | Distance | Ball | 2 | 2.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.5 | 7 | Distance | Ball | 4 | 2.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.5 | 7 | Distance | Ball | 64 | 3.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.0 | 7 | Distance | $k$-d | 4 | - | Manhattan | 91.17 | 0.995 | 0.915 | 0.115 | 0.953 | 0.256 |
| 3.0 | 7 | Distance | Ball | 128 | 3.5 | Minkowski | 91.17 | 0.995 | 0.915 | 0.115 | 0.953 | 0.256 |
| 3.0 | 7 | Uniform | Ball | 64 | 3.5 | Minkowski | 91.17 | 0.995 | 0.915 | 0.115 | 0.953 | 0.256 |
| 3.0 | 5 | Distance | Ball | 2 | - | Manhattan | 90.87 | 0.988 | 0.917 | 0.147 | 0.951 | 0.256 |
| 3.0 | 6 | Distance | $k$-d | 128 | 3.0 | Minkowski | 90.87 | 0.989 | 0.917 | 0.141 | 0.951 | 0.256 |
| 3.0 | 7 | Distance | $k$-d | 16 | 3.0 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.0 | 7 | Distance | $k$-d | 16 | 3.5 | Minkowski | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.256 |
| 3.0 | 7 | Uniform | $k$-d | 2 | 4.0 | Minkowski | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 7 | Distance | Ball | 256 | - | Euclidean | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 7 | Distance | Ball | 4 | - | Manhattan | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.5 | 7 | Distance | $k$-d | 8 | 1.5 | Minkowski | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.5 | 7 | Uniform | Ball | 32 | 4.0 | Minkowski | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |

*Continued on next page*

Table S3 – *Continued from previous page*

| $\psi$ | $k$ | Weights | Algorithm | Leaf Size | $p$ | Metric | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 6 | Distance | Ball | 256 | 3.0 | Minkowski | 90.80 | 0.987 | 0.918 | 0.154 | 0.951 | 0.255 |
| 3.0 | 7 | Distance | $k$-d | 8 | 2.0 | Minkowski | 90.99 | 0.991 | 0.917 | 0.135 | 0.952 | 0.255 |
| 3.0 | 7 | Distance | Ball | 2 | - | Euclidean | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 7 | Distance | Ball | 16 | - | Euclidean | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 7 | Uniform | $k$-d | 256 | - | Manhattan | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 7 | Distance | $k$-d | 32 | - | Manhattan | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 7 | Distance | $k$-d | 16 | 1.5 | Minkowski | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 7 | Distance | $k$-d | 64 | 2.0 | Minkowski | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 7 | Uniform | Ball | 32 | 3.5 | Minkowski | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 7 | Distance | Ball | 2 | 3.5 | Minkowski | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 7 | Distance | Ball | 4 | 3.5 | Minkowski | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 7 | Distance | Ball | 8 | 3.5 | Minkowski | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 7 | Distance | $k$-d | 64 | 4.0 | Minkowski | 91.05 | 0.992 | 0.916 | 0.128 | 0.953 | 0.255 |
| 3.0 | 5 | Distance | Ball | 2 | 2.0 | Minkowski | 90.68 | 0.985 | 0.918 | 0.160 | 0.950 | 0.254 |
| 3.0 | 7 | Uniform | Ball | 8 | 1.5 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.254 |
| 3.0 | 7 | Uniform | Ball | 128 | 1.5 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.254 |
| 3.5 | 7 | Uniform | Ball | 1 | 2.0 | Minkowski | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.253 |
| 3.5 | 7 | Distance | Ball | 1 | 2.5 | Minkowski | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.253 |
| 3.5 | 7 | Distance | Ball | 16 | 4.0 | Minkowski | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.253 |
| 3.5 | 7 | Distance | Ball | 128 | 2.5 | Minkowski | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.252 |
| 3.5 | 7 | Distance | $k$-d | 128 | - | Euclidean | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.252 |
| 3.5 | 7 | Distance | $k$-d | 64 | - | Manhattan | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.252 |
| 3.5 | 7 | Uniform | Ball | 1 | 3.0 | Minkowski | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.252 |
| 3.5 | 7 | Distance | Ball | 8 | 1.5 | Minkowski | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.252 |

*Continued on next page*

Table S3 – *Continued from previous page*

| $\psi$ | $k$ | Weights | Algorithm | Leaf Size | $p$ | Metric | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 7 | Uniform | Ball | 64 | 3.0 | Minkowski | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.252 |
| 3.5 | 7 | Uniform | Ball | 256 | 3.5 | Minkowski | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.252 |
| 3.5 | 7 | Distance | *k*-d | 64 | 1.5 | Minkowski | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.252 |
| 3.0 | 5 | Uniform | *k*-d | 4 | 2.5 | Minkowski | 90.68 | 0.985 | 0.918 | 0.160 | 0.950 | 0.252 |
| 3.0 | 7 | Distance | *k*-d | 1 | 3.0 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.252 |
| 3.0 | 6 | Distance | Ball | 32 | 3.5 | Minkowski | 90.74 | 0.986 | 0.918 | 0.154 | 0.951 | 0.252 |
| 3.0 | 7 | Uniform | *k*-d | 1 | 2.0 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.5 | 7 | Distance | Ball | 128 | - | Euclidean | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.5 | 7 | Uniform | Ball | 8 | 2.5 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.5 | 7 | Distance | *k*-d | 16 | 2.0 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.5 | 7 | Distance | *k*-d | 64 | 4.0 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.0 | 5 | Uniform | Ball | 4 | 3.0 | Minkowski | 90.56 | 0.983 | 0.919 | 0.167 | 0.950 | 0.251 |
| 3.0 | 7 | Distance | Ball | 256 | - | Manhattan | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.0 | 7 | Distance | *k*-d | 16 | 2.5 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.0 | 7 | Distance | Ball | 8 | 3.0 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.0 | 7 | Uniform | *k*-d | 2 | 3.0 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.0 | 7 | Uniform | *k*-d | 1 | 3.5 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.0 | 7 | Distance | *k*-d | 4 | 3.5 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.0 | 7 | Uniform | Ball | 16 | 4.0 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.0 | 7 | Distance | Ball | 1 | 4.0 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.5 | 7 | Distance | Ball | 64 | - | Euclidean | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.5 | 7 | Distance | *k*-d | 64 | - | Euclidean | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.5 | 7 | Uniform | Ball | 1 | 1.5 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.0 | 7 | Distance | Ball | 128 | - | Manhattan | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |

*Continued on next page*

Table S3 – *Continued from previous page*

| $\psi$ | $k$ | Weights | Algorithm | Leaf Size | $p$ | Metric | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 7 | Distance | Ball | 64 | 1.5 | Minkowski | 91.05 | 0.993 | 0.916 | 0.122 | 0.953 | 0.251 |
| 3.0 | 7 | Uniform | $k$-d | 1 | - | Manhattan | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.251 |
| 3.0 | 7 | Distance | Ball | 16 | 1.5 | Minkowski | 90.99 | 0.991 | 0.916 | 0.128 | 0.952 | 0.250 |
| 3.0 | 7 | Uniform | Ball | 32 | - | Manhattan | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.250 |
| 3.0 | 7 | Distance | Ball | 4 | - | Euclidean | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.250 |
| 3.0 | 7 | Distance | Ball | 32 | - | Euclidean | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.250 |
| 3.0 | 7 | Distance | Ball | 64 | 3.5 | Minkowski | 91.11 | 0.994 | 0.915 | 0.115 | 0.953 | 0.250 |
| 3.0 | 6 | Distance | $k$-d | 128 | 4.0 | Minkowski | 90.87 | 0.989 | 0.917 | 0.141 | 0.951 | 0.250 |

Table S4.: MCC and $F_1$-score performance measures, together with accuracy, sensitivity, precision and specificity, for all parameter combinations for the SVM classifier, with a MCC greater or equal to 0.15.

| $\psi$ | C | Kernel | Degree | Gamma | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 1.500 | Polynomial | 2 | 2.00 | 89.90 | 0.977 | 0.917 | 0.147 | 0.946 | 0.201 |
| 3.5 | 1.500 | Polynomial | 5 | 1.00 | 90.50 | 0.988 | 0.914 | 0.109 | 0.950 | 0.199 |
| 3.0 | 0.100 | Polynomial | 4 | 1.00 | 89.96 | 0.979 | 0.916 | 0.141 | 0.946 | 0.198 |
| 3.5 | 0.025 | Polynomial | 2 | 0.10 | 89.96 | 0.979 | 0.916 | 0.141 | 0.946 | 0.196 |
| 3.5 | 0.025 | Polynomial | 3 | 0.50 | 90.02 | 0.980 | 0.916 | 0.135 | 0.947 | 0.194 |
| 3.0 | 0.800 | Polynomial | 2 | 0.10 | 89.59 | 0.973 | 0.917 | 0.154 | 0.944 | 0.194 |
| 3.5 | 2.000 | Polynomial | 2 | 2.00 | 89.72 | 0.975 | 0.917 | 0.147 | 0.945 | 0.194 |
| 3.0 | 1.000 | Polynomial | 2 | 2.00 | 89.59 | 0.973 | 0.917 | 0.154 | 0.944 | 0.193 |
| 3.5 | 0.025 | Linear | - | - | 89.35 | 0.970 | 0.917 | 0.160 | 0.943 | 0.189 |
| 3.0 | 2.000 | Polynomial | 2 | 2.00 | 89.29 | 0.969 | 0.917 | 0.160 | 0.943 | 0.187 |
| 3.0 | 0.500 | Polynomial | 2 | 2.00 | 89.41 | 0.971 | 0.917 | 0.154 | 0.943 | 0.186 |
| 3.5 | 1.000 | Polynomial | 5 | 1.00 | 90.14 | 0.983 | 0.914 | 0.115 | 0.948 | 0.185 |
| 3.0 | 0.250 | Polynomial | 2 | 0.10 | 89.35 | 0.971 | 0.917 | 0.154 | 0.943 | 0.183 |
| 3.0 | 0.025 | Polynomial | 3 | 1.00 | 89.66 | 0.975 | 0.916 | 0.141 | 0.945 | 0.183 |
| 3.5 | 0.100 | Linear | - | - | 89.35 | 0.971 | 0.917 | 0.154 | 0.943 | 0.183 |
| 3.5 | 0.025 | Polynomial | 4 | 0.50 | 90.08 | 0.982 | 0.915 | 0.122 | 0.947 | 0.183 |
| 3.5 | 0.250 | Polynomial | 3 | 0.50 | 90.02 | 0.981 | 0.915 | 0.122 | 0.947 | 0.183 |
| 3.5 | 0.025 | Polynomial | 2 | 1.00 | 89.78 | 0.977 | 0.916 | 0.135 | 0.945 | 0.182 |
| 3.5 | 0.800 | Polynomial | 5 | 0.25 | 90.14 | 0.983 | 0.914 | 0.115 | 0.948 | 0.182 |
| 3.5 | 1.500 | Polynomial | 4 | 2.00 | 90.02 | 0.981 | 0.915 | 0.122 | 0.947 | 0.181 |
| 3.0 | 0.500 | Polynomial | 2 | 1.00 | 89.59 | 0.975 | 0.916 | 0.141 | 0.944 | 0.181 |
| 3.5 | 0.500 | Polynomial | 2 | 0.50 | 89.29 | 0.970 | 0.917 | 0.154 | 0.943 | 0.181 |

*Continued on next page*

Table S4 – *Continued from previous page*

| $\psi$ | C | Kernel | Degree | Gamma | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 0.800 | Linear | - | - | 89.59 | 0.975 | 0.916 | 0.141 | 0.944 | 0.180 |
| 3.5 | 1.500 | Polynomial | 2 | 0.10 | 89.84 | 0.979 | 0.915 | 0.128 | 0.946 | 0.180 |
| 3.0 | 1.000 | Polynomial | 2 | 0.50 | 89.53 | 0.974 | 0.916 | 0.141 | 0.944 | 0.178 |
| 3.0 | 0.800 | Polynomial | 2 | 1.00 | 89.17 | 0.969 | 0.917 | 0.154 | 0.942 | 0.177 |
| 3.5 | 0.250 | Polynomial | 2 | 2.00 | 89.66 | 0.976 | 0.915 | 0.135 | 0.945 | 0.177 |
| 3.5 | 0.100 | Polynomial | 4 | 0.25 | 90.08 | 0.983 | 0.914 | 0.115 | 0.947 | 0.176 |
| 3.0 | 1.500 | Polynomial | 2 | 1.00 | 89.35 | 0.971 | 0.916 | 0.147 | 0.943 | 0.176 |
| 3.0 | 1.000 | Polynomial | 2 | 0.10 | 89.59 | 0.975 | 0.915 | 0.135 | 0.944 | 0.176 |
| 3.0 | 2.000 | Polynomial | 3 | 1.00 | 89.78 | 0.978 | 0.915 | 0.128 | 0.945 | 0.176 |
| 3.0 | 0.025 | Polynomial | 3 | 0.50 | 89.78 | 0.978 | 0.915 | 0.128 | 0.945 | 0.176 |
| 3.5 | 0.500 | Polynomial | 5 | 0.25 | 90.14 | 0.984 | 0.914 | 0.109 | 0.948 | 0.176 |
| 3.5 | 0.500 | Polynomial | 2 | 0.10 | 89.59 | 0.975 | 0.915 | 0.135 | 0.944 | 0.174 |
| 3.5 | 0.800 | Polynomial | 2 | 0.10 | 89.59 | 0.975 | 0.915 | 0.135 | 0.944 | 0.174 |
| 3.0 | 0.500 | Polynomial | 2 | 0.25 | 89.41 | 0.973 | 0.916 | 0.141 | 0.943 | 0.174 |
| 3.0 | 1.000 | Polynomial | 2 | 0.25 | 89.41 | 0.973 | 0.916 | 0.141 | 0.943 | 0.174 |
| 3.5 | 0.500 | Polynomial | 2 | 2.00 | 89.72 | 0.977 | 0.915 | 0.128 | 0.945 | 0.174 |
| 3.5 | 2.000 | Linear | - | - | 89.23 | 0.970 | 0.916 | 0.147 | 0.942 | 0.174 |
| 3.0 | 1.500 | Polynomial | 3 | 0.10 | 89.72 | 0.977 | 0.915 | 0.128 | 0.945 | 0.174 |
| 3.0 | 2.000 | Polynomial | 3 | 0.10 | 89.41 | 0.973 | 0.916 | 0.141 | 0.943 | 0.174 |
| 3.0 | 0.100 | Polynomial | 3 | 2.00 | 89.72 | 0.977 | 0.915 | 0.128 | 0.945 | 0.173 |
| 3.0 | 0.250 | Linear | - | - | 89.23 | 0.970 | 0.916 | 0.147 | 0.942 | 0.173 |
| 3.5 | 1.500 | Linear | - | - | 89.35 | 0.972 | 0.916 | 0.141 | 0.943 | 0.172 |
| 3.0 | 0.025 | Polynomial | 3 | 0.10 | 89.53 | 0.975 | 0.915 | 0.135 | 0.944 | 0.172 |
| 3.5 | 1.500 | Polynomial | 4 | 0.10 | 90.08 | 0.983 | 0.914 | 0.109 | 0.947 | 0.172 |

*Continued on next page*

Table S4 – *Continued from previous page*

| $\psi$ | C | Kernel | Degree | Gamma | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 0.500 | Polynomial | 4 | 2.00 | 90.08 | 0.983 | 0.914 | 0.109 | 0.947 | 0.171 |
| 3.0 | 0.025 | Polynomial | 2 | 0.25 | 89.35 | 0.972 | 0.916 | 0.141 | 0.943 | 0.171 |
| 3.5 | 1.000 | Polynomial | 4 | 0.25 | 90.08 | 0.983 | 0.914 | 0.109 | 0.947 | 0.171 |
| 3.5 | 0.250 | Polynomial | 3 | 2.00 | 89.66 | 0.977 | 0.915 | 0.128 | 0.945 | 0.171 |
| 3.0 | 1.000 | Linear | - | - | 88.99 | 0.967 | 0.916 | 0.154 | 0.941 | 0.170 |
| 3.0 | 0.800 | Polynomial | 3 | 1.00 | 89.47 | 0.974 | 0.915 | 0.135 | 0.944 | 0.170 |
| 3.5 | 0.025 | Polynomial | 3 | 2.00 | 89.90 | 0.981 | 0.914 | 0.115 | 0.946 | 0.170 |
| 3.0 | 0.250 | Polynomial | 2 | 2.00 | 88.81 | 0.964 | 0.917 | 0.160 | 0.940 | 0.170 |
| 3.5 | 2.000 | Polynomial | 2 | 0.10 | 89.59 | 0.976 | 0.915 | 0.128 | 0.944 | 0.170 |
| 3.5 | 0.250 | Polynomial | 4 | 1.00 | 89.90 | 0.981 | 0.914 | 0.115 | 0.946 | 0.170 |
| 3.5 | 2.000 | Polynomial | 2 | 0.50 | 89.29 | 0.971 | 0.916 | 0.141 | 0.943 | 0.169 |
| 3.5 | 0.025 | Polynomial | 4 | 2.00 | 89.90 | 0.981 | 0.914 | 0.115 | 0.946 | 0.169 |
| 3.5 | 0.100 | Polynomial | 2 | 0.25 | 89.59 | 0.976 | 0.915 | 0.128 | 0.944 | 0.169 |
| 3.5 | 1.000 | Polynomial | 5 | 0.25 | 90.02 | 0.983 | 0.914 | 0.109 | 0.947 | 0.169 |
| 3.0 | 1.500 | Polynomial | 2 | 0.10 | 89.11 | 0.969 | 0.916 | 0.147 | 0.942 | 0.168 |
| 3.5 | 1.000 | Polynomial | 3 | 0.25 | 89.72 | 0.978 | 0.914 | 0.122 | 0.945 | 0.168 |
| 3.0 | 0.800 | Polynomial | 5 | 0.25 | 89.90 | 0.981 | 0.914 | 0.115 | 0.946 | 0.168 |
| 3.5 | 0.250 | Polynomial | 3 | 1.00 | 89.59 | 0.976 | 0.915 | 0.128 | 0.944 | 0.168 |
| 3.5 | 1.000 | Polynomial | 5 | 0.50 | 90.02 | 0.983 | 0.914 | 0.109 | 0.947 | 0.168 |
| 3.0 | 0.025 | Linear | - | - | 88.93 | 0.966 | 0.916 | 0.154 | 0.940 | 0.168 |
| 3.0 | 0.500 | Polynomial | 4 | 0.10 | 89.84 | 0.980 | 0.914 | 0.115 | 0.946 | 0.167 |
| 3.0 | 0.100 | Linear | - | - | 89.05 | 0.968 | 0.916 | 0.147 | 0.941 | 0.167 |
| 3.5 | 0.500 | Polynomial | 3 | 0.50 | 89.84 | 0.980 | 0.914 | 0.115 | 0.946 | 0.167 |
| 3.5 | 0.250 | Polynomial | 3 | 0.10 | 89.96 | 0.982 | 0.914 | 0.109 | 0.947 | 0.167 |

*Continued on next page*

Table S4 – *Continued from previous page*

| $\psi$ | C | Kernel | Degree | Gamma | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 0.025 | Polynomial | 2 | 0.10 | 89.23 | 0.971 | 0.916 | 0.141 | 0.942 | 0.167 |
| 3.0 | 0.025 | Polynomial | 5 | 0.10 | 89.96 | 0.982 | 0.914 | 0.109 | 0.947 | 0.166 |
| 3.0 | 0.100 | Polynomial | 2 | 1.00 | 89.53 | 0.975 | 0.915 | 0.128 | 0.944 | 0.166 |
| 3.0 | 0.100 | Polynomial | 4 | 0.25 | 89.84 | 0.980 | 0.914 | 0.115 | 0.946 | 0.165 |
| 3.5 | 0.025 | Polynomial | 3 | 0.10 | 89.78 | 0.979 | 0.914 | 0.115 | 0.945 | 0.165 |
| 3.0 | 2.000 | Linear | - | - | 88.81 | 0.965 | 0.916 | 0.154 | 0.940 | 0.164 |
| 3.5 | 1.000 | Polynomial | 4 | 0.10 | 89.96 | 0.982 | 0.914 | 0.109 | 0.947 | 0.164 |
| 3.5 | 1.000 | Polynomial | 5 | 0.10 | 89.90 | 0.981 | 0.914 | 0.109 | 0.946 | 0.164 |
| 3.5 | 0.500 | Linear | - | - | 89.17 | 0.970 | 0.916 | 0.141 | 0.942 | 0.164 |
| 3.5 | 0.100 | Polynomial | 3 | 1.00 | 89.66 | 0.977 | 0.914 | 0.122 | 0.945 | 0.163 |
| 3.5 | 0.800 | Polynomial | 3 | 0.25 | 89.66 | 0.977 | 0.914 | 0.122 | 0.945 | 0.163 |
| 3.0 | 1.000 | Polynomial | 2 | 1.00 | 89.11 | 0.969 | 0.915 | 0.141 | 0.942 | 0.163 |
| 3.0 | 1.000 | Polynomial | 3 | 0.10 | 89.47 | 0.975 | 0.915 | 0.128 | 0.944 | 0.163 |
| 3.0 | 0.500 | Polynomial | 3 | 0.10 | 89.29 | 0.972 | 0.915 | 0.135 | 0.943 | 0.163 |
| 3.0 | 0.800 | Polynomial | 2 | 0.25 | 89.29 | 0.972 | 0.915 | 0.135 | 0.943 | 0.163 |
| 3.0 | 2.000 | Polynomial | 2 | 0.50 | 89.11 | 0.969 | 0.915 | 0.141 | 0.942 | 0.163 |
| 3.0 | 0.100 | Polynomial | 2 | 0.50 | 88.93 | 0.967 | 0.916 | 0.147 | 0.941 | 0.162 |
| 3.0 | 1.000 | Polynomial | 3 | 2.00 | 89.29 | 0.972 | 0.915 | 0.135 | 0.943 | 0.162 |
| 3.5 | 0.500 | Polynomial | 3 | 2.00 | 89.72 | 0.979 | 0.914 | 0.115 | 0.945 | 0.162 |
| 3.0 | 0.800 | Polynomial | 2 | 0.50 | 89.29 | 0.972 | 0.915 | 0.135 | 0.943 | 0.161 |
| 3.5 | 0.025 | Polynomial | 2 | 0.50 | 89.41 | 0.974 | 0.915 | 0.128 | 0.943 | 0.161 |
| 3.5 | 1.000 | Polynomial | 3 | 0.50 | 89.72 | 0.979 | 0.914 | 0.115 | 0.945 | 0.161 |
| 3.5 | 1.000 | Polynomial | 2 | 0.25 | 89.41 | 0.974 | 0.915 | 0.128 | 0.943 | 0.161 |
| 3.5 | 0.800 | Polynomial | 2 | 0.25 | 89.59 | 0.977 | 0.914 | 0.122 | 0.944 | 0.161 |

*Continued on next page*

Table S4 – *Continued from previous page*

| $\psi$ | C | Kernel | Degree | Gamma | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 0.250 | Polynomial | 3 | 0.25 | 89.53 | 0.976 | 0.914 | 0.122 | 0.944 | 0.161 |
| 3.5 | 0.250 | Polynomial | 2 | 1.00 | 89.35 | 0.973 | 0.915 | 0.128 | 0.943 | 0.160 |
| 3.0 | 0.500 | Polynomial | 4 | 0.25 | 89.53 | 0.976 | 0.914 | 0.122 | 0.944 | 0.160 |
| 3.5 | 1.000 | Polynomial | 2 | 2.00 | 89.35 | 0.973 | 0.915 | 0.128 | 0.943 | 0.160 |
| 3.5 | 1.500 | Polynomial | 2 | 2.00 | 89.35 | 0.973 | 0.915 | 0.128 | 0.943 | 0.159 |
| 3.5 | 0.100 | Polynomial | 4 | 2.00 | 89.84 | 0.981 | 0.914 | 0.109 | 0.946 | 0.159 |
| 3.5 | 0.500 | Polynomial | 5 | 2.00 | 90.08 | 0.985 | 0.913 | 0.096 | 0.947 | 0.159 |
| 3.0 | 2.000 | Polynomial | 3 | 2.00 | 89.47 | 0.975 | 0.914 | 0.122 | 0.944 | 0.159 |
| 3.5 | 0.025 | Polynomial | 5 | 0.10 | 90.08 | 0.985 | 0.913 | 0.096 | 0.947 | 0.158 |
| 3.0 | 0.500 | Polynomial | 3 | 0.25 | 89.66 | 0.978 | 0.914 | 0.115 | 0.945 | 0.158 |
| 3.5 | 0.500 | Polynomial | 5 | 0.10 | 89.84 | 0.981 | 0.914 | 0.109 | 0.946 | 0.158 |
| 3.0 | 0.800 | Polynomial | 3 | 2.00 | 89.66 | 0.978 | 0.914 | 0.115 | 0.945 | 0.158 |
| 3.5 | 0.800 | Polynomial | 3 | 1.00 | 89.66 | 0.978 | 0.914 | 0.115 | 0.945 | 0.158 |
| 3.0 | 0.250 | Polynomial | 3 | 1.00 | 89.29 | 0.973 | 0.915 | 0.128 | 0.943 | 0.158 |
| 3.5 | 0.025 | Polynomial | 3 | 1.00 | 89.78 | 0.980 | 0.913 | 0.109 | 0.946 | 0.158 |
| 3.5 | 0.800 | Polynomial | 4 | 0.50 | 89.96 | 0.983 | 0.913 | 0.103 | 0.947 | 0.158 |
| 3.5 | 0.025 | Polynomial | 2 | 2.00 | 89.66 | 0.978 | 0.914 | 0.115 | 0.945 | 0.158 |
| 3.5 | 0.025 | Polynomial | 4 | 0.10 | 89.78 | 0.980 | 0.913 | 0.109 | 0.946 | 0.157 |
| 3.5 | 0.100 | Polynomial | 5 | 1.00 | 89.90 | 0.982 | 0.913 | 0.103 | 0.946 | 0.157 |
| 3.5 | 2.000 | Polynomial | 3 | 1.00 | 89.78 | 0.980 | 0.913 | 0.109 | 0.946 | 0.157 |
| 3.5 | 0.800 | Polynomial | 2 | 2.00 | 89.29 | 0.973 | 0.915 | 0.128 | 0.943 | 0.156 |
| 3.0 | 2.000 | Polynomial | 2 | 0.10 | 89.11 | 0.970 | 0.915 | 0.135 | 0.942 | 0.156 |
| 3.5 | 1.500 | Polynomial | 3 | 0.50 | 89.59 | 0.977 | 0.914 | 0.115 | 0.944 | 0.156 |
| 3.0 | 0.500 | Polynomial | 2 | 0.50 | 88.93 | 0.967 | 0.915 | 0.141 | 0.941 | 0.156 |

*Continued on next page*

Table S4 – *Continued from previous page*

| $\psi$ | C | Kernel | Degree | Gamma | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 0.100 | Polynomial | 3 | 2.00 | 89.72 | 0.979 | 0.913 | 0.109 | 0.945 | 0.156 |
| 3.5 | 0.500 | Polynomial | 3 | 0.25 | 89.59 | 0.977 | 0.914 | 0.115 | 0.944 | 0.155 |
| 3.0 | 0.800 | Linear | - | - | 88.87 | 0.967 | 0.915 | 0.141 | 0.940 | 0.155 |
| 3.5 | 2.000 | Polynomial | 5 | 0.10 | 89.72 | 0.979 | 0.913 | 0.109 | 0.945 | 0.155 |
| 3.5 | 0.500 | Polynomial | 3 | 0.10 | 89.41 | 0.975 | 0.914 | 0.122 | 0.943 | 0.155 |
| 3.0 | 0.100 | Polynomial | 3 | 0.10 | 89.41 | 0.975 | 0.914 | 0.122 | 0.943 | 0.155 |
| 3.0 | 0.800 | Polynomial | 3 | 0.50 | 89.05 | 0.969 | 0.915 | 0.135 | 0.941 | 0.155 |
| 3.5 | 0.250 | Polynomial | 4 | 2.00 | 89.78 | 0.980 | 0.913 | 0.109 | 0.946 | 0.155 |
| 3.5 | 2.000 | Polynomial | 3 | 0.10 | 89.72 | 0.979 | 0.913 | 0.109 | 0.945 | 0.155 |
| 3.0 | 0.100 | Polynomial | 3 | 0.50 | 89.23 | 0.972 | 0.915 | 0.128 | 0.942 | 0.155 |
| 3.5 | 0.100 | Polynomial | 5 | 0.25 | 89.84 | 0.981 | 0.913 | 0.103 | 0.946 | 0.155 |
| 3.5 | 0.250 | Polynomial | 4 | 0.10 | 89.90 | 0.982 | 0.913 | 0.103 | 0.946 | 0.154 |
| 3.0 | 0.025 | Polynomial | 5 | 2.00 | 89.72 | 0.979 | 0.913 | 0.109 | 0.945 | 0.154 |
| 3.5 | 2.000 | Polynomial | 2 | 0.25 | 89.23 | 0.972 | 0.915 | 0.128 | 0.942 | 0.154 |
| 3.0 | 0.800 | Polynomial | 3 | 0.10 | 89.23 | 0.972 | 0.915 | 0.128 | 0.942 | 0.154 |
| 3.5 | 0.800 | Polynomial | 2 | 1.00 | 89.41 | 0.975 | 0.914 | 0.122 | 0.943 | 0.154 |
| 3.0 | 1.000 | Polynomial | 3 | 0.25 | 89.23 | 0.972 | 0.915 | 0.128 | 0.942 | 0.154 |
| 3.0 | 0.100 | Polynomial | 3 | 1.00 | 89.41 | 0.975 | 0.914 | 0.122 | 0.943 | 0.154 |
| 3.5 | 0.100 | Polynomial | 2 | 0.50 | 89.41 | 0.975 | 0.914 | 0.122 | 0.943 | 0.153 |
| 3.0 | 2.000 | Polynomial | 5 | 0.50 | 89.66 | 0.979 | 0.913 | 0.109 | 0.945 | 0.153 |
| 3.5 | 0.250 | Polynomial | 5 | 0.10 | 89.84 | 0.981 | 0.913 | 0.103 | 0.946 | 0.153 |
| 3.0 | 1.500 | Polynomial | 2 | 0.25 | 89.05 | 0.969 | 0.915 | 0.135 | 0.941 | 0.153 |
| 3.5 | 0.250 | Polynomial | 5 | 0.25 | 89.72 | 0.979 | 0.913 | 0.109 | 0.945 | 0.153 |
| 3.5 | 0.800 | Polynomial | 5 | 1.00 | 89.96 | 0.983 | 0.913 | 0.096 | 0.947 | 0.153 |

*Continued on next page*

Table S4 – *Continued from previous page*

| $\psi$ | C | Kernel | Degree | Gamma | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 0.250 | Polynomial | 2 | 0.50 | 89.23 | 0.972 | 0.915 | 0.128 | 0.942 | 0.153 |
| 3.0 | 2.000 | Polynomial | 3 | 0.50 | 89.41 | 0.975 | 0.914 | 0.122 | 0.943 | 0.153 |
| 3.0 | 0.025 | Polynomial | 4 | 0.50 | 89.53 | 0.977 | 0.914 | 0.115 | 0.944 | 0.152 |
| 3.5 | 2.000 | Polynomial | 4 | 0.10 | 89.66 | 0.979 | 0.913 | 0.109 | 0.945 | 0.152 |
| 3.5 | 1.000 | Linear | - | - | 88.99 | 0.969 | 0.915 | 0.135 | 0.941 | 0.152 |
| 3.0 | 0.250 | Polynomial | 3 | 0.25 | 89.35 | 0.974 | 0.914 | 0.122 | 0.943 | 0.152 |
| 3.0 | 2.000 | Polynomial | 2 | 1.00 | 88.99 | 0.969 | 0.915 | 0.135 | 0.941 | 0.152 |
| 3.5 | 0.100 | Polynomial | 2 | 1.00 | 89.35 | 0.974 | 0.914 | 0.122 | 0.943 | 0.152 |
| 3.0 | 0.500 | Polynomial | 3 | 0.50 | 89.35 | 0.974 | 0.914 | 0.122 | 0.943 | 0.151 |
| 3.0 | 0.250 | Polynomial | 5 | 0.50 | 89.66 | 0.979 | 0.913 | 0.109 | 0.945 | 0.151 |
| 3.5 | 2.000 | Polynomial | 4 | 0.25 | 90.08 | 0.985 | 0.912 | 0.090 | 0.947 | 0.151 |
| 3.5 | 0.100 | Polynomial | 4 | 1.00 | 90.02 | 0.985 | 0.912 | 0.090 | 0.947 | 0.151 |
| 3.0 | 1.000 | Polynomial | 4 | 1.00 | 89.47 | 0.976 | 0.914 | 0.115 | 0.944 | 0.150 |
| 3.0 | 1.000 | Polynomial | 4 | 2.00 | 89.66 | 0.979 | 0.913 | 0.109 | 0.945 | 0.150 |

Table S5.: MCC and $F_1$-score performance measures, together with accuracy, sensitivity, precision and specificity, for all parameter combinations for the RDF classifier, with a MCC greater or equal to 0.20. IG stands for information gain.

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 13 | Gini | $\sqrt{n}$ | 8 | 91.35 | 0.995 | 0.917 | 0.135 | 0.954 | 0.287 |
| 3.0 | 7 | IG | $n$ | $\infty$ | 91.11 | 0.989 | 0.919 | 0.160 | 0.953 | 0.286 |
| 3.0 | 11 | IG | $n$ | $\infty$ | 91.11 | 0.989 | 0.919 | 0.160 | 0.953 | 0.281 |
| 3.0 | 15 | Gini | 1 | $\infty$ | 91.35 | 0.997 | 0.915 | 0.115 | 0.954 | 0.280 |
| 3.0 | 15 | IG | $n$ | 8 | 91.17 | 0.991 | 0.918 | 0.147 | 0.953 | 0.279 |
| 3.5 | 18 | IG | $\log_2 n$ | $\infty$ | 91.35 | 0.997 | 0.915 | 0.109 | 0.954 | 0.278 |
| 3.0 | 13 | IG | $n$ | 8 | 91.11 | 0.991 | 0.918 | 0.147 | 0.953 | 0.272 |
| 3.0 | 18 | Gini | $\log_2 n$ | $\infty$ | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.268 |
| 3.0 | 19 | IG | $\sqrt{n}$ | 8 | 91.29 | 0.997 | 0.914 | 0.103 | 0.954 | 0.267 |
| 3.0 | 18 | Gini | 1 | $\infty$ | 91.29 | 0.997 | 0.914 | 0.103 | 0.954 | 0.266 |
| 3.0 | 12 | IG | $\sqrt{n}$ | 8 | 91.23 | 0.996 | 0.915 | 0.109 | 0.954 | 0.265 |
| 3.5 | 19 | IG | $\sqrt{n}$ | 8 | 91.23 | 0.995 | 0.915 | 0.115 | 0.954 | 0.265 |
| 3.5 | 19 | Gini | 1 | $\infty$ | 91.29 | 0.998 | 0.914 | 0.096 | 0.954 | 0.264 |
| 3.5 | 18 | IG | 1 | $\infty$ | 91.29 | 0.998 | 0.914 | 0.096 | 0.954 | 0.264 |
| 3.0 | 16 | IG | 1 | $\infty$ | 91.23 | 0.996 | 0.915 | 0.109 | 0.954 | 0.262 |
| 3.5 | 15 | Gini | $\log_2 n$ | $\infty$ | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.262 |
| 3.5 | 19 | IG | $\log_2 n$ | $\infty$ | 91.23 | 0.996 | 0.915 | 0.109 | 0.954 | 0.262 |
| 3.0 | 13 | Gini | $\log_2 n$ | $\infty$ | 91.17 | 0.995 | 0.915 | 0.115 | 0.953 | 0.261 |
| 3.0 | 14 | Gini | $\log_2 n$ | $\infty$ | 90.99 | 0.990 | 0.917 | 0.141 | 0.952 | 0.260 |
| 3.0 | 13 | Gini | $\sqrt{n}$ | 8 | 91.17 | 0.994 | 0.916 | 0.122 | 0.953 | 0.259 |
| 3.0 | 14 | IG | $n$ | $\infty$ | 90.62 | 0.983 | 0.919 | 0.173 | 0.950 | 0.258 |
| 3.0 | 19 | IG | $n$ | 5 | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.258 |

*Continued on next page*

Table S5 – *Continued from previous page*

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 16 | Gini | 1 | $\infty$ | 91.23 | 0.997 | 0.914 | 0.096 | 0.954 | 0.256 |
| 3.0 | 9 | IG | $n$ | 5 | 90.99 | 0.991 | 0.917 | 0.135 | 0.952 | 0.256 |
| 3.0 | 5 | IG | $n$ | $\infty$ | 90.38 | 0.979 | 0.920 | 0.186 | 0.949 | 0.255 |
| 3.5 | 17 | IG | $\log_2 n$ | $\infty$ | 91.17 | 0.995 | 0.915 | 0.109 | 0.953 | 0.255 |
| 3.0 | 10 | IG | $\log_2 n$ | $\infty$ | 91.11 | 0.993 | 0.916 | 0.122 | 0.953 | 0.255 |
| 3.0 | 17 | IG | $n$ | 8 | 90.87 | 0.988 | 0.917 | 0.147 | 0.951 | 0.255 |
| 3.0 | 19 | IG | 1 | $\infty$ | 91.23 | 0.998 | 0.913 | 0.090 | 0.954 | 0.254 |
| 3.5 | 8 | IG | 1 | $\infty$ | 90.93 | 0.991 | 0.915 | 0.122 | 0.952 | 0.253 |
| 3.5 | 16 | IG | $\sqrt{n}$ | 5 | 91.17 | 0.996 | 0.914 | 0.103 | 0.953 | 0.252 |
| 3.0 | 17 | Gini | $n$ | 8 | 90.80 | 0.987 | 0.917 | 0.147 | 0.951 | 0.252 |
| 3.0 | 15 | IG | 1 | $\infty$ | 91.23 | 0.998 | 0.913 | 0.090 | 0.954 | 0.251 |
| 3.0 | 17 | IG | $n$ | $\infty$ | 90.87 | 0.989 | 0.917 | 0.141 | 0.951 | 0.250 |
| 3.0 | 10 | IG | $n$ | $\infty$ | 90.56 | 0.983 | 0.919 | 0.167 | 0.950 | 0.250 |
| 3.0 | 19 | IG | $\log_2 n$ | $\infty$ | 91.17 | 0.996 | 0.914 | 0.103 | 0.953 | 0.250 |
| 3.0 | 9 | Gini | $n$ | 5 | 90.99 | 0.991 | 0.916 | 0.128 | 0.952 | 0.250 |
| 3.0 | 6 | IG | $n$ | $\infty$ | 89.29 | 0.961 | 0.924 | 0.237 | 0.942 | 0.250 |
| 3.0 | 15 | Gini | $\log_2 n$ | $\infty$ | 91.11 | 0.995 | 0.915 | 0.109 | 0.953 | 0.249 |
| 3.0 | 19 | Gini | $n$ | $\infty$ | 90.68 | 0.985 | 0.918 | 0.154 | 0.950 | 0.249 |
| 3.5 | 17 | Gini | 1 | $\infty$ | 91.17 | 0.997 | 0.914 | 0.096 | 0.953 | 0.249 |
| 3.0 | 14 | IG | $\log_2 n$ | $\infty$ | 91.11 | 0.995 | 0.915 | 0.109 | 0.953 | 0.248 |
| 3.0 | 19 | Gini | $\log_2 n$ | $\infty$ | 91.11 | 0.995 | 0.915 | 0.109 | 0.953 | 0.248 |
| 3.0 | 15 | IG | $\sqrt{n}$ | $\infty$ | 91.11 | 0.995 | 0.915 | 0.109 | 0.953 | 0.248 |
| 3.0 | 18 | IG | $\log_2 n$ | $\infty$ | 91.11 | 0.995 | 0.915 | 0.109 | 0.953 | 0.248 |
| 3.5 | 14 | Gini | $\sqrt{n}$ | 8 | 91.05 | 0.993 | 0.915 | 0.115 | 0.953 | 0.248 |

*Continued on next page*

Table S5 – *Continued from previous page*

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 16 | Gini | $\sqrt{n}$ | $\infty$ | 90.87 | 0.989 | 0.916 | 0.135 | 0.951 | 0.248 |
| 3.0 | 6 | Gini | $n$ | 8 | 89.96 | 0.973 | 0.921 | 0.199 | 0.946 | 0.248 |
| 3.5 | 15 | IG | 1 | $\infty$ | 91.17 | 0.997 | 0.914 | 0.096 | 0.953 | 0.248 |
| 3.5 | 16 | IG | 1 | $\infty$ | 91.17 | 0.997 | 0.914 | 0.096 | 0.953 | 0.248 |
| 3.0 | 18 | IG | $n$ | 8 | 90.87 | 0.990 | 0.916 | 0.128 | 0.952 | 0.247 |
| 3.0 | 16 | IG | $n$ | $\infty$ | 90.80 | 0.988 | 0.917 | 0.141 | 0.951 | 0.247 |
| 3.5 | 17 | IG | 1 | $\infty$ | 91.17 | 0.997 | 0.913 | 0.090 | 0.953 | 0.246 |
| 3.5 | 16 | Gini | $\sqrt{n}$ | 8 | 91.05 | 0.994 | 0.915 | 0.109 | 0.953 | 0.245 |
| 3.0 | 13 | IG | $n$ | $\infty$ | 90.93 | 0.991 | 0.916 | 0.128 | 0.952 | 0.245 |
| 3.5 | 15 | IG | $n$ | 8 | 90.99 | 0.992 | 0.916 | 0.122 | 0.952 | 0.244 |
| 3.0 | 9 | Gini | $\sqrt{n}$ | $\infty$ | 90.74 | 0.987 | 0.917 | 0.141 | 0.951 | 0.243 |
| 3.5 | 13 | Gini | $\log_2 n$ | $\infty$ | 91.11 | 0.995 | 0.914 | 0.103 | 0.953 | 0.243 |
| 3.0 | 15 | Gini | $n$ | 5 | 90.93 | 0.991 | 0.916 | 0.128 | 0.952 | 0.243 |
| 3.5 | 18 | Gini | 1 | $\infty$ | 91.11 | 0.995 | 0.914 | 0.103 | 0.953 | 0.243 |
| 3.5 | 19 | Gini | $\log_2 n$ | 8 | 91.17 | 0.998 | 0.913 | 0.083 | 0.953 | 0.242 |
| 3.5 | 13 | IG | 1 | $\infty$ | 91.17 | 0.998 | 0.913 | 0.083 | 0.953 | 0.242 |
| 3.0 | 8 | IG | $\sqrt{n}$ | $\infty$ | 90.50 | 0.983 | 0.918 | 0.160 | 0.949 | 0.242 |
| 3.0 | 16 | Gini | $\log_2 n$ | $\infty$ | 91.05 | 0.994 | 0.915 | 0.109 | 0.953 | 0.242 |
| 3.0 | 6 | IG | $n$ | 8 | 89.66 | 0.969 | 0.921 | 0.205 | 0.944 | 0.241 |
| 3.0 | 10 | Gini | $n$ | 8 | 90.50 | 0.983 | 0.918 | 0.160 | 0.949 | 0.241 |
| 3.5 | 9 | IG | $\sqrt{n}$ | $\infty$ | 90.93 | 0.991 | 0.915 | 0.122 | 0.952 | 0.240 |
| 3.5 | 11 | IG | $\log_2 n$ | $\infty$ | 90.99 | 0.993 | 0.915 | 0.115 | 0.952 | 0.240 |
| 3.5 | 11 | IG | $\sqrt{n}$ | $\infty$ | 90.99 | 0.993 | 0.915 | 0.109 | 0.952 | 0.240 |
| 3.5 | 11 | Gini | $\sqrt{n}$ | $\infty$ | 90.93 | 0.992 | 0.915 | 0.115 | 0.952 | 0.240 |

*Continued on next page*

Table S5 – *Continued from previous page*

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 8 | Gini | $\sqrt{n}$ | 5 | 91.05 | 0.995 | 0.914 | 0.103 | 0.953 | 0.239 |
| 3.0 | 10 | Gini | $n$ | 5 | 90.74 | 0.988 | 0.916 | 0.135 | 0.951 | 0.239 |
| 3.5 | 16 | Gini | $\sqrt{n}$ | $\infty$ | 90.93 | 0.991 | 0.915 | 0.122 | 0.952 | 0.238 |
| 3.0 | 17 | Gini | $\sqrt{n}$ | $\infty$ | 90.87 | 0.991 | 0.915 | 0.122 | 0.952 | 0.238 |
| 3.5 | 15 | IG | $\sqrt{n}$ | $\infty$ | 91.05 | 0.995 | 0.914 | 0.103 | 0.953 | 0.238 |
| 3.0 | 14 | Gini | $\sqrt{n}$ | 5 | 91.05 | 0.995 | 0.914 | 0.103 | 0.953 | 0.238 |
| 3.0 | 17 | Gini | $\sqrt{n}$ | 5 | 91.11 | 0.997 | 0.913 | 0.090 | 0.953 | 0.237 |
| 3.5 | 15 | Gini | 1 | $\infty$ | 91.11 | 0.997 | 0.913 | 0.090 | 0.953 | 0.237 |
| 3.0 | 19 | Gini | $n$ | 5 | 90.93 | 0.992 | 0.915 | 0.115 | 0.952 | 0.237 |
| 3.0 | 18 | Gini | $\sqrt{n}$ | 5 | 91.11 | 0.997 | 0.913 | 0.090 | 0.953 | 0.237 |
| 3.0 | 9 | IG | $n$ | $\infty$ | 90.50 | 0.983 | 0.918 | 0.154 | 0.949 | 0.237 |
| 3.0 | 13 | IG | $\sqrt{n}$ | 5 | 91.11 | 0.997 | 0.913 | 0.090 | 0.953 | 0.236 |
| 3.0 | 17 | Gini | $\log_2 n$ | $\infty$ | 90.99 | 0.993 | 0.915 | 0.109 | 0.952 | 0.236 |
| 3.5 | 15 | Gini | $\sqrt{n}$ | 5 | 90.99 | 0.993 | 0.915 | 0.109 | 0.952 | 0.236 |
| 3.0 | 10 | IG | $n$ | 5 | 90.87 | 0.991 | 0.915 | 0.122 | 0.952 | 0.236 |
| 3.5 | 13 | Gini | $\sqrt{n}$ | $\infty$ | 90.99 | 0.993 | 0.915 | 0.109 | 0.952 | 0.236 |
| 3.0 | 10 | Gini | $\sqrt{n}$ | 8 | 90.87 | 0.991 | 0.915 | 0.122 | 0.952 | 0.235 |
| 3.5 | 13 | IG | $\sqrt{n}$ | $\infty$ | 90.87 | 0.991 | 0.915 | 0.122 | 0.952 | 0.235 |
| 3.0 | 16 | IG | $\sqrt{n}$ | $\infty$ | 90.80 | 0.989 | 0.916 | 0.128 | 0.951 | 0.235 |
| 3.5 | 17 | IG | $\sqrt{n}$ | 8 | 91.05 | 0.995 | 0.914 | 0.096 | 0.953 | 0.235 |
| 3.0 | 19 | IG | $\sqrt{n}$ | $\infty$ | 91.05 | 0.995 | 0.914 | 0.096 | 0.953 | 0.234 |
| 3.5 | 14 | Gini | $\sqrt{n}$ | 5 | 91.05 | 0.995 | 0.914 | 0.096 | 0.953 | 0.234 |
| 3.5 | 18 | Gini | $\sqrt{n}$ | 5 | 91.05 | 0.995 | 0.914 | 0.096 | 0.953 | 0.234 |
| 3.5 | 14 | IG | $\sqrt{n}$ | $\infty$ | 90.99 | 0.994 | 0.914 | 0.103 | 0.952 | 0.234 |

*Continued on next page*

Table S5 – *Continued from previous page*

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 5 | Gini | $n$ | 5 | 90.44 | 0.983 | 0.918 | 0.154 | 0.949 | 0.233 |
| 3.5 | 10 | Gini | 1 | $\infty$ | 91.05 | 0.996 | 0.913 | 0.090 | 0.953 | 0.233 |
| 3.5 | 9 | IG | $n$ | 5 | 90.87 | 0.991 | 0.915 | 0.122 | 0.952 | 0.233 |
| 3.0 | 18 | IG | $n$ | $\infty$ | 90.68 | 0.987 | 0.916 | 0.135 | 0.950 | 0.233 |
| 3.0 | 13 | IG | $n$ | 5 | 90.99 | 0.994 | 0.914 | 0.103 | 0.952 | 0.232 |
| 3.5 | 16 | IG | $n$ | 5 | 90.99 | 0.994 | 0.914 | 0.103 | 0.952 | 0.232 |
| 3.0 | 10 | IG | 1 | $\infty$ | 90.99 | 0.994 | 0.914 | 0.103 | 0.952 | 0.232 |
| 3.0 | 19 | IG | $n$ | 8 | 90.99 | 0.994 | 0.914 | 0.103 | 0.952 | 0.232 |
| 3.5 | 18 | IG | $\sqrt{n}$ | $\infty$ | 90.93 | 0.993 | 0.914 | 0.109 | 0.952 | 0.232 |
| 3.0 | 12 | Gini | $\sqrt{n}$ | $\infty$ | 90.74 | 0.989 | 0.916 | 0.128 | 0.951 | 0.232 |
| 3.0 | 19 | IG | $n$ | $\infty$ | 90.68 | 0.989 | 0.915 | 0.122 | 0.951 | 0.232 |
| 3.0 | 16 | Gini | $\sqrt{n}$ | 5 | 91.11 | 0.997 | 0.913 | 0.083 | 0.953 | 0.231 |
| 3.0 | 14 | Gini | $n$ | 5 | 90.93 | 0.991 | 0.916 | 0.122 | 0.952 | 0.231 |
| 3.0 | 12 | IG | $\sqrt{n}$ | $\infty$ | 90.80 | 0.990 | 0.915 | 0.122 | 0.951 | 0.231 |
| 3.5 | 12 | IG | $\sqrt{n}$ | 8 | 91.05 | 0.996 | 0.913 | 0.090 | 0.953 | 0.230 |
| 3.0 | 15 | IG | $n$ | $\infty$ | 90.87 | 0.993 | 0.914 | 0.103 | 0.952 | 0.230 |
| 3.5 | 8 | IG | $\sqrt{n}$ | 5 | 91.05 | 0.995 | 0.914 | 0.096 | 0.953 | 0.230 |
| 3.5 | 16 | Gini | $\sqrt{n}$ | 5 | 91.05 | 0.995 | 0.914 | 0.096 | 0.953 | 0.230 |
| 3.0 | 17 | IG | $\log_2 n$ | $\infty$ | 90.99 | 0.995 | 0.913 | 0.096 | 0.952 | 0.230 |
| 3.0 | 9 | IG | 1 | $\infty$ | 91.05 | 0.996 | 0.913 | 0.090 | 0.953 | 0.230 |
| 3.5 | 16 | Gini | $\log_2 n$ | 8 | 91.05 | 0.996 | 0.913 | 0.090 | 0.953 | 0.229 |
| 3.0 | 14 | IG | 1 | $\infty$ | 91.11 | 0.998 | 0.912 | 0.077 | 0.953 | 0.229 |
| 3.0 | 17 | Gini | $\sqrt{n}$ | 8 | 90.93 | 0.993 | 0.914 | 0.103 | 0.952 | 0.229 |
| 3.0 | 11 | Gini | $\sqrt{n}$ | 8 | 90.99 | 0.994 | 0.914 | 0.103 | 0.952 | 0.229 |

*Continued on next page*

Table S5 – *Continued from previous page*

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 10 | Gini | $\log_2 n$ | $\infty$ | 90.87 | 0.991 | 0.915 | 0.115 | 0.952 | 0.229 |
| 3.5 | 16 | Gini | $\log_2 n$ | $\infty$ | 90.99 | 0.995 | 0.913 | 0.096 | 0.952 | 0.229 |
| 3.0 | 13 | IG | $\sqrt{n}$ | $\infty$ | 90.80 | 0.991 | 0.915 | 0.115 | 0.951 | 0.228 |
| 3.5 | 14 | IG | 1 | $\infty$ | 90.99 | 0.995 | 0.913 | 0.096 | 0.952 | 0.228 |
| 3.0 | 16 | Gini | $n$ | 5 | 90.68 | 0.988 | 0.916 | 0.128 | 0.951 | 0.228 |
| 3.0 | 14 | Gini | $n$ | $\infty$ | 90.02 | 0.976 | 0.919 | 0.173 | 0.947 | 0.228 |
| 3.0 | 8 | IG | $n$ | $\infty$ | 89.96 | 0.975 | 0.919 | 0.173 | 0.946 | 0.227 |
| 3.5 | 15 | IG | $\log_2 n$ | $\infty$ | 91.05 | 0.996 | 0.913 | 0.090 | 0.953 | 0.227 |
| 3.0 | 14 | IG | $n$ | 5 | 90.99 | 0.995 | 0.914 | 0.096 | 0.952 | 0.227 |
| 3.0 | 18 | IG | $\sqrt{n}$ | 5 | 91.05 | 0.997 | 0.913 | 0.083 | 0.953 | 0.227 |
| 3.0 | 7 | IG | $\log_2 n$ | $\infty$ | 90.68 | 0.988 | 0.916 | 0.128 | 0.951 | 0.227 |
| 3.5 | 14 | Gini | $\sqrt{n}$ | $\infty$ | 90.74 | 0.989 | 0.915 | 0.122 | 0.951 | 0.227 |
| 3.0 | 18 | IG | $n$ | 5 | 90.93 | 0.993 | 0.914 | 0.103 | 0.952 | 0.226 |
| 3.5 | 18 | IG | $\sqrt{n}$ | 8 | 90.93 | 0.993 | 0.914 | 0.103 | 0.952 | 0.226 |
| 3.0 | 8 | Gini | $n$ | $\infty$ | 89.35 | 0.965 | 0.921 | 0.205 | 0.943 | 0.226 |
| 3.0 | 15 | Gini | $n$ | $\infty$ | 90.32 | 0.981 | 0.918 | 0.154 | 0.948 | 0.226 |
| 3.0 | 6 | IG | $n$ | 5 | 90.68 | 0.988 | 0.916 | 0.128 | 0.951 | 0.226 |
| 3.0 | 17 | IG | 1 | $\infty$ | 91.05 | 0.997 | 0.913 | 0.083 | 0.953 | 0.226 |
| 3.5 | 19 | Gini | $\sqrt{n}$ | 8 | 90.93 | 0.993 | 0.914 | 0.103 | 0.952 | 0.225 |
| 3.0 | 11 | Gini | $\sqrt{n}$ | 5 | 90.93 | 0.993 | 0.914 | 0.103 | 0.952 | 0.225 |
| 3.0 | 11 | IG | $\log_2 n$ | $\infty$ | 90.93 | 0.993 | 0.914 | 0.103 | 0.952 | 0.225 |
| 3.0 | 17 | IG | $\sqrt{n}$ | $\infty$ | 90.93 | 0.993 | 0.914 | 0.103 | 0.952 | 0.225 |
| 3.5 | 16 | IG | $\sqrt{n}$ | 8 | 90.99 | 0.995 | 0.913 | 0.090 | 0.952 | 0.225 |
| 3.0 | 10 | Gini | 1 | $\infty$ | 90.80 | 0.991 | 0.914 | 0.109 | 0.951 | 0.225 |

*Continued on next page*

Table S5 – *Continued from previous page*

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 19 | Gini | $\sqrt{n}$ | $\infty$ | 90.87 | 0.992 | 0.914 | 0.109 | 0.952 | 0.225 |
| 3.5 | 18 | Gini | $n$ | 5 | 90.87 | 0.992 | 0.914 | 0.109 | 0.952 | 0.225 |
| 3.0 | 16 | Gini | $n$ | 8 | 90.44 | 0.984 | 0.917 | 0.141 | 0.949 | 0.224 |
| 3.5 | 12 | Gini | $\sqrt{n}$ | 5 | 90.99 | 0.995 | 0.914 | 0.096 | 0.952 | 0.224 |
| 3.0 | 18 | Gini | $\sqrt{n}$ | $\infty$ | 90.62 | 0.987 | 0.916 | 0.128 | 0.950 | 0.224 |
| 3.0 | 8 | Gini | $\sqrt{n}$ | $\infty$ | 90.14 | 0.979 | 0.918 | 0.160 | 0.947 | 0.223 |
| 3.0 | 12 | Gini | $\log_2 n$ | $\infty$ | 90.62 | 0.987 | 0.916 | 0.128 | 0.950 | 0.223 |
| 3.5 | 17 | IG | $n$ | $\infty$ | 90.74 | 0.991 | 0.914 | 0.109 | 0.951 | 0.223 |
| 3.0 | 8 | Gini | $n$ | 5 | 90.44 | 0.984 | 0.917 | 0.141 | 0.949 | 0.223 |
| 3.5 | 9 | Gini | $\sqrt{n}$ | $\infty$ | 90.62 | 0.987 | 0.916 | 0.128 | 0.950 | 0.222 |
| 3.0 | 16 | IG | $n$ | 8 | 90.80 | 0.991 | 0.915 | 0.115 | 0.951 | 0.222 |
| 3.5 | 10 | IG | $n$ | 5 | 90.93 | 0.994 | 0.913 | 0.096 | 0.952 | 0.222 |
| 3.0 | 15 | Gini | $\sqrt{n}$ | $\infty$ | 90.68 | 0.989 | 0.915 | 0.122 | 0.951 | 0.222 |
| 3.0 | 12 | IG | 1 | $\infty$ | 90.99 | 0.995 | 0.913 | 0.090 | 0.952 | 0.222 |
| 3.0 | 9 | IG | $\sqrt{n}$ | $\infty$ | 90.80 | 0.991 | 0.915 | 0.115 | 0.951 | 0.222 |
| 3.5 | 15 | Gini | $\sqrt{n}$ | $\infty$ | 90.80 | 0.991 | 0.914 | 0.109 | 0.951 | 0.222 |
| 3.5 | 14 | IG | $n$ | 5 | 90.87 | 0.993 | 0.914 | 0.103 | 0.952 | 0.221 |
| 3.5 | 19 | IG | $\sqrt{n}$ | $\infty$ | 90.80 | 0.991 | 0.914 | 0.109 | 0.951 | 0.221 |
| 3.0 | 15 | IG | $\sqrt{n}$ | 8 | 90.99 | 0.995 | 0.913 | 0.090 | 0.952 | 0.221 |
| 3.5 | 9 | Gini | $\sqrt{n}$ | 5 | 90.99 | 0.995 | 0.913 | 0.090 | 0.952 | 0.221 |
| 3.5 | 13 | IG | $\sqrt{n}$ | 8 | 90.99 | 0.995 | 0.913 | 0.090 | 0.952 | 0.220 |
| 3.0 | 17 | IG | $\sqrt{n}$ | 8 | 90.93 | 0.994 | 0.913 | 0.096 | 0.952 | 0.220 |
| 3.5 | 12 | Gini | 1 | $\infty$ | 90.93 | 0.994 | 0.913 | 0.096 | 0.952 | 0.220 |
| 3.5 | 10 | Gini | $\sqrt{n}$ | 5 | 90.93 | 0.994 | 0.913 | 0.096 | 0.952 | 0.220 |

*Continued on next page*

Table S5 – *Continued from previous page*

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 16 | IG | $n$ | $\infty$ | 90.68 | 0.989 | 0.915 | 0.122 | 0.951 | 0.220 |
| 3.5 | 16 | IG | $n$ | 8 | 90.74 | 0.990 | 0.915 | 0.115 | 0.951 | 0.220 |
| 3.5 | 9 | Gini | $\log_2 n$ | $\infty$ | 90.74 | 0.990 | 0.915 | 0.115 | 0.951 | 0.220 |
| 3.5 | 19 | IG | $n$ | 8 | 90.87 | 0.993 | 0.914 | 0.103 | 0.952 | 0.220 |
| 3.0 | 18 | Gini | $n$ | 5 | 90.74 | 0.990 | 0.915 | 0.115 | 0.951 | 0.219 |
| 3.0 | 8 | IG | $n$ | 5 | 90.62 | 0.988 | 0.915 | 0.122 | 0.950 | 0.219 |
| 3.0 | 9 | Gini | $n$ | 8 | 90.50 | 0.985 | 0.916 | 0.135 | 0.949 | 0.219 |
| 3.0 | 15 | IG | $\log_2 n$ | 8 | 90.99 | 0.996 | 0.912 | 0.083 | 0.952 | 0.219 |
| 3.5 | 17 | IG | $\log_2 n$ | 8 | 90.99 | 0.996 | 0.912 | 0.083 | 0.952 | 0.219 |
| 3.5 | 10 | IG | 1 | $\infty$ | 90.87 | 0.993 | 0.914 | 0.103 | 0.952 | 0.219 |
| 3.5 | 15 | IG | $\sqrt{n}$ | 8 | 90.87 | 0.993 | 0.914 | 0.103 | 0.952 | 0.219 |
| 3.5 | 19 | Gini | $n$ | 5 | 90.80 | 0.991 | 0.914 | 0.109 | 0.951 | 0.219 |
| 3.5 | 14 | Gini | $\log_2 n$ | $\infty$ | 90.87 | 0.993 | 0.914 | 0.103 | 0.952 | 0.219 |
| 3.0 | 10 | Gini | $\sqrt{n}$ | 5 | 90.93 | 0.994 | 0.913 | 0.096 | 0.952 | 0.218 |
| 3.0 | 6 | Gini | $\log_2 n$ | $\infty$ | 90.08 | 0.979 | 0.917 | 0.154 | 0.947 | 0.218 |
| 3.0 | 8 | IG | $\log_2 n$ | $\infty$ | 90.44 | 0.985 | 0.916 | 0.135 | 0.949 | 0.218 |
| 3.0 | 10 | Gini | $n$ | $\infty$ | 89.66 | 0.971 | 0.919 | 0.179 | 0.944 | 0.217 |
| 3.5 | 15 | Gini | $\log_2 n$ | 8 | 91.05 | 0.997 | 0.912 | 0.077 | 0.953 | 0.217 |
| 3.5 | 13 | IG | $n$ | 8 | 90.74 | 0.991 | 0.914 | 0.103 | 0.951 | 0.216 |
| 3.0 | 11 | IG | $\sqrt{n}$ | 8 | 90.87 | 0.993 | 0.913 | 0.096 | 0.952 | 0.216 |
| 3.0 | 8 | Gini | $\log_2 n$ | 8 | 90.87 | 0.993 | 0.913 | 0.096 | 0.952 | 0.216 |
| 3.5 | 19 | Gini | $\log_2 n$ | $\infty$ | 90.93 | 0.995 | 0.913 | 0.090 | 0.952 | 0.215 |
| 3.5 | 4 | IG | $n$ | 5 | 89.59 | 0.971 | 0.919 | 0.179 | 0.944 | 0.215 |
| 3.0 | 18 | IG | $\sqrt{n}$ | $\infty$ | 90.87 | 0.993 | 0.913 | 0.096 | 0.952 | 0.215 |

*Continued on next page*

Table S5 – *Continued from previous page*

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 7 | Gini | $\sqrt{n}$ | 5 | 90.87 | 0.993 | 0.913 | 0.096 | 0.952 | 0.215 |
| 3.5 | 15 | IG | $n$ | 5 | 90.87 | 0.994 | 0.913 | 0.090 | 0.952 | 0.214 |
| 3.0 | 17 | Gini | $n$ | 5 | 90.62 | 0.988 | 0.915 | 0.122 | 0.950 | 0.214 |
| 3.0 | 12 | Gini | 1 | $\infty$ | 90.80 | 0.993 | 0.913 | 0.096 | 0.951 | 0.214 |
| 3.5 | 9 | IG | $\sqrt{n}$ | 5 | 90.99 | 0.997 | 0.912 | 0.077 | 0.952 | 0.214 |
| 3.0 | 17 | Gini | 1 | $\infty$ | 90.99 | 0.997 | 0.912 | 0.077 | 0.952 | 0.214 |
| 3.5 | 12 | IG | $\log_2 n$ | $\infty$ | 90.80 | 0.992 | 0.914 | 0.103 | 0.951 | 0.214 |
| 3.0 | 18 | Gini | $n$ | $\infty$ | 90.02 | 0.978 | 0.917 | 0.154 | 0.947 | 0.214 |
| 3.5 | 17 | IG | $\sqrt{n}$ | $\infty$ | 90.87 | 0.993 | 0.913 | 0.096 | 0.952 | 0.214 |
| 3.5 | 11 | IG | $n$ | 8 | 90.56 | 0.988 | 0.915 | 0.115 | 0.950 | 0.213 |
| 3.5 | 12 | IG | $n$ | 8 | 90.68 | 0.989 | 0.915 | 0.115 | 0.951 | 0.213 |
| 3.5 | 18 | IG | $\sqrt{n}$ | 5 | 90.99 | 0.997 | 0.912 | 0.077 | 0.952 | 0.213 |
| 3.5 | 10 | IG | $\sqrt{n}$ | $\infty$ | 90.50 | 0.986 | 0.916 | 0.128 | 0.950 | 0.212 |
| 3.0 | 11 | IG | $n$ | 5 | 90.87 | 0.993 | 0.913 | 0.096 | 0.952 | 0.212 |
| 3.0 | 7 | Gini | $n$ | 5 | 90.62 | 0.989 | 0.915 | 0.115 | 0.950 | 0.212 |
| 3.0 | 12 | Gini | $\sqrt{n}$ | 8 | 90.68 | 0.990 | 0.914 | 0.109 | 0.951 | 0.212 |
| 3.0 | 14 | Gini | $n$ | 8 | 90.26 | 0.982 | 0.916 | 0.141 | 0.948 | 0.212 |
| 3.0 | 16 | IG | $\sqrt{n}$ | 5 | 91.05 | 0.997 | 0.912 | 0.077 | 0.953 | 0.211 |
| 3.0 | 7 | IG | 1 | $\infty$ | 90.87 | 0.994 | 0.913 | 0.090 | 0.952 | 0.211 |
| 3.0 | 12 | Gini | $\log_2 n$ | 8 | 90.80 | 0.993 | 0.913 | 0.090 | 0.951 | 0.211 |
| 3.0 | 7 | Gini | $\log_2 n$ | $\infty$ | 90.68 | 0.990 | 0.914 | 0.109 | 0.951 | 0.211 |
| 3.5 | 14 | IG | $n$ | 8 | 90.62 | 0.989 | 0.915 | 0.115 | 0.950 | 0.211 |
| 3.0 | 14 | Gini | 1 | $\infty$ | 90.93 | 0.995 | 0.912 | 0.083 | 0.952 | 0.210 |
| 3.0 | 8 | IG | $\log_2 n$ | 8 | 90.93 | 0.995 | 0.912 | 0.083 | 0.952 | 0.210 |

*Continued on next page*

Table S5 – *Continued from previous page*

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 17 | Gini | $\log_2 n$ | 8 | 90.93 | 0.995 | 0.912 | 0.083 | 0.952 | 0.210 |
| 3.5 | 8 | IG | $\log_2 n$ | $\infty$ | 90.56 | 0.987 | 0.915 | 0.122 | 0.950 | 0.210 |
| 3.0 | 11 | Gini | $n$ | 5 | 90.50 | 0.986 | 0.916 | 0.128 | 0.950 | 0.210 |
| 3.5 | 17 | Gini | $\sqrt{n}$ | $\infty$ | 90.74 | 0.991 | 0.914 | 0.103 | 0.951 | 0.210 |
| 3.5 | 19 | IG | 1 | $\infty$ | 90.99 | 0.997 | 0.912 | 0.077 | 0.952 | 0.210 |
| 3.5 | 11 | IG | $n$ | $\infty$ | 90.44 | 0.985 | 0.916 | 0.128 | 0.949 | 0.210 |
| 3.5 | 17 | Gini | $\sqrt{n}$ | 5 | 90.80 | 0.993 | 0.913 | 0.096 | 0.951 | 0.210 |
| 3.0 | 13 | Gini | $n$ | 8 | 90.38 | 0.985 | 0.916 | 0.128 | 0.949 | 0.209 |
| 3.0 | 16 | IG | $\log_2 n$ | $\infty$ | 90.80 | 0.993 | 0.913 | 0.096 | 0.951 | 0.209 |
| 3.5 | 18 | IG | $n$ | 8 | 90.62 | 0.989 | 0.914 | 0.109 | 0.950 | 0.208 |
| 3.5 | 9 | Gini | $\log_2 n$ | 8 | 90.87 | 0.995 | 0.912 | 0.083 | 0.952 | 0.208 |
| 3.0 | 8 | Gini | $\sqrt{n}$ | 8 | 90.50 | 0.987 | 0.915 | 0.122 | 0.950 | 0.208 |
| 3.0 | 3 | IG | $n$ | 5 | 90.14 | 0.981 | 0.916 | 0.141 | 0.947 | 0.208 |
| 3.0 | 5 | IG | $n$ | 8 | 90.08 | 0.979 | 0.917 | 0.147 | 0.947 | 0.208 |
| 3.0 | 11 | Gini | $n$ | $\infty$ | 90.08 | 0.979 | 0.917 | 0.147 | 0.947 | 0.208 |
| 3.5 | 12 | Gini | $\sqrt{n}$ | $\infty$ | 90.68 | 0.990 | 0.914 | 0.109 | 0.951 | 0.208 |
| 3.0 | 15 | Gini | $\sqrt{n}$ | 5 | 90.99 | 0.997 | 0.911 | 0.071 | 0.952 | 0.207 |
| 3.0 | 18 | IG | 1 | $\infty$ | 90.99 | 0.997 | 0.911 | 0.071 | 0.952 | 0.207 |
| 3.0 | 6 | Gini | $n$ | $\infty$ | 88.08 | 0.949 | 0.922 | 0.231 | 0.935 | 0.207 |
| 3.0 | 15 | Gini | $\sqrt{n}$ | 8 | 90.87 | 0.993 | 0.913 | 0.096 | 0.952 | 0.207 |
| 3.5 | 16 | IG | $\sqrt{n}$ | $\infty$ | 90.68 | 0.991 | 0.914 | 0.103 | 0.951 | 0.207 |
| 3.0 | 15 | IG | $\log_2 n$ | $\infty$ | 90.74 | 0.992 | 0.913 | 0.096 | 0.951 | 0.207 |
| 3.0 | 13 | Gini | $\sqrt{n}$ | $\infty$ | 90.80 | 0.993 | 0.913 | 0.090 | 0.951 | 0.206 |
| 3.0 | 19 | Gini | $\sqrt{n}$ | 5 | 90.93 | 0.996 | 0.912 | 0.077 | 0.952 | 0.206 |

*Continued on next page*

Table S5 – *Continued from previous page*

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.0 | 18 | Gini | $n$ | 8 | 90.44 | 0.986 | 0.915 | 0.122 | 0.949 | 0.206 |
| 3.0 | 17 | Gini | $n$ | $\infty$ | 90.14 | 0.981 | 0.916 | 0.141 | 0.947 | 0.206 |
| 3.0 | 5 | IG | $\log_2 n$ | 8 | 90.93 | 0.996 | 0.912 | 0.077 | 0.952 | 0.206 |
| 3.5 | 8 | IG | $\sqrt{n}$ | $\infty$ | 90.02 | 0.979 | 0.917 | 0.147 | 0.947 | 0.206 |
| 3.0 | 11 | Gini | $\sqrt{n}$ | $\infty$ | 90.62 | 0.989 | 0.914 | 0.109 | 0.950 | 0.205 |
| 3.0 | 9 | Gini | 1 | $\infty$ | 90.87 | 0.995 | 0.912 | 0.083 | 0.952 | 0.205 |
| 3.5 | 12 | IG | $n$ | 5 | 90.74 | 0.992 | 0.913 | 0.096 | 0.951 | 0.205 |
| 3.0 | 9 | IG | $\log_2 n$ | $\infty$ | 90.68 | 0.991 | 0.914 | 0.103 | 0.951 | 0.205 |
| 3.5 | 17 | Gini | $n$ | 8 | 90.44 | 0.986 | 0.915 | 0.122 | 0.949 | 0.204 |
| 3.5 | 19 | IG | $n$ | $\infty$ | 90.80 | 0.993 | 0.913 | 0.090 | 0.951 | 0.204 |
| 3.0 | 8 | Gini | $n$ | 8 | 89.53 | 0.971 | 0.918 | 0.167 | 0.944 | 0.204 |
| 3.5 | 18 | IG | $\log_2 n$ | 8 | 90.93 | 0.996 | 0.912 | 0.077 | 0.952 | 0.204 |
| 3.5 | 13 | IG | $\log_2 n$ | $\infty$ | 90.87 | 0.995 | 0.912 | 0.083 | 0.952 | 0.204 |
| 3.0 | 7 | IG | $\log_2 n$ | 8 | 90.93 | 0.996 | 0.912 | 0.077 | 0.952 | 0.203 |
| 3.5 | 10 | Gini | $\sqrt{n}$ | $\infty$ | 90.20 | 0.982 | 0.916 | 0.135 | 0.948 | 0.203 |
| 3.5 | 18 | Gini | $\sqrt{n}$ | 8 | 90.68 | 0.991 | 0.914 | 0.103 | 0.951 | 0.203 |
| 3.5 | 6 | Gini | $\sqrt{n}$ | $\infty$ | 89.29 | 0.967 | 0.919 | 0.179 | 0.942 | 0.203 |
| 3.0 | 14 | IG | $\sqrt{n}$ | $\infty$ | 90.56 | 0.989 | 0.914 | 0.109 | 0.950 | 0.203 |
| 3.0 | 11 | IG | $\sqrt{n}$ | 5 | 90.93 | 0.997 | 0.911 | 0.071 | 0.952 | 0.203 |
| 3.0 | 14 | IG | $\sqrt{n}$ | 5 | 90.99 | 0.998 | 0.911 | 0.064 | 0.953 | 0.202 |
| 3.0 | 13 | Gini | $n$ | $\infty$ | 89.90 | 0.977 | 0.917 | 0.147 | 0.946 | 0.202 |
| 3.5 | 4 | IG | $\sqrt{n}$ | 8 | 89.90 | 0.977 | 0.917 | 0.147 | 0.946 | 0.202 |
| 3.0 | 19 | Gini | $\sqrt{n}$ | 8 | 90.80 | 0.994 | 0.912 | 0.083 | 0.951 | 0.202 |
| 3.5 | 14 | IG | $n$ | $\infty$ | 90.20 | 0.982 | 0.916 | 0.135 | 0.948 | 0.202 |

*Continued on next page*

Table S5 – *Continued from previous page*

| $\psi$ | #Estimators | Criterion | #Max Features | Max Depth | Accuracy | Sensitivity | Precision | Specificity | $F_1$-score | MCC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 6 | IG | $n$ | 8 | 89.17 | 0.966 | 0.919 | 0.179 | 0.942 | 0.202 |
| 3.5 | 13 | Gini | $n$ | $\infty$ | 90.32 | 0.985 | 0.915 | 0.122 | 0.949 | 0.201 |
| 3.0 | 17 | IG | $\log_2 n$ | 8 | 90.93 | 0.997 | 0.911 | 0.071 | 0.952 | 0.201 |
| 3.0 | 16 | Gini | 1 | $\infty$ | 90.80 | 0.994 | 0.912 | 0.083 | 0.951 | 0.201 |
| 3.0 | 13 | Gini | 1 | $\infty$ | 90.93 | 0.997 | 0.911 | 0.071 | 0.952 | 0.201 |
| 3.0 | 19 | Gini | 1 | $\infty$ | 90.93 | 0.997 | 0.911 | 0.071 | 0.952 | 0.201 |
| 3.0 | 14 | IG | $n$ | 8 | 90.50 | 0.987 | 0.915 | 0.115 | 0.950 | 0.200 |
| 3.5 | 7 | IG | $n$ | $\infty$ | 89.90 | 0.977 | 0.917 | 0.147 | 0.946 | 0.200 |
| 3.5 | 9 | IG | $\log_2 n$ | 5 | 90.93 | 0.997 | 0.911 | 0.064 | 0.952 | 0.200 |