
Network optimization based on Genetic Algorithms for high-level classification via complex networks

Janayna Moura Fernandes



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Uberlândia
2023

Janayna Moura Fernandes

**Network optimization based on Genetic
Algorithms for high-level classification via
complex networks**

Dissertation submitted to the Faculty of Computing
at the Federal University of Uberlândia in accordance
with the requirements of the program for the degree of
Master in Computer Science.

Concentration area: Computer Science

Advisor: Prof. Dr. Murillo Guimarães Carneiro

Co-advisor: Prof. Dr. Gina Maira Barbosa de Oliveira

Uberlândia

2023

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

F363
2023

Fernandes, Janayna Moura, 1994-
Network Optimization based on Genetic Algorithms for
high-level classification via Complex Networks [recurso
eletrônico] / Janayna Moura Fernandes. - 2023.

Orientador: Murillo Guimarães Carneiro.
Coorientadora: Gina Maira Barbosa de Oliveira.
Dissertação (Mestrado) - Universidade Federal de
Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: <http://doi.org/10.14393/ufu.di.2023.146>
Inclui bibliografia.

1. Computação. I. Carneiro, Murillo Guimarães, 1988-,
(Orient.). II. Oliveira, Gina Maira Barbosa de, -,
(Coorient.). III. Universidade Federal de Uberlândia.
Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

To all those who have spent their lives seeking knowledge

Acknowledgements

I would like to express my sincere gratitude to my advisor, Professor Murillo Carneiro, for his guidance and support throughout the timeframe taken to complete this work. His valuable insights, feedback, support and encouragement helped me a lot to delve into the right theoretical approaches and bring out relevant insights through this master dissertation.

I would also like to express my appreciation to all of the lecturers and professors at the Computer Science Department of the Federal University of Uberlândia who helped me learn about this topic and inspired me to do more research on it.

Finally, I would like to express my gratitude to my family and friends for their constant support and encouragement throughout the entirety of my time spent pursuing my educational career.

“His dreams have taken hold of his research. His dreams have worn him out, exhausted him so that he sometimes cannot tell whether he is awake or asleep.”
(Alan Lightman, Einstein’s Dreams, 1992)



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Ciência da Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br



ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação de Mestrado 4/2023, PPGCO				
Data:	1 de março de 2023	Hora de início:	09:05	Hora de encerramento:	11:45
Matrícula do Discente:	12112CCP015				
Nome do Discente:	Janayna Moura Fernandes				
Título do Trabalho:	Network optimization based on Genetic Algorithms for high-level classification via complex networks				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Inteligência Artificial				
Projeto de Pesquisa de vinculação:	-				

Reuniu-se, por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Paulo Henrique Ribeiro Gabriel - FACOM/UFU, Fabrício Aparecido Breve - UNESP, Gina Maira Barbosa de Oliveira - FACOM/UFU (Coorientadora) e Murillo Guimarães Carneiro - FACOM/UFU, orientador da candidata.

Os examinadores participaram desde as seguintes localidades: Fabrício Aparecido Breve - Rio Claro/SP; Paulo Henrique Ribeiro Gabriel, Gina Maira Barbosa de Oliveira e Murillo Guimarães Carneiro - Uberlândia/MG. A discente participou da cidade de Uberlândia/MG.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Murillo Guimarães Carneiro, apresentou a Comissão Examinadora e a candidata, agradeceu a presença do público, e concedeu à Discente a palavra para a exposição do seu trabalho. A duração da apresentação da Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir a candidata. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando a candidata:

Aprovada

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.



Documento assinado eletronicamente por **Murillo Guimarães Carneiro, Professor(a) do Magistério Superior**, em 06/03/2023, às 10:42, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Fabrizio Aparecido Breve, Usuário Externo**, em 06/03/2023, às 10:53, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Paulo Henrique Ribeiro Gabriel, Professor(a) do Magistério Superior**, em 06/03/2023, às 11:35, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Gina Maira Barbosa de Oliveira, Professor(a) do Magistério Superior**, em 06/03/2023, às 16:24, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4291057** e o código CRC **DA0CD1A2**.

Abstract

Network-based classification has demonstrated its value especially due to its inherent capacity to capture the properties of networked data (e.g., structural and dynamical). However, its performance depends heavily on the network structure. In this sense, we present GANet, a technique of network structural optimization based on genetic algorithms (GAs) for the classification via characterization of importance. The importance based classification is a recent network classification technique that employs the pagerank measure to capture the underlying data relationship. In particular, we hypothesize that the prominent characteristics of GAs, such as their robust search mechanism and binary representation, may provide a more effective network architecture. Further, in an effort to capture the relationships between the networked data, we also analyze, despite pagerank, other network measures, namely degree, betweenness, closeness, and shortest path length. In summary, experimental findings using real data sets demonstrated that the proposed algorithm outperforms the widely used k -nearest neighbors graph method in terms of classification accuracy. They also show competitive results against a state-of-the-art network optimization technique based on swarm intelligence. Meanwhile, for the network measures, results revealed that pagerank and degree produced the best outcomes and statistically outperformed all other network measures in terms of predictive capability and robustness. Our technique was also applied to the detection of autism spectrum disorder from salivary data processed by the attenuated total reflectance Fourier-transform infrared (ATR-FTIR) spectroscopy. In the experiments, GANet outperformed both linear discriminant analysis, a widely adopted technique in ATR-FTIR analysis, and support vector machine, a state-of-the-art technique for such problems. Moreover, these results give evidence about the potential of our approach in dealing with such a difficult problem, characterized by high-dimensional data and arbitrary distributions.

Keywords: Complex Networks, Genetic Algorithms, Network Optimization, Data Classification, Graph construction, Graph Optimization, ASD, Autism, ATR-FTIR.

List of Figures

Figure 1 – A data set consisting of two classes that form distinct patterns.	18
Figure 2 – k NN technique idea	24
Figure 3 – Example of a directed graph with self-loop.	26
Figure 4 – The steps of a simple GA.	30
Figure 5 – The three principal individual representations: binary, integer and real.	30
Figure 6 – Roulette selection technique	31
Figure 7 – Crossover techniques: two points crossover and uniform crossover	31
Figure 8 – The five different network construction techniques.	35
Figure 9 – Examples of network measures used for the network-based classification through characterization of importance.	44
Figure 10 – Evaluation of the predictive accuracy of the network measures in rela- tion to the various metrics under study.	47
Figure 11 – Comparing γ variation across network measures on data sets	49
Figure 12 – Comparing k variation across network measures on data sets	51
Figure 13 – Structural optimization method via GA applied to network learning	54
Figure 14 – Illustrative example of <i>Map-all</i>	55
Figure 15 – Critical Nemenyi diagram comparing the average predictive perfor- mance ranking of analyzed models.	58
Figure 16 – Comparison between the networks obtained by k NNG and GANet-C	59
Figure 17 – Data pre-processing techniques plot of saliva samples where blue indi- cates control data and orange ASD data.	64
Figure 18 – GA-encoded salivary FTIR data for ASD detection.	65
Figure 19 – Initial generation of individuals example	65
Figure 20 – Individual encoding representing a network	66
Figure 21 – GANet overall concept	67

List of Tables

Table 1 – Main steps of the classification via Importance Characterization	41
Table 2 – Brief description of the real data sets in terms of the number of data items, number of attributes and number of classes	46
Table 3 – The predictive capability with standard deviation (SD) is determined by the network’s best γ and k configuration. The best results are boldfaced.	46
Table 4 – Predictive capability study of the network measures with positive significant differences as determined by the statistical Wilcoxon test	48
Table 5 – Robustness of the network measures in terms of averaged predictive capability along all parameter configurations. Best results are boldfaced.	50
Table 6 – Analysis of the robustness of network metrics with substantial positive (“#Wins”) differences according to the Wilcoxon test.	50
Table 7 – Brief description of the datasets being analyzed in terms of number of data items, attributes and number of classes	56
Table 8 – Different configurations under investigation composed of the parameters γ , q , selection, reinsertion and crossover methods.	57
Table 9 – Accuracy and standard deviation for GANet model configurations compared to k NNG technique	57
Table 10 – Topological measures of Mean Assortativity, Closeness, Shortest Mean Path, and Clustering Coefficient in different network configurations . . .	59
Table 11 – Comparison of the best configuration of GANet versus PSONet, a state-of-the-art method for structural optimization of networks.	60
Table 12 – Group-specific demographics characteristics	63
Table 13 – Results using <i>degree</i> as the network measure for importance concept . .	68
Table 14 – Classification results of low and high level techniques	68

List of acronyms

ACO Ant colony optimization

ANN Artificial neural networks

ASD Autism spectrum disorder

ATR-FTIR Attenuated total reflection-Fourier transform infrared

***b*Matching** *b*Matching graph

***b*Dash** *b*Dash graph

Cr Crossover rate

DT Decision tree

ϵ NG ϵ -radius neighborhood graph

GA Genetic algorithm

***k*NN** *k*-nearest neighbors algorithm

***k*NNG** *k*-nearest neighbors graph

ML Machine learning

M*k*NNG Mutual *k*-nearest neighbors graph

MST Minimum spanning tree graph

MSTs Multiple MST

NB Naive Bayes

Pmut Mutation rate

PSO Swarm optimization

SL-PSO Social learning particle swarm optimization

SVM Support vector machine

Sk NNG Symmetric k -nearest neighbor graph

T_p Population size

List of symbols

\mathcal{X}	Data set
\mathcal{G}	Graph
\mathcal{V}	Vertices set
\mathcal{E}	Edges set
\mathcal{W}	Weights matrix
\mathcal{P}	Particle
N	Number of samples in class
v	A vertex in a network
e	An edge in a network
c	Number of classes
l	A vertex label
β	Damping factor
k_i^{in}	In-degree of vertex i
k_i^{out}	Out-degree of vertex i
η	Number of vertex in the shortest path
α	A component
ξ	Local efficiency
\mathcal{F}	Efficiency function
\mathcal{I}	Importance classifier

Contents

1	INTRODUCTION	17
1.1	Hypotheses	19
1.2	Objectives	20
1.3	Contributions	21
1.4	Dissertation organization	21
2	BACKGROUND	23
2.1	Supervised machine learning	23
2.1.1	<i>k</i> -Nearest Neighbors (<i>k</i> NN)	24
2.1.2	Support vector machine (SVM)	24
2.1.3	Linear Discriminant Analysis (LDA)	24
2.2	Complex networks	25
2.2.1	Pagerank	26
2.2.2	Betweenness centrality	27
2.2.3	Closeness centrality	27
2.2.4	Degree centrality	27
2.2.5	Shortest path length	28
2.3	Bio-inspired optimization	28
2.3.1	Swarm intelligence	28
2.3.2	Genetic algorithms	29
3	RELATED WORK	33
3.1	Data classification in complex networks	33
3.2	Network construction	34
3.3	Network optimization	35
3.4	Network-based classification	37
3.4.1	Classification via pattern conformation	37
3.4.2	Classification via importance characterization	39

3.5	Characterization of network measures for network-based classification	41
4	EVALUATION OF CENTRALITY MEASURES FOR CHARACTERIZATION OF IMPORTANCE	43
4.1	Overview	43
4.2	Experimental design	45
4.3	Analyses and results	45
4.3.1	Predictive capability	45
4.3.2	Predictive robustness	47
4.4	Chapter summary	52
5	GANET: A GA FOR NETWORK STRUCTURAL OPTIMIZATION	53
5.1	Overview	53
5.2	Experimental design	54
5.3	Analyses and results	56
5.4	Chapter summary	60
6	GANET FOR DETECTION OF AUTISM SPECTRUM DISORDER FROM VIBRATIONAL SPECTROSCOPY SALIVARY SAMPLES	61
6.1	Overview	61
6.2	Attenuated total reflection-Fourier transform infrared spectroscopy	62
6.3	Data processing	62
6.4	GA structural optimization for ASD detection	64
6.5	Results and Discussion	67
6.6	Chapter summary	69
7	CONCLUSION	70
7.1	Main findings	70
7.2	Bibliographic contributions	72
	BIBLIOGRAPHY	73

Introduction

Data classification is one of the most well-known tasks in machine learning, having several associations to human cognitive behavior (CARBONELL; MICHALSKI; MITCHELL, 1983). In a natural sense, humans are always classifying something: the people in their social circle (coworkers, family, friends, college classmates, etc.), weather (sunny, cloudy, windy, rainy, stormy, etc.), and so on. In fact, they are frequently associated not only with the physical similarity between objects, but also with the semantic concept that one desires to represent. When we see a table, regardless of its substance, type, or format, we still recognize it as a table and label it as such (CARNEIRO; ZHAO, 2017).

Data classification in machine learning involves learning a model or function called a classifier from a set of known/labeled data to classify new data items into one of the classes. Examples of traditional classification techniques include Decision tree (DT), Artificial neural networks (ANN), Naive Bayes (NB) k -nearest neighbors algorithm (k NN) and Support vector machine (SVM) (CARNEIRO; ZHAO, 2018).

Traditional techniques (a.k.a. low-level techniques) successfully perform the classification task for several problems. However, there are scenarios or problems in which the use of networks brings advantages, for example when there is a lot of overlap in classes or very arbitrary data distribution (SILVA; ZHAO, 2012; CARNEIRO; ZHAO, 2017). Such techniques perform classification considering only the physical characteristics (distance or distribution), which means that they may have difficulties in capturing semantic properties of the data, such as pattern formation. While deep learning techniques, such as convolutional neural networks, can capture high-level patterns, by manipulating simultaneously millions of parameters, they are extremely dependant of such parameters as well as their multiple layers of neurons, which requires high computational time as well as non-trivial and non-intuitive decisions, such as several hyper-parameters choices and architectures decisions. Indeed, how exactly deep learning architectures map an input to an output is still a puzzle. Another issue is that many of these low-level techniques assume that all objects are of equal relevance, that's because the object's individual importance which may be relevant to the problem understanding is being ignored. (CARNEIRO;

ZHAO, 2017).

In contrast, it is natural to perform structural and topological research in a network, also the data network representation is semantically natural and intuitive when considering its organization solely in terms of vertices and edges. Representing data as network is natural and it provides us with different concepts and heuristics for the classification task, due to its ability to characterize not only the physical attributes, but also the data semantic structure (CARNEIRO; ZHAO, 2017). In this way, through the functional, spatial and topological characteristics obtained from the network based representation, we can understand, for example, the implicit pattern within the data (CARNEIRO et al., 2019). The Fig.1(a) shows a simple example of a classification task on toy data, in which there are very well defined patterns. Its objective is to classify the test object, \square (green). Low-level techniques struggle in this scenario as they essentially consider physical features of the data (similarity, distance or distribution), neglecting the structure and relationship between the data (CARNEIRO; ZHAO, 2017). Thus, such algorithms have serious limitations to associate the test object with the \triangle class (blue).

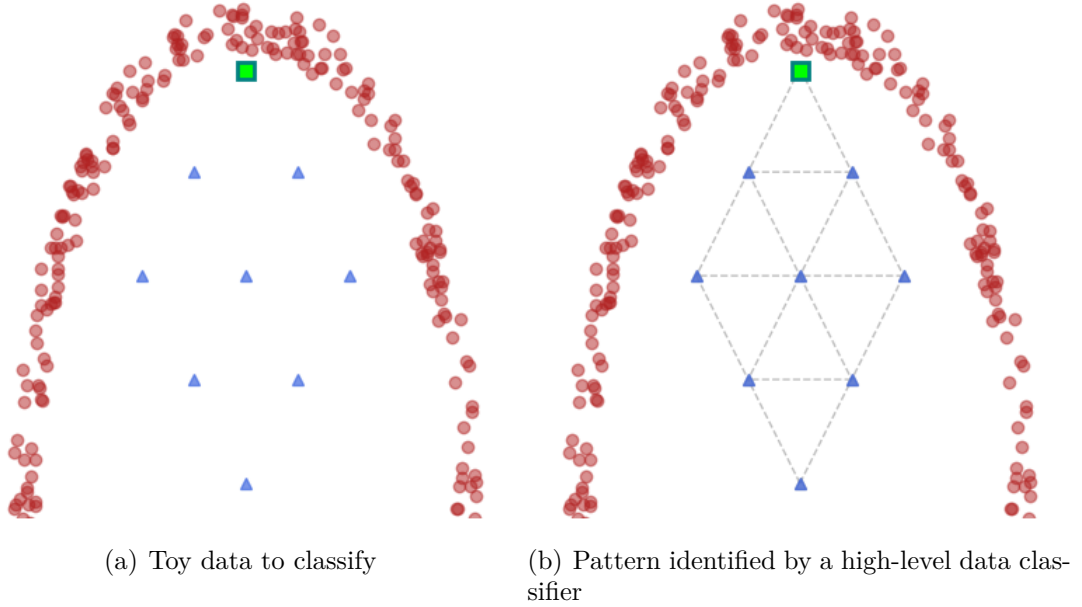


Figure 1 – Green test object (\square) to be classified between red (\circ) or blue (\triangle) class. (a) Traditional classifiers have difficulties due to the proximity the test object has to the circle class. (b) Example of high-level classification: the test object belongs to the diamond pattern.

On the other hand, the use of networks allows, besides examining physical attributes of the data, also to consider topological information within the data configuration in a network. This type of classification is called *high-level classification* (SILVA; ZHAO, 2012) and such techniques are able to detect semantic relationships such as the pattern formation shown in Fig. 1(a) coming out of the structural analysis of the data in the network, as presented in Fig. 1(b).

Even though high-level classification explores structural and topological characteristics, it is still dependent on a graph that is sufficiently representative. As the class patterns are directly extracted from the network, the graph needs to be able to reflect these data relationships well. The graph construction is done in accordance with a function that exploits some information of the networked data, e.g., spatial, structural or dynamical. In this line, there are some graph construction methods to deal with vector-based data in the supervised learning literature, such as k -associated graphs (Bertini Jr et al., 2011), k -nearest neighbors graph (CARNEIRO; GAMA; RIBEIRO, 2021), weighted matrix graph (CUPERTINO et al., 2018) and particle swarm optimized network (CARNEIRO et al., 2019).

While most techniques are essentially based in general purposes, like k -nearest neighbors or ϵ -neighborhood graphs, the objective of this study is to develop a solution based on genetic algorithms (GA) for structural optimization of networks, that is, to find the most adequate configuration to represent the connections between the vertices in the network.

The primary motivation behind this research is to develop an efficient mechanism for representing discrete optimization problems by using genetic algorithms (GAs). Such a mechanism would enable the representation of complex network connections within the problem, providing more accurate solutions.

Another of the key advantages of using GAs is their ability to handle problems with a large number of variables and constraints. In this work, the relationship between the variables is complex and nonlinear, which further highlights the suitability of GAs as an optimization tool.

By leveraging the unique strengths of GAs, this research aims to provide an effective solution to discrete optimization problems in the network context. By naturally reflecting the connections between variables, the proposed mechanism has the potential to improve the accuracy and efficiency of optimization solutions, paving the way for a wide range of applications in fields such as engineering, economics, and computer science.

1.1 Hypotheses

Recent research reveals several methods and resources in which we can associate structural and physical properties derived from complex networks measures for the classification task, e.g. (SILVA; ZHAO, 2012; CARNEIRO; ZHAO, 2017). Even with the noticeable aspects of the characterization of importance technique, only a limited evaluation of one centrality measure was conducted in that paper, using particularly the pagerank measure to score the data importance. In this sense, one of our investigated hypothesis states that, besides pagerank measure, other network measures are also able to perform the characterization of importance. Therefore, the investigation contemplated here aims to extend the previous one by considering and evaluating other network measures possible

in the literature.

H₁: Degree, betweenness, closeness and shortest path length are network measures able to perform the characterization of importance at the same level of pagerank.

Bio-inspired optimization techniques are able, jointly with network measures, to obtain more accurate results in the high-level classification process via structural optimization of graphs, such as PSONet (CARNEIRO et al., 2019). In this sense, one of the investigated hypothesis states that Genetic algorithm (GA) can provide more appropriate networks for classification via importance characterization than other heuristically generated, such as the k NNG, at the same time, they can be competitive with other state-of-the-art methods by allowing direct manipulation of network configurations based on the representation and operation of binary variables rather than continuous ones. It is mainly motivated by the efficient mechanism of GAs for representing discrete problems, which can naturally reflect network connections.

H₂: GA for graph construction provides optimized networks competitive with traditional and state-of-the-art methods of graph construction.

We also aim to analyze and compare the predictive performance of the proposed structural optimization GANet for autism detection. In the proposed technique, the salivary samples processed by the attenuated total reflectance Fourier-transform infrared (ATR-FTIR) technique are pre-processed and mapped to a network, so that each spectra sample represents a vertex and the edges between them are defined based on some similarity criterion. Different low-level methods, pre-processing strategies and network measures are evaluated in the classification of autism spectra data.

H₃: The proposed method, GANet, has the potential to improve the detection of autism using spectra obtained by ATR-FTIR from saliva samples.

1.2 Objectives

The main objective is to employ genetic algorithms in association with heuristics based on complex networks in the context of data classification, in particular high-level classification via importance characterization. More specifically:

- ❑ To investigate the contribution of different centrality measures to the high-level classification based on importance characterization.
- ❑ To develop a GA-based bio-inspired optimization technique to perform network structural optimization;
- ❑ To evaluate the proposed technique in relation to other state-of-the-art methods using as data saliva samples processed by FTIR.

1.3 Contributions

This dissertation presents empirical findings that support the use of network measures in conjunction with GA for high-level data classification across two aspects: Network measure to characterize the vertex's role in the network and GA to design an optimized network. Our contributions are summarized as follows:

- Evaluate a variety of network measures, including pagerank, degree, betweenness, closeness and shortest path length that outlines the semantic embeddings within the data.
- Present a learning framework that allows for high-level classification strategies using different GA configurations and evaluate the tradeoffs between each implementation.
- Compare the suggested technique, GANet, with other state-of-the-art classification techniques in the domain of FTIR-processed saliva samples as data.

1.4 Dissertation organization

This dissertation presents seven chapters starting with the present introduction chapter, followed by a background section and five chapters that describe the research studies, finalized with a conclusion section. The remainder of this chapter introduces each chapter briefly.

- **Chapter 2** presents core concepts and techniques for a proper understanding of this work, which encompasses machine learning, traditional supervised algorithms, complex networks, data classification through complex networks, network measures and bio-inspired optimization. It also outlines the significance of this dissertation related to these fields.
- **Chapter 3** reviews prior work that are particularly closely related to research done for this dissertation which includes network construction benchmark techniques as well as network structural optimization techniques, high-level data classification, and network measures evaluation.
- **Chapter 4** presents a new heuristic for high level-classification as our contribution to the machine learning literature in the context of network measures evaluation. Complex network measures, like pagerank, degree, betweenness, closeness and shortest path length were selected. Experiments revealed that pagerank and degree typically generated the greatest results and statistically surpassed every other predictive measure.

- **Chapter 5** outlines the hypothesis about the potential use of genetic algorithms for graph optimization in the context of high-level classification. It also provides a set of experiments that support the claim that optimized networks can increase accuracy results. We compared our algorithm with existing approaches. The results suggest that our method outperforms common approaches, especially in terms of representation.
- **Chapter 6** presents the classification method used to discover autism spectrum patterns in saliva samples processed by FTIR. In addition to that, pre-processing methods are discussed and evaluated.
- Finally, **Chapter 7** closes this dissertation. It summarizes the contributions of the dissertation and discusses future directions.

Background

In this chapter, the terms and several concepts associated with high-level classification and structural optimization of networks are presented and described. Summarily, it starts with supervised machine learning and then introduces complex networks and network-based classification. Finally, bio-inspired optimization is also introduced, focusing on swarm intelligence approaches and genetic algorithms.

2.1 Supervised machine learning

Machine learning (ML) is an umbrella term for the application of statistical methods and algorithms to automatically convert experience (historical data) into knowledge (MITCHELL, 1997). Some examples are supervised learning (SAMUEL, 1959), unsupervised learning (BARLOW, 1989) and reinforcement learning (SUTTON; BARTO, 2018). ML application areas are vast: financial market (FERNANDES et al., 2019), industry (LEI et al., 2020; ELSISI et al., 2021; DOGAN; BIRANT, 2021), medicine (ALYASSERI et al., 2022; SAMMUT et al., 2022), telecommunications (PUSTOKHINA et al., 2021) and science (physics, astronomy, biology, computer science) (LIU; ARUNACHALAM; TEMME, 2021; HUANG et al., 2021; LAWSON et al., 2021; GREENER et al., 2022).

Supervised learning usually uses statistics theories and mathematical models since the main purpose is to infer knowledge from a set of data samples (SARKER, 2021). Thus, the input to a supervised learning algorithm is the data set, which represents the experience, and its output is the model learned or deduced.

Classification is one of the most important data mining research topics. The algorithm utilizes the training data to create a classification model that will operate correctly on new data. The ability to generalize is a critical feature: a collection of historic data is used to perform well on future data. Classification's main objective is to predict the labels of new data instances using previously labeled data instances as training examples (MITCHELL, 1997). In recent decades, numerous classification systems have been developed for real world applications. A selection of them are detailed below:

2.1.1 k -Nearest Neighbors (k NN)

The k NN method (COVER; HART, 1967) is frequently utilized in data mining and machine learning applications because of its simple implementation. The basic idea of k NN is to predict the label of a test data instance using the majority rule, i.e., by associating the label of the test instance with the majority class of its k most similar training data instances in the feature space. k NN algorithm has at least two open challenges to define, namely the proximity measurement between two data points and the determination of the k value.

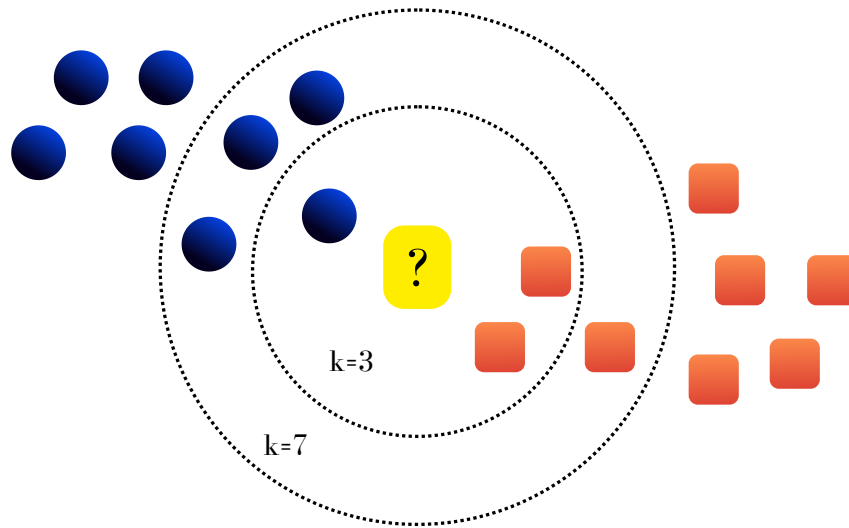


Figure 2 – k NN technique: The test instance is assigned to the class, which has the nearest neighbors. In the preceding example, when $k=3$, the test instance is classified as orange/ \square , whereas when $k=7$, the test instance is classified as blue/ \circ .

2.1.2 Support vector machine (SVM)

A SVM classifier performs classification by differentiating two or more classes by identifying the “boundary line” that separates them. The SVM model is produced by projecting the input vectors onto a new higher-dimensional feature space. Then, a kernel function $K(x_i, x_j)$ is used to construct an optimal separating hyperplane in the new feature space (VAPNIK, 1999). The ideal hyperplane maximizes the margin while fully separating the classes. Prior to constructing the SVM classifier, the kernel function must be selected, and different kernel functions can result in varying prediction results.

2.1.3 Linear Discriminant Analysis (LDA)

LDA technique (RAO et al., 1973) attempts to learn a linear transformation matrix for mapping while preserving the original data’s structure. LDA searches for vectors in the underlying space that discriminate classes most effectively. Given a set of independent

features, LDA defines two metrics: 1) one is referred to as the within-class scatter matrix, as given by

$$S_w = \sum_{j=1}^c \sum_{i=1}^{N_j} (x_i^j - \mu_j)(x_i^j - \mu_j)^T, \quad (1)$$

where x_i^j is the i th sample of class j , μ_j is the mean of class j , c is the number of classes, and N_j is the number of samples in class j ; and 2) the other is referred to as the between-class scatter matrix:

$$S_b = \sum_{j=1}^c (\mu_j - \mu)(\mu_j - \mu)^T, \quad (2)$$

where μ is the mean of all classes. The aim is to build a linear combination that maximizes the mean differences between the target classes while minimizing the within-class measure.

2.2 Complex networks

The Internet, social networks, the human brain, the stock market, blackouts and earthquakes share an important aspect in common: they are complex systems whose representation and modeling can be performed through complex networks (LI et al., 2021). Complex networks bring together a set of tools to represent and model such systems characterized by the existence of non-trivial connection patterns: neither completely regular nor chaotic, thus being called complex (CARNEIRO, 2017).

Similarities can be identified in different domains through representation in complex networks. A well-known example is the scale-free networks (BARABÁSI; ALBERT, 1999) that model preferred links present in various systems, such as the Internet and citation networks, for example (MAKAROV et al., 2021; WANG et al., 2022).

In the context of machine learning, the modeling and analysis of networked data have contributed to a number of tasks, such as community detection, multi-label classification and transductive learning (SPORNS, 2022; ENGELEN; HOOS, 2020). In this sense, network measures are able to characterize different aspects, behaviors and characteristics underlying the data, since they are able to examine not only their physical properties (distance or distribution, for example), but also topological information from their configuration in a network (RESENDE; CARNEIRO, 2021; CARNEIRO; GAMA; RIBEIRO, 2021).

The basic outline for capturing the global features of such systems is to describe them as graphs whose vertices represent the dynamical entities and whose links indicate their interactions. Graphs are very generic representation models naturally capable of characterizing various phenomena and applications in the real world.

A graph can be undirected or directed (digraph) depending on whether or not there is a directional dependence between the vertices, in other words, a source vertex and a

destination vertex (BIGGS; LLOYD; WILSON, 1986). A graph can also be weighted, with a weight assigned to each edge to specify the link strength between the vertices. The basic terminology of a graph \mathcal{G} is described by the set of *vertices* \mathcal{V} that represent instances, *edges* \mathcal{E} that represent relationships between the vertices that \mathcal{G} contains (BIGGS; LLOYD; WILSON, 1986) and \mathcal{W} representing the weight matrix associated with the edge set \mathcal{E} :

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W}) \quad (3)$$

A practical example of an undirected graph is a social network of friends; and for a directed one, a social network of followers. A graph can also contain self-loop or not. A self-loop is an edge from a vertex to itself. The Fig. 3 illustrates a graph with four vertices, $\mathcal{V} = \{1, 2, 3, 4\}$ and seven edges, $\mathcal{E} = \{(1, 1), (2, 3), (3, 1), (3, 2), (4, 1), (4, 2), (4, 3)\}$.

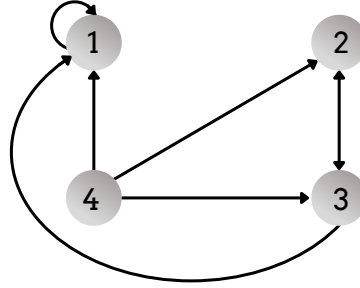


Figure 3 – Example of a directed graph with self-loop.

Network representation allows interpretation of associativity and measurement of the relations among data that are not easy to express with conventional representations, such as vectors. Overlapping dense and sparse regions in networks of data items, the emergence of new densities in sparse regions as a result of recursive interactions, and the possibility of approaching densities from different directions are all factors that can influence the predominance of certain characteristics and, ultimately, the classification of these regions through network measures. A selection of these measures are detailed below:

2.2.1 Pagerank

Centrality indices are fundamental network analysis metrics because they quantify the importance of a network vertex (LANGVILLE; MEYER, 2011; LANGVILLE; MEYER, 2004). Pagerank is a metric that describes the importance of a certain vertex in a network. It is determined by the number of connections received. In general, the greater the number of in-edges a vertex has, the more important it is. Also, a vertex is regarded as important if it has in-edges from other important vertices. This is one of the ranking factors now employed by Google's search engine, and its definition is as follows:

$$\mathcal{PG}_j = \beta \sum_{i \rightarrow j} \frac{\mathcal{PG}_i}{k_i^{out}} + \frac{1 - \beta}{N}, \quad (4)$$

where the \mathcal{PG}_j is the pagerank of vertex v_j , $i \rightarrow j$ denotes a link from vertex i to j , k_i^{out} is the out-degree of vertex i and β is the damping factor.

2.2.2 Betweenness centrality

Since this measure is based on shortest path computations, it provides information regarding the global significance of the network's topology (BRANDES, 2001). Betweenness quantifies the vertex's contribution to the network communication along the shortest paths. It is the total of all pairwise shortest paths that pass through a vertex. Specifically, the Betweenness value of a vertex $v \in V$, given a graph \mathcal{G} , is defined as:

$$\mathcal{BC}_i = \sum_{u,v \in V-i} \frac{\eta_{uv}^i}{\eta_{uv}}, \quad (5)$$

where η is the number of occurrences of vertex i along the shortest paths between vertices u and v .

2.2.3 Closeness centrality

The closeness metric estimates the average distance between a vertex and all other network vertices (OKAMOTO; CHEN; LI, 2008). In a basic view, two vertices are close to one another if they are adjacent to one another. In network theory, closeness is an advanced measure of centrality in a topological space that indicates a vertex's core position inside the network. The definition of closeness is the sum of the geodesic distances to all other vertices reachable from a given vertex, where $d(i, j)$ is the shortest path distance between vertices n_i and n_j . If a vertex is located at the network's edge, its total distance to all reachable vertices will be high, resulting in a low closeness value, and vice versa (ZHANG; LUO, 2017).

$$\mathcal{CC}_i = \left(\frac{n-1}{\sum_{j=1}^n d(i, j)} \right). \quad (6)$$

2.2.4 Degree centrality

Reflects the local attributes of the structural information based on the number of connections at each vertex (SNIJDERS, 1981). As the number of connections increases, so does the degree value. If the network is directed, a vertex's degree, denoted by k_i , is the number of edges leading into that vertex; otherwise, it is the total number of adjacent vertices.

$$\mathcal{DC}_i = k_i^{in} + k_i^{out}. \quad (7)$$

2.2.5 Shortest path length

Shortest path is the geodesic length between two pairs of vertices having the shortest path between them, $d(i, j)$ (BRANDES, 2008). For example, the distances can be determined using the Dijkstra or Bellman-Ford algorithm. The average shortest path of a vertex i is given by:

$$\mathcal{SP}_i = \frac{1}{N-1} \sum_{i \neq j} d(i, j). \quad (8)$$

2.3 Bio-inspired optimization

Optimization problems can be extremely challenging to solve because of the harsh time needs, nonlinear constraints, high set of parameters and also because these problems are often highly nonlinear (TANG; LIU; PAN, 2021). Traditional optimization approaches, such as Lagrangian methods and gradient-based methods, are incapable of dealing well with such types of problems (SUN et al., 2019). The use of bio-inspired optimization techniques is a promising alternative (DHIMAN, 2021).

Nature-inspired computing is an interdisciplinary area that brings together the fields of computer science, biology and statistics and it has two related goals: to explore natural phenomena as a source of inspiration for optimizing the search for solutions and the use of computers to model these phenomena (XIAO, 2016).

The search for solutions in the search space is explored coming out of a random sampling of its points using mechanisms based on *swarm intelligence* or mechanisms inspired by *biological evolution*, like Darwin's theory, and genetic operations such as crossover and mutation. Their main application is in problems where there is a very large space of solutions but there is not much information about this space (YANG, 2020).

Genetic algorithms (GA) (HOLLAND, 1992) and Swarm optimization (PSO) (KENNEDY; EBERHART, 1995) are well known for their success in dealing with high-dimensional cost functions and together provide a solid view on bio-inspired optimization's state of the art (DARWISH, 2018).

2.3.1 Swarm intelligence

Swarm Intelligence is a branch of bio-inspired computing that consists of a simple, decentralized and self-organizing method based on the collective behavior of animals such as colony of fireflies, bees swarms, flocks of birds and shoals (BANSAL, 2019). However, approaches inspired by the immune system (NI et al., 2019) and crowds of people (SALIH et al., 2018) can also be found in the literature. PSO is one of the techniques that are part of Swarm Intelligence, another technique example is the Ant colony optimization (ACO) (DORIGO; BIRATTARI; STUTZLE, 2006).

The intelligent behavior (global) of the swarm often emerges from the distributed interactions (local) between the particles, as a result of applying the principles of cultural adaptation: evaluation, comparison and imitation (POLI; KENNEDY; BLACKWELL, 2007). Evaluation is the particle's ability to quantify how good it is in relation to some parameter or objective. Comparison is the ability to evaluate themselves in relation to other particles. And imitation (not only imitating but also understanding their purposes) is the ability to change behavior when they perceive/observe that another particle is more successful. In other words, the particles learn from their own experiences and from other particles, evaluate each other, compare and imitate other (better) particles in the swarm.

Each particle \mathcal{P}_i is associated with a velocity vector and a position vector. In addition, it has a memory of its best positions called $pbest_i$. The best position of all particles is called $gbest$. The probability of the particle making a decision depends on how successful it was ($pbest_i$) and also on the social influences of the results of its neighbors ($gbest$).

This collective behavior has the advantage of allowing single particles to learn from each other (BANSAL, 2019), enabling faster learning as it is not necessary for each particle to learn alone by trial and error.

Social learning particle swarm optimization (SL-PSO) (CHENG; JIN, 2015) is a variant algorithm of PSO, whose objective is to optimize problems with large numbers of local optima or high dimensional problems, in which the traditional PSO performs poorly. With SL-PSO, each particle only learns from particles better than itself in the current swarm, as opposed to classical PSO, where particles learn from better individual and global solutions over iterations. The SL-PSO basically consists of calculating the fitness function, ordering the particles and successively updating them. It takes advantage of mechanisms such as imitation and enhancement that are capable of accelerating learning rates, especially when the objective (behavior) to be learned is complex (high dimensionality). Among the PSO principles, imitation is the most fundamental concept for SL-PSO, according to the authors (CHENG; JIN, 2015). Imitation can lead to behavior matching, at the population level, such as culture or tradition.

2.3.2 Genetic algorithms

GAs (HOLLAND, 1992) are a generic search and optimization framework in which the solution search space is explored from a random sampling of its points using mechanisms that are based on the abstraction of the concept of evolution and genetic operations such as the generation of a new being and the possibility of mutation (KATOCH; CHAUHAN; KUMAR, 2021).

Such a framework considers a certain period of generations where a random population of candidate solutions is submitted to genetic operations (selection, crossover and mutation). Each individual in the population represents a hypothesis from the solution search space, which is evaluated through a performance measure that indicates how good that

hypothesis is for solving the problem. The performance measure is also known as *fitness function* or *fitness* (MIRJALILI, 2019). A conventional GA can be described succinctly through the flowchart in Fig. 4.

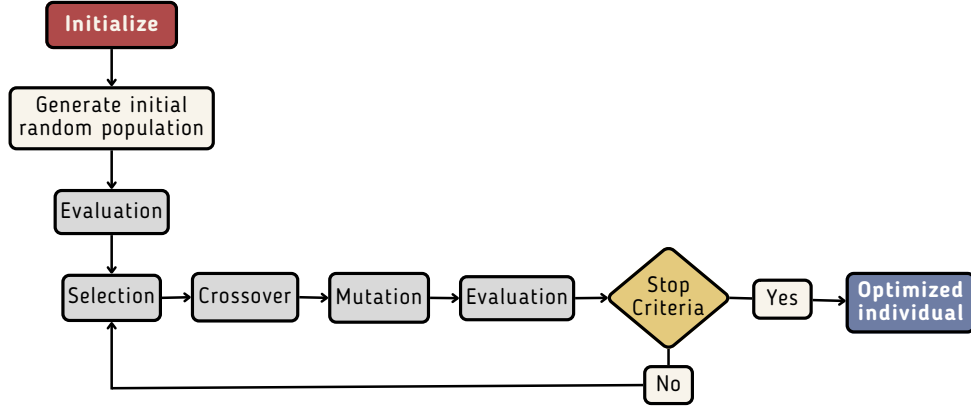


Figure 4 – The steps of a simple GA.

In the Fig. 4, first, the population of candidate solutions is randomly initialized, later evaluated using a fitness function. Then the optimization process starts for a certain number of generations or other stopping criteria. This process is characterized by the following steps: *selection*, in which individuals with the best evaluation will have preference to participate in the crossover stage; *crossover*, in which the selected individuals will contribute to the formation of new individuals; *mutation*, which alters, with some probability, the new individuals generated, allowing a global exploration of the search space. Then the individuals obtained are evaluated and the *reinsertion* occurs, with the objective of selecting the individuals that will be part of the next generation. At the end of this process, the individual with the best evaluation is returned as the solution.

One of the great challenges in the design of a GA refers to the representation and evaluation of individuals, as well as the consequent choice of methods adopted in the genetic operators of selection, crossover, mutation and reinsertion (KATOCH; CHAUHAN; KUMAR, 2021). The individual representations are mainly through the binary, integer, and real numbers. In the Fig. 5 examples of them can be seen.

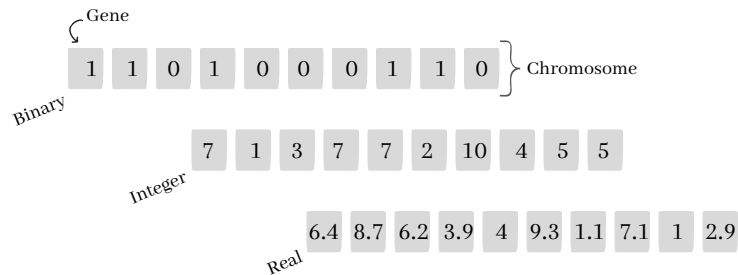


Figure 5 – The three principal individual representations: binary, integer and real.

The **selection** of the best individuals for recombination is fundamentally stochastic and can be done through methods such as Roulette, Ranking, Tournament, Truncation

and among others (YADAV; SOHAL, 2017). In the *roulette* selection method, illustrated in Fig.6, each individual has the probability of being selected according to its associated fitness.

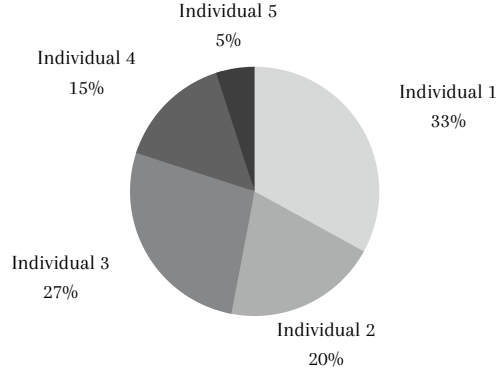


Figure 6 – Roulette selection technique: selection strategy for distinguishing the merits of the data based on their fitness. Because *Individual 1* has the highest fitness, it receives the largest piece of the roulette wheel.

In **crossover** the selected individuals are used for genetic recombination and thus will generate the children that will compose the total population ($Tp + Cr$). Each child is made up of their parents' material and the crossover technique determines which parent's genes will be used. Cyclic Crossover, PMX, Two Point Crossover and Uniform Crossover are examples of crossover techniques (KORA; YADLAPALLI, 2017). The below Fig.7 illustrates both two points and uniform crossover.

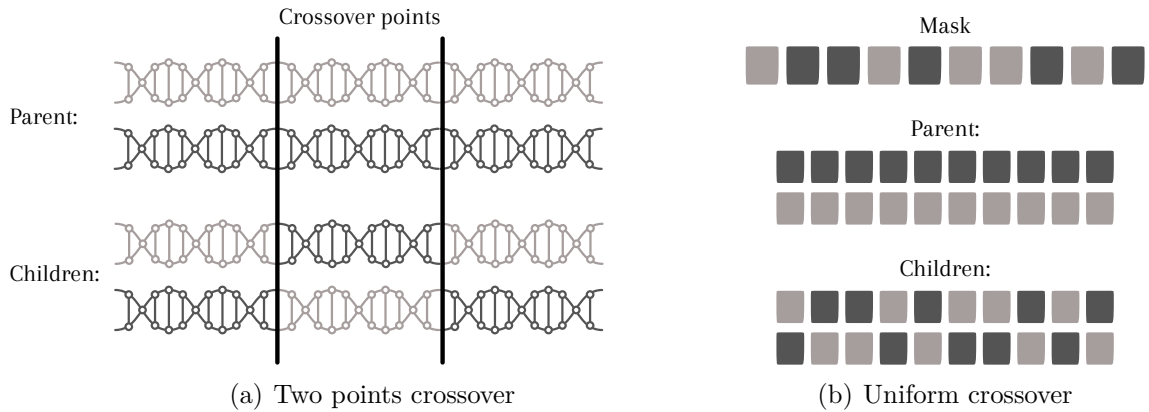


Figure 7 – Crossover techniques: (a) two points crossover: two crossover points are randomly selected, in which, the individuals only swap the bits between the two crossover points; and (b) uniform crossover: a mask indicates that the bit is copied from the first parent otherwise from the second parent.

Biologically, “**mutation**” means an abrupt change in the characteristics of a gene on a chromosome. In this algorithm phase, the children generated can undergo genetic variations according to a predefined probability. These genetic variations can occur, depending on the representation of the individual, by the alteration of a randomly chosen

gene or by the random replacement of one gene by another. The mutation aims to add diversity and increase exploration of the search space (MIRJALILI, 2019).

In **reinsertion**, the selection takes place among the total individuals ($Tp + Cr$) which will compose the next generation (Tp) based on the fitness of the individuals. Examples are Ordered Reinsertion and Pure Reinsertion (MIRJALILI, 2019).

Related works

In this chapter, research problems and related works are discussed, such as network construction, structural optimization of networks and analysis of centrality measures that characterize the objects in a network for data classification.

3.1 Data classification in complex networks

In supervised learning, the input data of a classification algorithm is a set of instances each denoted by the tuple (x, y) , where x denotes the attributes and y the target class. Thus, the number of classes is known and the algorithm receives information from the classes during training (supervised learning). The goal is to map a set of attributes x to its class label y , in order to correctly classify new instances whose attributes (x) are known, but class (y) is not.

Unlike the more well-known classification techniques that are based only on the physical features of the data (e.g. distance or distribution), network-based learning is also able to consider topological patterns of the data when representing them in a graph (CARNEIRO; ZHAO, 2017; CARNEIRO et al., 2019). To represent the data set \mathcal{X} from a vector of attributes, \mathcal{X} is transformed into a graph $\mathcal{G} = \mathcal{V}, \mathcal{E}$, where each vertex $v_i \in \mathcal{V}$ represents a data item $i \in \mathcal{X}$ and each edge $e_{i,u} \in \mathcal{E}$ represents a connection between the vertices $v_i, v_u \in \mathcal{X}$. Therefore, \mathcal{G} plays a fundamental role in obtaining the results since the class patterns are directly extracted from it.

The k NNG technique generates a directed graph in which a vertex is connected to the closest k vertices, as long as the objects are of the same class. The ε -radius neighborhood technique generates an undirected graph in which a threshold is defined to determine the connections between objects. Both techniques are able to generate graph from the input data in the form of attribute vector however make strong assumptions about the data, such as that their relations can be mapped from the same number of links (k or ε) between the vertices (CARNEIRO et al., 2019). Carneiro et al. (2019) proposed a more recent alternative to deal with such limitations, the structural optimization of networks,

which is a framework for structural optimization of networks based on particle swarms. According to the study's discussions, finding the best network configuration to learn a given problem is essential to improve the results obtained during high-level classification.

3.2 Network construction

To perform high-level classification on non-graph data (e.g., attribute vectors, images, text, etc.), the first step is to generate a network (or graph) in which the vertices and the edges represent, respectively, the objects and the relationships between them. The network's construction is a crucial step, since it is from the network that the data underlying information will be extracted for the classification process (CARNEIRO et al., 2019). The most used graph construction techniques in the literature are k -nearest neighbors (k NN) graph and the radius neighborhood (ε -radius) graph.

In general, graph construction involves selecting a similarity function or kernel to estimate the affinity between the instances. The most used network formation methods are k NNG and ε -radius neighborhood graph (ε NG) (CARNEIRO, 2017). An example of the five different graph construction techniques described here is shown in Fig. 8.

□ **Neighborhood graphs.** A (dis)similarity function is used to measure the affinity between the vertices. The k nearest neighbors edge set of v_i is denoted as $k\text{NNG}(v_i)$. The most popular choice for any graph-based machine learning problem is possibly the k NNG method due to its simplicity and effectiveness (CAI; ZHENG; CHANG, 2018). But because of the data distribution or the sampling density it might not be appropriate to use the same neighborhood size for all vertices. Also, it requires the k parameter definition. Some different neighborhood graphs are defined as follows:

1. **k NNG:** In a k NN graph a vertex has edges to its first k -nearest neighbors: the vertex v_i connects to v_j if $v_j \in k\text{NNG}(v_i)$, thus the k NNG method produces a directed graph.
2. **Sk NNG:** The Symmetric k -nearest neighbor graph (Sk NNG) is a variation of k NNG that produces an undirected graph: vertex v_i e vertex v_j are connected if $v_i \in k\text{NNG}(v_j)$ or $v_j \in k\text{NNG}(v_i)$.
3. **Mk NNG:** The Mutual k -nearest neighbors graph (Mk NNG) is also a variation of k NNG in which two vertices v_i and v_j are connected if $v_i \in k\text{NNG}(v_j)$ and $v_j \in k\text{NNG}(v_i)$.

□ **ε -radius neighborhood graph.** In a ε -radius graph, all vertices within a certain distance ε are connected; otherwise, they do not connect (CARNEIRO; ZHAO, 2018), where ε is a predefined distance radius. This method can produce densely

connected components by generating an excessive number of edges. It is both sensitive to the similarity function and dataset.

1. **ϵ NG:** In a ϵ N graph a vertex has edges to the vertices under a ϵ -radius distance: the vertex v_i connects to v_j if $d(v_i, v_j) \leq \epsilon$, the ϵ NG method produces a undirected graph.

□ **Minimum spanning tree graphs.** A Minimum spanning tree graph (MST) produces an acyclic connected graph by searching a minimum sum of the pairwise scores based on the dissimilarity matrix. Despite its being effectively employed to model and analyze complex brain networks (STAM et al., 2014), the traditional MST is too sparse, containing only $n - 1$ edges, where n is the vertex set size (GABOW et al., 1986).

1. **MST:** In a MST graph all vertices are connected minimizing their total distance D :

$$\sum_{i,j} D_{i,j} \quad (9)$$

2. **MSTs:** To address the problem of producing very sparse graphs, (ZEMEL; CARREIRA-PERPIÑÁN, 2004) proposed a Multiple MST (MSTs) method that involves an ensemble of MST from randomly perturbed data. In MSTs the data is perturbed, multiple MST are generated from it and then the final graph is generated from the MST ensemble by creating an edge if $e_{i,j} > 0$ in any MST.

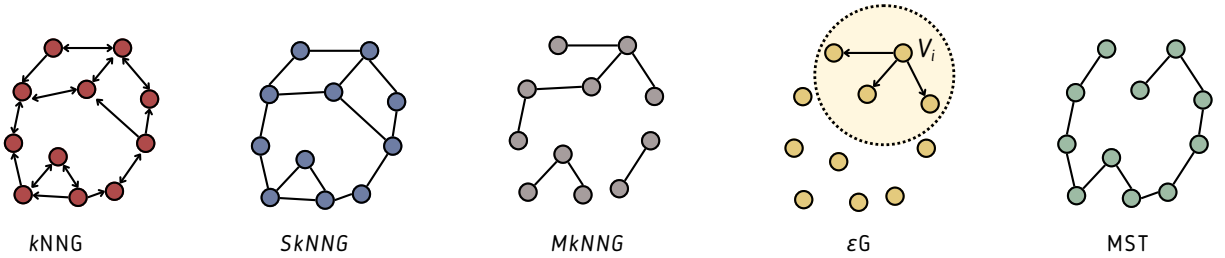


Figure 8 – The five different network construction techniques.

There are also other works of metric combination in the literature, for instance, the authors (CUPERTINO et al., 2018) developed a framework where a weighted and undirected network without self-loops is constructed calculating a weighted matrix using a distance function (e.g., Euclidean distance).

3.3 Network optimization

Carneiro et al. (2019) investigated structural optimization in the context of complex networks for data classification. They propose a particle swarm-based optimization frame-

work that is responsible for building an optimized network from data in the form of an attribute vector. The SL-PSO (CHENG; JIN, 2015) algorithm is employed to build the network from the training data optimizing the quality function under a validation dataset. Initially, SL-PSO creates a swarm of particles, where each particle represents a network. Then, at each iteration, the particles are evaluated and updated according to the quality function, the classification via characterization of importance. In the end, the SL-PSO returns the particle with the best evaluation value. This study concluded that the classification in networks built by the optimization process presents better performance than those networks generated heuristically.

In (CARNEIRO et al., 2016), a bio-inspired framework for network structural optimization is studied on the high level conformation data classification model (please refer to 3.4.1) which classifies data instances by verifying the conformation of data pattern formed by the training data, which means a test instance receives the label from the data network whose structure is kept unmodified or is barely modified after the insertion of the test instance. During the optimization process, the framework is also anticipated to build up the network and adjust the parameters of the hybrid model. The quality function and complex network measures, such as closeness and assortativity, are also used to evaluate structural optimization. Experiments on two real-world data sets revealed that the network optimization approach not only improves the classification through SL-PSO but also outperforms k NNG.

Carneiro et al. (2017) studied an optimization framework for graph-based supervised dimensionality reduction in which the graph connections are updated repeatedly using the SL-PSO optimization approach. Experiments were conducted out on a variety of real-world image classification problems. Unlike most previous work, where a specific value defines the number of connections for all vertices, their approach can construct preserving and penalty graphs while optimizing the performance of the graph embedding approach. Experiments on a variety of real-world image classification problems revealed that the method outperformed the Sk NNG and Mk NNG generation methods. Furthermore, the optimization framework produced significant dimension reduction rates while simultaneously obtaining good predictive performance.

Strumberger et al. (2019) proposed a firefly bio-inspired swarm optimization approach for building a satisfying CNNs network architecture by optimizing the network hyper-parameters such as the number of convolutional and dense layers, the number of kernels per layer, and the kernel size. Finding the convolutional neural network architecture that achieves the highest performance for a specific application is one of the most difficult jobs in this task, because the values of a network's hyper-parameters have a significant influence on overall network performance. The empirical results obtained utilizing the MNIST dataset for hand-written digits against other metaheuristics methods demonstrated that the proposed framework is a robust and a high-performing technique.

Kotary, Nanda e Gupta (2021) proposes MaOWOA, a swarm-based many-objective optimization method based on whale optimization, to handle robust distributed clustering in wireless sensor networks (WSN) while optimizing privacy, security, and technical constraints. The suggested approach's simulation result was evaluated on six DTLZ functions from the DTLZ suit test problem against the MaOPSO, NSGA-III, NSGA III, VaEA and MOEA/D many-objective methods. The proposed method outperforms many-objective methods in terms of simulation time and provides competitive results. Furthermore, the MaOWOA is used in WSNs to provide robust distributed clustering (DMaOWOA). A weighted method is incorporated in this approach to discover and eliminate outliers, and the diffusion method of cooperation is used for distributed clustering. Four datasets are used to test the proposed DMaOWOA. In terms of the Silhouette index, DMaOWOA-based clustering outperformed PSO-based many objective distributed clustering and distributed K-Means clustering algorithms. The distributed approach eliminates the requirement of a central processor or base station. However, as network and data size grow, so does computing complexity.

3.4 Network-based classification

In high-level classification, the purpose is to investigate the properties and measurements of complex networks in order to characterize the emergence of network patterns during the classification of each test item. The techniques developed have several advantages over low-level techniques, including the absence of parameters and the capacity to discover classes even when they overlap and to capture spatial, topological, and functional links within the data. Next, two network-based classification strategies are presented. The first method uses complex networks measures in order to identify data pattern formation. And the last one uses the concept of importance to classify the test object.

3.4.1 Classification via pattern conformation

The classification technique proposed in (SILVA; ZHAO, 2012) is a hybrid classification technique that combines low-level and high-level learning concepts. Briefly, low-level classification can be implemented by any traditional classification technique (SVM, DT, ANN, NB), exploiting the physical characteristics of the data. On the other hand, high-level classification exploits the complex topological properties underlying the network constructed from the input data. The great advantage of networks, as data representation tools, is the ability to describe the topological structure of the data.

The proposed method can be divided into two steps: training and classification. In training, the training set is mapped into a graph using a combination of graph formation techniques: ϵ NG and k NNG. The ϵ NG technique creates a connection between two vertices if they are at a distance ϵ , while the k NNG technique establishes a connection

between the vertices i and j , if i is one of the k vertices closest to j or vice versa. The purpose of combining the ray- ϵ and k NNG techniques is to treat both dense (ϵ NG) and sparse (k NNG) regions. With this mechanism, each class is expected to generate a single component. If the region is sparse enough, that is, if the distance ϵ from a vertex i has less than k neighbors to connect then the k NNG approach is employed.

Still in the training phase, network measurements are calculated to try to characterize the structure and topological properties of these components. The authors used a combination of network measures, namely assortativity, clustering coefficient and mean degree. Whereas the average degree measure characterizes the network in local terms with strict information on the number of connections from each vertex in the network; that the clustering coefficient of each vertex captures local structures by defining how close the component is to a click of a complete graph by the current vertex and any of its neighbors; and that the assortativity coefficient measures the tendency of a vertex to connect to another at a global level; it can be seen that the three measures characterize the topological properties of the network both locally and globally. Thus, it is expected that the combination of these measures will be able to systematically capture the formation of the pattern present in the network. It is noteworthy that other network measures could also be adopted in the high-level classifier.

The idea for verifying the pattern compliance of a test instance is to analyze whether its insertion causes a large variation in the network measures that characterize the component. In other words, if there is a small change in the network measurements, the test instance conforms to all other vertices that make up that component, that is, it reproduces the same pattern as the original members of that class, or even enhances that pattern. On the other hand, if its insertion is responsible for a disturbance of the component's network measurements, then the test instance probably does not belong to that component.

In the classification phase, the object is received to predict, inserted into the network (based on the radius ϵ or k NNG). Once the data item is inserted, the impact of inserting this data item in its respective component is analyzed separately. If little or no change occurs, the high-level classifier produces a high binding value for that test instance in this component. On the other hand, if inserting the test object into the component drastically changes the component's pattern, the high-level classifier will produce a small association value in this component.

At the same time as the prediction made by the high-level classifier, a low-level classifier also predicts the probability of the test instance for each class of problem. Then, the predictions produced by the high and low-level classifiers are combined through a linear fit to derive the final prediction. A low compliance term makes the final hybrid classifier decision more based on spatial assumptions. When a high value is used, the dominant feature that the hybrid classifier tries to emphasize is the patterns that the classes exhibit.

It should be emphasized that when the classes are totally distinct and each class

is appropriately positioned, no assistance from the high-level classifier is required, and intuitively, the high-level classifier is also not needed. As the complexity of the class configuration increases, the compliance term must be increased to obtain the correct classification. This suggests that the high-level classifier is especially useful in complex classification situations and that improvements in level classification can be achieved by combining the two levels of learning.

3.4.2 Classification via importance characterization

Classification via importance characterization is a high-level classification technique that individually assesses the importance of data items to determine a label of a new instance proposed in (CARNEIRO; ZHAO, 2017). Furthermore, it takes advantage of both spatial and structural properties when representing data in graph form.

In this technique, the concept of *importance* is derived from a centrality measure called *pagerank*, which characterizes the importance that a given object has in a network based on the number of in-edges, this implies that higher the number of in-edges, the more important it is (PAGE et al., 1999).

The classification via importance characterization consists of two main phases: the training phase, where the graph is constructed from the input data (for example in the form of a vector of attributes), efficiency of information flow measure is calculated and then the importance based on pagerank is calculated; and the prediction phase, which consists of the virtual insertion of a given test object in the component based on improving the efficiency of the information flow and assigning the importance value to the test object, so that it will be classified to the component class in which it received the highest importance value (CARNEIRO; ZHAO, 2017).

First, the network is constructed using a certain approach whereas ensuring the connected vertices belong to the same class. Given two vertices v_i and v_j , a connection from v_i to v_j , denoted as e_{ij} , is defined by:

$$e_{ij} = \begin{cases} 1, & \text{if } x_j \in k\text{-NN}(x_i) \text{ and } l_i = l_j, \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

where $k\text{-NN}(x_i)$ represents the k -nearest neighbors of v_i and l_i the class label associated with v_i .

After the network construction, the efficiency is calculated, it measures how fast the information flows between the vertices of a component and it is given by:

$$\mathcal{E} = \frac{1}{N^\alpha} \sum_{y \in \alpha} \xi_i^\alpha, \quad (11)$$

where N^α denotes the number of vertices in the component α and ξ_i^α , the local efficiency, which is defined as:

$$\xi_i^\alpha = \frac{1}{N_i} \sum_{i \rightarrow j} D_{i,j}, \quad (12)$$

where N_i represents the total number of connections from i and $D_{i,j}$ is the distance between vertices v_i and v_j .

Then, the importance concept is computed. The *importance* of a test item y , denoted by \mathcal{I} , in relation to the class $l \in \mathcal{L}$ is given by:

$$\mathcal{I}_y^{(l)} = \sum_{j \in \Lambda_y^{(l)}} \mathcal{I}_j, \quad (13)$$

where $x_j \in X_{train}$ denotes a labeled vertex, $\Lambda_{x_k}^{(l)}$ is a set of vertices belonging to the class l in which x_k is temporarily connected, and \mathcal{I}_j represents the *importance* of vertex v_j .

In the prediction phase, each test instance $x_k \in X_{test}$ is temporarily inserted into the network components by assessing physical and topological properties embedded in the input data, such as the following:

$$\Lambda_y^{(l)} \cup j | \mathcal{F}(y, j) \geq 0, \quad (14)$$

where $\mathcal{F}(y, j)$ is a function that determines if a link between y and j increases or decreases the efficiency of the component. Vertex j is included in the $\Lambda_{x_k}^{(l)}$ set if it enhances efficiency. $\mathcal{F}(y, j)$ is given by:

$$\mathcal{F}(x_k, j) = \mathcal{E}_{j \in \alpha}^\alpha \cdot \gamma - D_{x_k, j}, \quad (15)$$

where $D_{x_k, j}$ is the distance between vertices x_k and j , \mathcal{E}^α is the efficiency of the component α , and γ is a parameter that corresponds to the network building approach's component efficiency values. In this regard, a parameter γ controls the trade-off between physical and topological properties, indicating how much physical and structural information should be considered (CARNEIRO; ZHAO, 2017). A high γ value indicates that the structural qualities represented by the network measure are more relevant. A small value indicates the physical attributes are more important. When $\gamma = 1.0$, the strategy employs the efficiency patterns as discovered through network generation methods.

The importance that the vertex v_k receives from its neighbors is provided via a score produced during the training phase and derived from the pagerank metric. Finally, x_k receives the label of the class where it was considered more important.

A brief of the technique's six steps is presented in Table 1. Given N input elements, the first step has a computational order of $\mathcal{O}(N^2)$ and the others have an order of $\mathcal{O}(N)$, except for the last one, which is constant. Taking the higher order term, the framework results in computational cost of order $\mathcal{O}(N^2)$ but using a Lanczos graph this complexity can be reduced between $\mathcal{O}(N^{1.06})$ and $\mathcal{O}(N^{1.33})$ (CARNEIRO et al., 2019).

Table 1 – Main steps of the classification via Importance Characterization

Step	Description
Network Construction	A graph is constructed (using some technique such as k NNG or radius ε)
Efficiency Calculation	Heuristic proposed by the authors, <i>differential efficiency of structural space</i> , which measures how fast is the flow of information between the vertices of a component, the closer the vertices the higher the efficiency
Importance calculation	The <i>pagerank</i> associated with each object in the graph
Insert the query	Insert the test object temporarily based on the efficiency measure to link the new object, check if inserting the query into that component increases or maintains the efficiency then inserts it temporarily, on the other hand if efficiency decreases then it is not an ideal candidate to be inserted temporarily
Importance calculation	The test object receives the importance of each of the vertices that connected to it
Query sort	Object is sorted on component label where it is connected to most important vertices

3.5 Characterization of network measures for network-based classification

Carneiro, Gama e Ribeiro (2021) characterized and comparatively evaluated the predictive performance of eight measures of complex networks in the context of data classification in terms of predictive capacity and robustness, namely assortativity, average local clustering coefficient, average degree, the betweenness, the mean of the shortest path, the closeness, the global clustering coefficient and the eigenvector centrality. The network measurements were evaluated using the pattern conformation high-level classification technique (SILVA; ZHAO, 2012) under different network formation techniques based on k -nearest neighbors and radius-neighborhood heuristics. According to the authors, certain complex network measures have better predictive performance than others.

The technique consists of building the network from data in the form of an attribute vector and then calculating the network measure for each component. Then, each test item is virtually inserted into the network, the network measure is recalculated, and the label of the component in which the insertion caused the least variation of the measure is assigned to the test item.

The results obtained showed that the shortest mean path and assortativity measures reached high predictive capacity, even in datasets with higher noise levels, and also ro-

bustness to the change of the graph structure, that is, less influenced by the variation of the graph structure. Furthermore, this research is the first contribution towards supporting the selection of complex network measures most appropriate for network-based data classification.

Evaluation of centrality measures for data classification via characterization of importance

The importance characterization technique, in the original work, takes into consideration the individual importance of each input data by using the pagerank measure, whereas other relevant centrality measures were not even examined. In this chapter, the prediction performance of other relevant measures are evaluated. From the literature, five complex network measures were chosen: pagerank, degree, betweenness, closeness and shortest path length.

4.1 Overview

The network-based classification technique investigated here consists of two major steps. In the training step, a graph construction method is selected to build up a network from the feature vector data and in the test step, a new instance is inserted into the network and it is classified according to the concept of importance.

PageRank is one successful and well-publicized metric that has applications in search, browsing, user navigation, and traffic estimation (PAGE et al., 1999). In spite of the fact that pagerank has been the only measure embraced within the related works, in this work we hypothesized that other network measures can have the capacity to perform the characterization of importance. Fig. 9 presents a basic outline of a two-class dataset in which three network measures are embraced to implement the characterization of importance (please, refer to 3.4.2): degree, pagerank and closeness. By the figure, one can see that after the network is learned, each class is defined by a component: the left component speaks for the blue class, and the right component for the red one. There is also an uncolored item, which speaks for the test case. The figure displays that by considering distinctive measures, we may accomplish different results, for example, degree and close-

ness, which classifies the test case as blue class, and pagerank, which classifies it as red class.

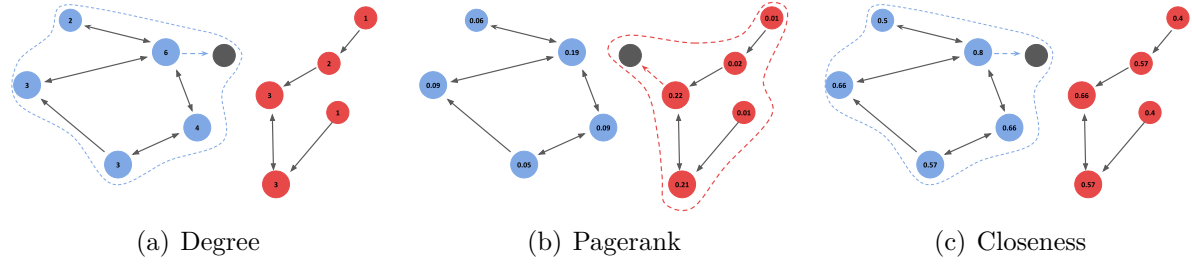


Figure 9 – Examples of network measures used for the network-based classification through characterization of importance.

After the network is constructed, Carneiro e Zhao (2017) proposed two extra measurement steps: efficiency and importance characterization. Efficiency measures how quick the information flows between the vertices of a component by analyzing physical and topological underlying aspects within the input data. In this sense, a parameter, γ , governs the trade-off between physical and topological features, it determines how much physical and structural information should be taken into account (CARNEIRO; ZHAO, 2017). A high γ value indicates that the structural characteristics obtained by the network measure are more relevant. A small value indicates that the physical characteristics are more relevant.

The primary goal of this chapter is the development and evaluation of new heuristics based on complex network measures for data classification via importance characterization. Recent research has revealed numerous methods for combining structural and physical features obtained from complex network measures for characterization tasks such as the one suggested in (SILVA; ZHAO, 2012; CARNEIRO; ZHAO, 2017). Despite the noteworthy elements of the characterization of importance technique, the authors did a limited evaluation in that study, using only the PageRank measure to score the data importance. As a result, the current chapter attempts to expand on the previous one by studying and evaluating different network measures accessible in the literature. The following are the primary contributions of this chapter research:

- Analyze the most well-known centrality measures of complex networks in the literature, comparing their performance and use in the context of data classification via importance characterization.
- Empirically evaluate new approaches for determining vertex importance from measures other than PageRank in terms of predicting capability and robustness.

4.2 Experimental design

Importance characterization evaluates the importance of each vertex in function of its pagerank ranking. This work follows the general stages of (CARNEIRO; ZHAO, 2017), besides the importance characterization, in which, in addition to pagerank, four alternative centrality metrics with various biases are investigated: degree, betweenness, closeness and average shortest path length.

The network is built using the k -nearest neighbors graph (k NNG) approach from vector features in the input training data. Such a method builds a graph in which each vertex is connected to its k -nearest vertices based on some affinity criterion as long as the vertices have the same class.

During the prediction phase, each test instance $x_k \in X_{test}$ is temporarily inserted into the network components based on the efficiency characterization. The importance that the vertex v_k receives from its neighbors is determined by a score produced during the training phase, which is taken from one of the network measures described previously in Section 2.2. Ultimately, x_k acquires the label of the class in which it was considered more important.

In the simulations, we utilized a 10-fold stratified cross-validation procedure that was averaged over three executions, for a total of 30 runs. For the k -nearest neighbors graph, we chose the range $\{1, 2, \dots, 15\}$ for the k parameter. For the trade-off between structural and physical attributes in our high-level classifier, the parameter γ was set to the interval $\{0, 0.5, 1, 2, 5\}$. For all simulations, we also employ the Euclidean distance as the affinity criterion.

4.3 Analyses and results

This section presents the simulations done to test the important characterization given five relevant network metrics across eight real-world datasets shown in Table 2, which can be found in (ASUNCION; NEWMAN, 2007). The datasets were selected in order to accommodate a wide variety of data domains, instances, properties, and classes.

4.3.1 Predictive capability

Here, we analyze the predictive capabilities of the network measures' ability to identify patterns in networked data. Toward our goal of understanding the effects of the network measures as importance concept we framed a guiding question to formalize the analyses: "how effective a specific network measure may be in classifying via importance characterization?"

Table 3 provides a summary of the accuracy attained by each network measure across the eight datasets studied. In the table, while boldfaced results show the best predicted

Table 2 – Brief description of the real data sets in terms of the number of data items ($\#Inst.$), number of attributes ($\#Attr.$) and number of classes ($\#Classes$).

Name (Abbrev.)	$\#Inst.$	$\#Attr.$	$\#Classes$
Appendicitis (App.)	106	7	2
Balance (Bal.)	625	4	3
Digits (Dig.)	5620	64	10
Ecoli (Eco.)	336	7	8
Glass (Gla.)	205	23	5
Iris	150	4	3
Sonar (Son.)	208	60	2
Thyroid (Thy.)	215	5	3

accuracy attained for a specific dataset, the value inside $()$ stands for the optimal γ value. The \mathcal{PG} column represents our state-of-the-art results (CARNEIRO; ZHAO, 2017). The table demonstrates that such a measure brought in favorable outcomes, providing the best results in four data sets. In five of the eight data sets, degree centrality (\mathcal{DC}) produced the best results.

Another notable aspect in Table 3 is that neither $\gamma=0$ nor $\gamma=5$ produced the best results, highlighting the importance of balancing the structural and topological qualities of the data with the physical ones. In actuality, $\gamma \in \{0.5, 1, 2\}$ produced the best results.

Table 3 – The predictive capability with standard deviation (SD) is determined by the network’s best γ and k configuration. The best results are boldfaced.

Dataset	\mathcal{PG}	\mathcal{BC}	\mathcal{CC}	\mathcal{DC}	\mathcal{SP}
App.	83.8±7.9 (1)	83.0±8.4 (5)	79.9±10.8 (0)	83.4±6.6 (1)	79.9±10.8 (0)
Bal.	94.9±2.5 (0.5)	93.6±2.7 (0)	95.0±2.5 (0.5)	95.5±2.4 (0.5)	94.8±2.6 (0)
Dig.	98.7±0.4 (0.5)	98.5±0.4 (0.5)	98.7±0.4 (0.5)	98.7±0.4 (0.5)	98.7±0.4 (0.5)
Eco.	87.1±5.4 (1)	85.4±5.6 (1)	81.5±5.7 (0.5)	87.5±5.2 (1)	82.6±5.5 (0.5)
Gla.	73.3±8.7 (0)	71.8±9.3 (0)	74.3±7.8 (1)	72.7±8.6 (0)	74.6±8.2 (0.5)
Iris	98.4±3.3 (2)	98.4±3.3 (1)	98.4±3.3 (1)	98.4±2.8 (1)	96.2±5.4 (0)
Son.	84.0±7.1 (0.5)	81.9±7.9 (0)	84.0±7.2 (0.5)	83.6±7.4 (0)	83.6±7.4 (0)
Thy.	96.8±4.0 (2)	94.9±4.3 (1)	96.8±3.6 (1)	97.2±3.5 (2)	96.5±3.7 (2)
Average	89.62	88.43	88.57	89.62	88.36

Fig. 10 provides a comparison of the five analyzed network measures. In the figure, the prediction accuracy is illustrated as a function of the graph construction parameter k , γ being the value that yielded the best results in Table 3. According to the figure, betweenness achieved outcomes comparable to pagerank and degree for the Appendicitis and Ecoli data sets, although producing the worst results for the Glass and Sonar datasets.

According to Table 3, pagerank and degree achieved the best outcomes for all data sets except for the Glass dataset, where shortest path length and closeness fared better. It should also be observed that, regardless of the network metric, the Balance database produces consistent results.

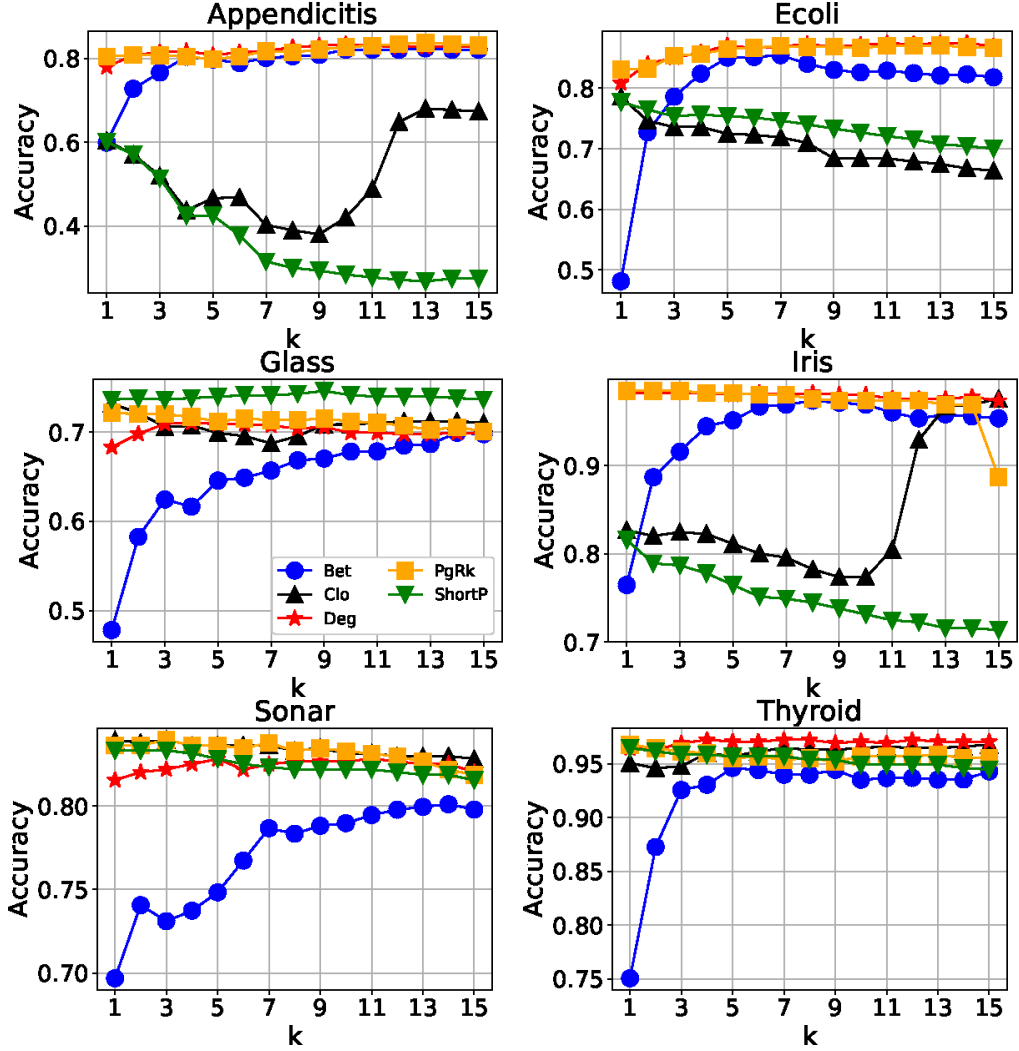


Figure 10 – Evaluation of the predictive accuracy of the network measures in relation to the various metrics under study.

The Wilcoxon test (DEMŠAR, 2006) was utilized to conduct a statistical analysis of results of Table 3. Using a degree of confidence of 95% ($\alpha = 0.05$), Table 4 ranks the network measures performance according to the number of favorable comparisons. As seen in the table, pagerank occupied the first position, closely followed by degree and closeness, while average shortest path length and betweenness occupied the last position.

4.3.2 Predictive robustness

Here, we analyze the predictive robustness of the network measures' ability to identify patterns in networked data. To guide our analyses, we frame our guiding question

Table 4 – Predictive capability study of the network measures with positive significant differences (wins) as determined by the statistical Wilcoxon test.

Pos	Measure	#Wins \uparrow	#Draws	#Losses
1	PageRank	2	2	0
2	Degree	1	3	0
	Closeness	1	3	0
4	Avg.Short.Path	0	2	2
	Betweenness	0	2	2

as: “How robust a given network measure may be in classification via importance characterization while maintaining a strong prediction capability as a function of parameter variation?”

First, the centrality metrics are analyzed in terms of the γ parameter. In general, shortest path, betweenness, and closeness are more sensitive to the variation of such a parameter, as illustrated in Fig. 11. In addition, the accuracy of closeness and shortest path typically degrades significantly for higher values of γ , such as 1, 2, and 5, indicating that such measures do not benefit from vast amounts of structural information. By the contrary, as evidenced by the Appendicitis, Iris, and Thyroid data sets, degree and pagerank are less sensitive to the change of such a parameter and may benefit from higher γ values.

In addition to $\gamma \in \{0, 0.5, 1, 2, 5\}$, we also evaluate the robustness of network measures as a function of the variation of the graph construction parameter $k \in \{1, 2, \dots, 15\}$. In other words, we assess how sensible a given network measure is based on the graph topology. This approach is interesting because it may provide network measures that are less dependent on complex parameter selection. Table 5 displays the average accuracy of each network measure over all possible parameter combinations. It demonstrates that degree and pagerank achieved the greatest results in five and three of eight data sets, respectively. In contrast, the other three network measures were unable to deliver any best result of robustness, including closeness, which our prior statistical test identified as one of the best measures in terms of predictive capability. The last line is the average of the column results.

Fig. 12 shows a graphical representation of the network measures for three datasets in function of the γ and k parameters. These three datasets as well the network measures were chosen in consideration of the greatest diversity of patterns. For the Appendicitis dataset, it is evident that: degree achieved competitive outcomes with higher γ values (1 and 5); betweenness also achieved the best results with the highest γ value (5); and closeness got the best results with the smallest γ value (0). For the Ecoli dataset, pagerank and degree produced the best results with $\gamma = 1$ and the worst results with $\gamma = 5$; betweenness also produced the best results with $\gamma = 1$; and closeness produced competitive

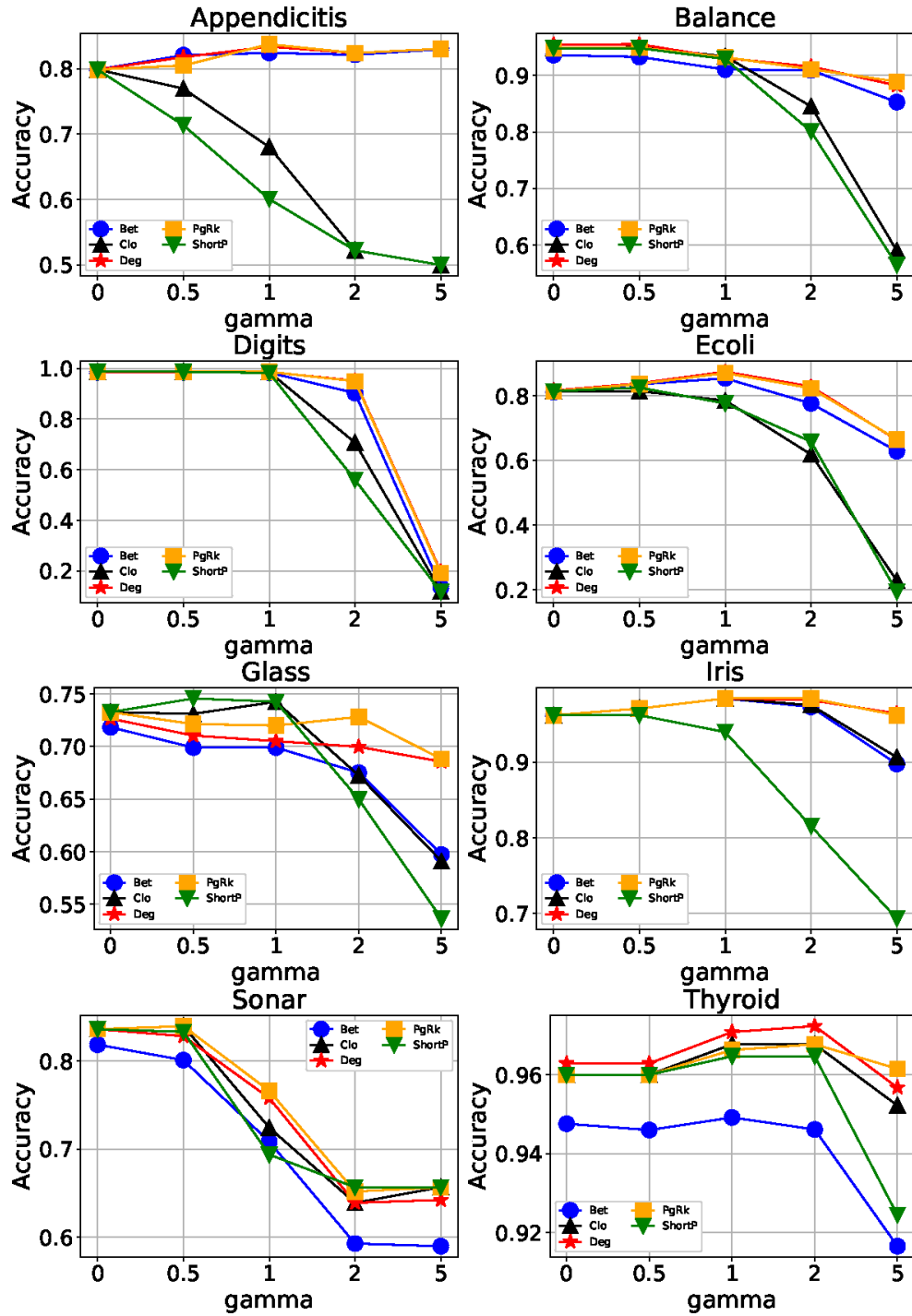


Figure 11 – Comparative evaluation of the variation of γ over the network measures on the data sets.

Table 5 – Robustness of the network measures in terms of averaged predictive capability along all parameter configurations. Best results are boldfaced.

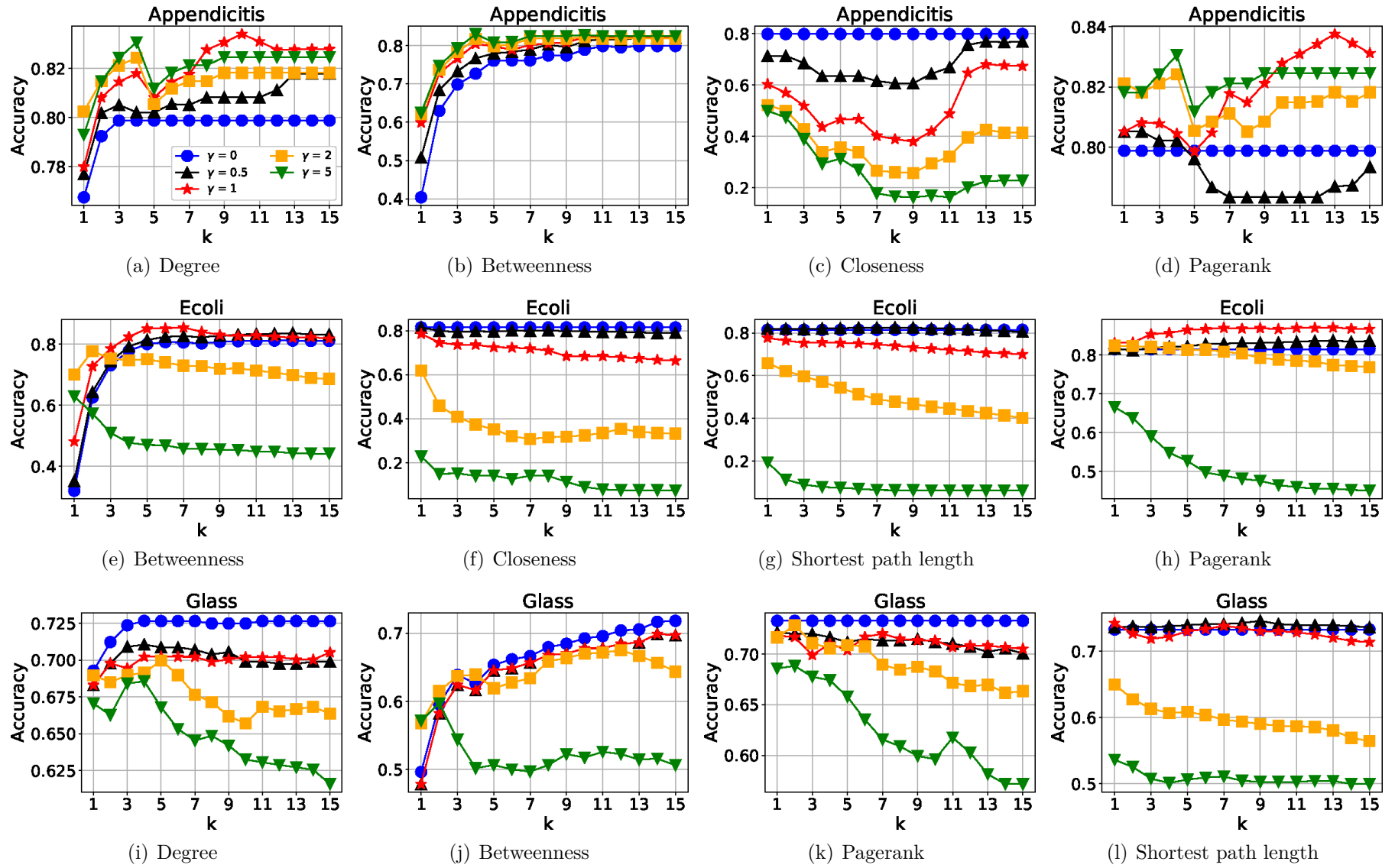
Dataset	\mathcal{PG}	\mathcal{BC}	\mathcal{CC}	\mathcal{DC}	\mathcal{SP}
App.	80.7±1.4	77.2±8.0	57.2±21.8	80.9±1.4	50.4±24.7
Bal.	91.9±3.7	87.3±8.3	77.6±27.0	92.0±3.8	81.1±17.7
Dig.	79.4±34.1	76.5±33.7	74.7±34.4	79.6±34.0	67.2±38.1
Eco.	77.1±12.2	71.5±14.6	60.3±27.0	77.1±12.4	62.7±27.2
Gla.	70.1±4.1	63.1±6.8	65.3±10.6	69.5±2.9	67.3±8.9
Iris	96.1±2.3	89.2±10.6	90.4±9.6	96.4±1.9	86.9±11.5
Son.	73.9±10.8	68.7±8.7	71.8±13.3	73.5±10.5	72.1±11.5
Thy.	95.8±0.6	91.3±4.9	95.6±1.5	95.9±1.1	94.7±2.7
Average	83.12	78.10	74.11	83.11	72.8

outcomes with lesser γ values (0 and 0.5). For the Glass dataset, the opposite happens: shortest path achieved competitive results with smaller values of γ (0, 0.5, and 1), whereas pagerank, degree, and betweenness achieved the best results with the smallest value of γ (0). Also, betweenness and degree typically produce better results for smaller k values.

Table 6 displays the results of the Wilcoxon test with a confidence level of 95% ($\alpha = 0.05$) for statistical analysis. The robustness of the metrics in the table is ranked by the number of favorable comparisons (“#Wins”). As in Table 4, pagerank and degree centrality ranked first here as well. Other than that, betweenness ranked worse in both predictive capability and robustness. Interestingly, despite earlier research not suggesting the adoption of the degree measure in the classification by pattern conformation (CARNEIRO; GAMA; RIBEIRO, 2021), our findings identified both pagerank and degree as the most suitable measures in the context of data classification via characterization of importance. We believe this may be inherently related to the bias of the characterization of importance, which differs from previous works, like conformation pattern, lies on the individual importance of each instance.

Table 6 – Analysis of the robustness of network metrics with substantial positive (“#Wins”) differences according to the Wilcoxon test.

Pos	Measure	#Wins \uparrow	#Draws	#Losses
1	PageRank	3	1	0
	Degree	3	1	0
3	Avg.Short.Path	0	2	2
	Closeness	0	2	2
	Betweenness	0	2	2

Figure 12 – Comparative evaluation of the variation of k over the network measures on the data sets.

4.4 Chapter summary

In conclusion, we provide a comprehensive investigation about the complex network measures considered in this study. The high-level classification via characterization of importance is one of the strategies capable of sophisticated data relationship analysis. As prior works have only explored pagerank in this context, this work contributes to the literature by investigating other network measures, including degree, betweenness, closeness, and shortest path length.

An experimental setup with multiple real-world datasets was provided to examine the potential of each network measure for high-level classification via importance characterization in order to study the measures. The results were evaluated based on both their predictive capability and their robustness. Statistical investigation validated pagerank as one of the most effective metrics and, interestingly, also identified the degree as performing at a comparable level. Therefore, both centrality measures are the most recommended according to our research, as they demonstrated the highest predictive potential and predictive robustness.

Since the most important aspect of high-level classification is to capture the data semantic characteristics in a given network, we can consider how to characterize importance in terms of various aspects of centrality, as opposed to assuming that all vertices have the same significance in the entire network. In order for high-level classification methods to be utilized in a broader range of applications, it is necessary to devise diverse measures for maximizing the capabilities of network features. According to our analysis, some centrality metrics show significantly superior predictive performance than others. Additionally, it is necessary to consider the computational complexity. In this regard, degree and pagerank offer several advantages over betweenness and closeness, as degree is a very simple metric and pagerank, in addition to being simple, can also account for dynamic relationships between vertices.

Finally, this research contributes to the field of network-based learning, namely in the design of supervised learning techniques employing new heuristic for high-level classification via characterization of importance.

GANet: A GA for network structural optimization

In this chapter, the GA was investigated for network construction and optimization, using the work (CARNEIRO et al., 2019) as a basis, whose main contribution was the development of a generic framework based on particle swarm for network formation optimization. Here, several configurations of GA were explored in real databases in order to obtain a good network configuration that can be applied together with the data classification via importance characterization.

5.1 Overview

To perform high-level classification on non-graph data (e.g., attribute vectors, images, text, etc.), the first step is to generate a network. The construction of the network is a crucial step, as the underlying information of the data is extracted from the graph for the classification process (CARNEIRO et al., 2019).

The objective of this study is to develop a solution based on GA for structural optimization of networks, that is, to find the most adequate configuration for representing the connections of instances in a network. The main motivation for the research is the efficient mechanism for representing discrete problems of GAs, which can naturally reflect the connections in a network. Thus, the investigated hypothesis states that GAs can provide more appropriate networks for classification via importance characterization than traditional network formation techniques, such as the k NN network, at the same time that they can be competitive compared to other state-of-the-art methods for allowing the direct manipulation of network configurations from the representation and manipulation of binary variables instead of continuous ones.

5.2 Experimental design

The Fig. 13 presents the GA-based network optimization method for network based classification, called GANet. The method is divided into two main phases: optimization and test. During training or optimization, from the training data, the GA is used to obtain a good network configuration in terms of a fitness function. In the test phase, the best solution obtained in training is used to create the network and applying the classification technique via importance characterization to classify the data.

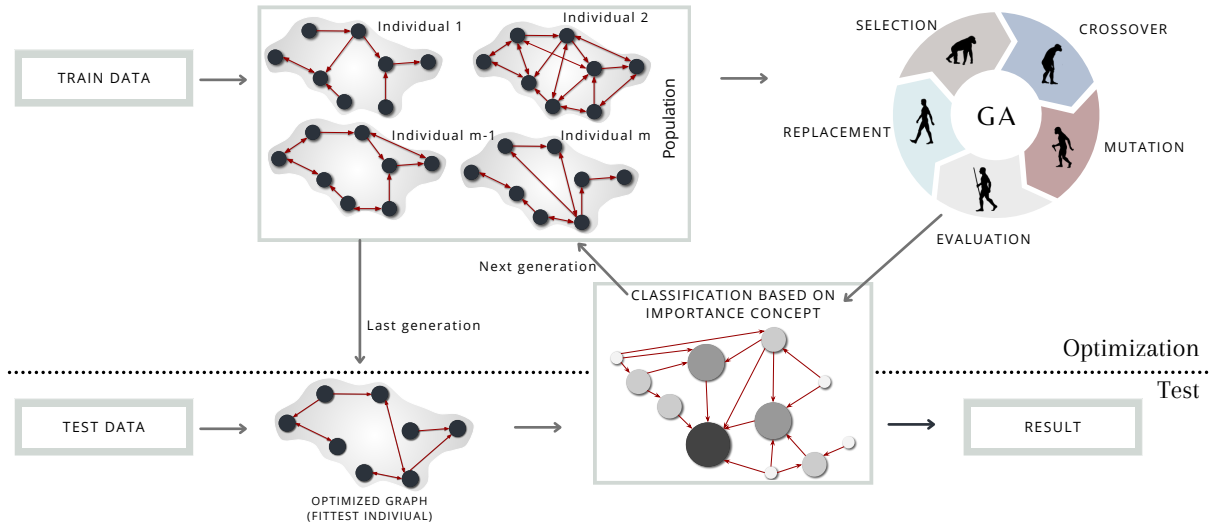


Figure 13 – Illustration of the structural optimization method via GA applied to network learning. Adapted from (FERNANDES; OLIVEIRA; CARNEIRO, 2023).

In a formal way, the GA manipulates a population of individuals $P = \{I_1, I_2, \dots, I_m\}$, in which each individual $I_i \in P$ is denoted by:

$$I_i = \{v_1, v_2, \dots, v_n\}, \quad (16)$$

where $v_i \in I_i$ represents the connections of a given vertex v_i (associated with an object $x_i \in X$) defined by:

$$v_i = \{e_{i1}, e_{i2}, \dots, e_{iq}\}, \quad (17)$$

where $j \in \{1, 2, \dots, q\}$ denotes the possible q connections of v_i and $e_{ij} \in \{0, 1\}$ the existence or absence of a given connection from vertex v_i to neighboring vertex Map_{ij} in the network, as illustrated in Fig. 14. The neighbors of each vertex v_i are defined based on the mapping heuristic *MapAll*, originally proposed in (CARNEIRO et al., 2019) and which defines the matrix *Map* through the following steps:

- Calculate the similarity between each pair of data items;
- Select for each vertex v_i its most similar q vertices;
- Given that $1 \leq z \leq q$, create the matrix $Map_{n \times q}$ such that:

$$Map_{iz} = \begin{cases} v_z & \text{se } l_i = l_z \\ \emptyset & \text{otherwise.} \end{cases} \quad (18)$$

where, $l \in \mathcal{L}$ is the instance label. Note that Map_{iz} is empty if the vertex v_i does not belong to the same class as v_z . From the formulation presented, it is also important to note that, unlike the continuous optimization method presented in (CARNEIRO et al., 2019), the GA developed here performs the optimization in a discrete space of solutions.

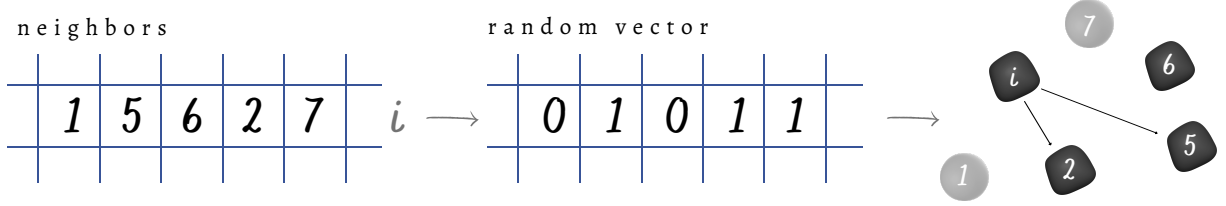


Figure 14 – Illustrative example of *Map-all* where from a random vector vertices of the same class are connected. In this example there is no connection between i and 7 vertices as they are of different classes, however, even if i and 6 vertices are of the same class there is no connection between them because there is no such connection in *Map-all* vector.

In Fig. 13 it can be seen that, first, the data is divided into training, validation and testing, and a random population is generated with m individuals based on the MapAll mapping heuristic from the training dataset. Each individual in the population goes through the process of evolution (selection, crossover and mutation) for m generations. During the optimization phase each individual \mathcal{I}_k is evaluated through a fitness function under the validation dataset. \mathcal{I}_k is converted to a network $\mathcal{G}_k = \{V_k, E_k\}$, where $V_k = \{1, \dots, n\}$ represents the vertices associated with each data item and E_k the edges between such vertices. The fitness function adopted was *the classification via importance characterization* proposed in (CARNEIRO; ZHAO, 2017) because it is a high-level classification technique that captures both topological and physical characteristics of the data and uses the pagerank measure to classify the object test on the most important component, and also because it has fewer parameters when compared to pattern conformation classification (please see section 3.4.1). In the test phase, the best individual obtained at the end of the m generations is used to map the optimized network to be applied for the high-level classification of the test items.

Regarding the genetic operators adopted by GANet, they are presented below:

Selection. Two methods were evaluated: roulette and tournament (YADAV; SOHAL, 2017). In *roulette*, the probability of each individual to be selected is given according to its fitness value. In the *tournament*, t individuals are randomly selected to form groups and then the best individual of each group is selected for the crossover step (YADAV; SOHAL, 2017).

Crossover. Two methods were evaluated: two points and uniform (KORA; YAD-LAPALLI, 2017). The *two-point* crossover is one of the simplest crossover methods in

which two cut points are randomly chosen at the same positions for the two parents, the genetic material is exchanged between the points and two new children are generated from them. In the *uniform* crossover, a vector of the size of the individual is randomly generated, indicating whether the gene will come from the parent 1 or 2 (child 1), and also the complement of this vector (child 2).

Mutation. New individuals are randomly selected for mutation: The bit that represents the connection between two vertices is switched, in other words, if there is a connection, it will be removed, and vice-versa.

Replacement. Two methods of reinsertion were evaluated for the replacement step: pure and ordered (MIRJALILI, 2019). In the *pure* reinsertion, only a percentage (*elit*) of the original population is kept, while the rest of the original individuals are replaced by the new generated individuals. In the *sorted* reinsertion the total population, both original and generated individuals, is evaluated and sorted, selecting the best *Tp*.

5.3 Analyses and results

To build an optimized network, we designed our experiment and employed data to compare among the conventional network construction technique *k*NNG and the PSO optimization technique. Here we present the results followed by a exploratory analyses.

For the accomplishment of the experiments, six real databases were considered. The actual databases used are publicly available in the UCI Machine Learning data repository (ASUNCION; NEWMAN, 2007) and are briefly presented in Table 7 through a meta-description of the data in terms of instance numbers, attributes and target classes numbers.

Table 7 – Brief description of the datasets being analyzed in terms of number of data items (*#Inst.*), attributes (*#Atrib.*) and number of classes (*#Classes*).

Nome	#Inst.	#Atrib.	#Classes
Iris	150	4	3
Teaching	151	5	3
Glass	214	9	7
Libras	360	90	15
Appendicitis	106	7	2
Balance	625	4	3

The experimental environment was developed in the Python language. Each experiment was performed five times. The parameters tested were γ and q , related to the mapping heuristic; two selection operators: tournament ($t = 3$) and roulette; two reinsertion operators: the ordered and the pure (*elit* = 20%); for the crossover, the two-point and the uniform were used. Regarding the network structure, the weighted graph was

considered and the pagerank network measure was employed as the importance concept, the population size and the number of generations used were 100, the crossover percentage was 80% and one bit switched in mutation phase with 30% of probability. The tested parameter combinations are shown in Table 8, totaling 16 different configurations.

Table 8 – Different configurations under investigation composed of the parameters γ , q , selection, reinsertion and crossover methods.

Configuration	γ value	q value	Selection	Replacement	Crossover
GANet-A	1	3	Tournament	Ordered	Two points
GANet-B	1	3	Roulette	Ordered	Two points
GANet-C	1	3	Tournament	Pure	Two points
GANet-D	1	3	Tournament	Ordered	Uniform
GANet-E	1	5	Tournament	Ordered	Two points
GANet-F	1	5	Roulette	Ordered	Two points
GANet-G	1	5	Tournament	Pure	Two points
GANet-H	1	5	Tournament	Ordered	Uniform
GANet-I	2	3	Tournament	Ordered	Two points
GANet-J	2	3	Roulette	Ordered	Two points
GANet-K	2	3	Tournament	Pure	Two points
GANet-L	2	3	Tournament	Ordered	Uniform
GANet-M	2	5	Tournament	Ordered	Two points
GANet-N	2	5	Roulette	Ordered	Two points
GANet-O	2	5	Tournament	Pure	Two points
GANet-P	2	5	Tournament	Ordered	Uniform

Table 9 – Accuracy (%) followed by standard deviation for the different configurations of the GANet model compared to the k NNG technique. The best results are in bold.

Alg.	Iris	Teaching	Glass	Libras	Appendicitis	Balance	Avg. Rank
k NNG	96.89 \pm 5.09	60.47 \pm 8.69	69.89 \pm 8.68	75.28 \pm 6.14	80.79 \pm 10.25	91.79 \pm 2.91	12.8 \pm 5.5
GANet-A	97.33 \pm 2.49	49.03 \pm 5.55	72.68 \pm 6.05	79.72 \pm 3.99	84.55 \pm 10.6	92.96 \pm 1.63	8.8 \pm 2.1
GANet-B	97.33 \pm 2.49	48.39 \pm 5.4	74.63 \pm 5.25	79.44 \pm 4.06	84.55 \pm 10.6	92.96 \pm 1.55	8.3 \pm 3.9
GANet-C	98.0 \pm 2.67	50.32 \pm 4.38	73.66 \pm 6.62	80.56 \pm 4.56	87.27 \pm 7.27	93.12 \pm 1.65	4.4 \pm 1.6
GANet-D	97.33 \pm 2.49	49.68 \pm 1.58	72.2 \pm 7.33	79.44 \pm 3.45	84.55 \pm 10.6	92.96 \pm 1.99	9.1 \pm 1.7
GANet-E	98.0 \pm 1.63	52.26 \pm 9.44	73.17 \pm 7.24	79.44 \pm 3.66	85.45 \pm 6.68	92.48 \pm 1.57	6.6 \pm 2.0
GANet-F	98.0 \pm 1.63	49.68 \pm 10.32	71.22 \pm 9.68	78.06 \pm 3.66	85.45 \pm 6.68	91.04 \pm 2.79	10.3 \pm 3.9
GANet-G	97.33 \pm 2.49	48.39 \pm 4.08	73.66 \pm 7.77	82.78 \pm 4.08	87.27 \pm 6.68	92.64 \pm 1.85	6.6 \pm 4.6
GANet-H	98.0 \pm 1.63	50.97 \pm 8.01	73.66 \pm 7.77	79.17 \pm 5.2	86.36 \pm 5.75	90.4 \pm 2.68	7.8 \pm 4.9
GANet-I	96.67 \pm 5.16	49.03 \pm 5.91	72.2 \pm 7.65	80.28 \pm 4.76	85.45 \pm 6.68	91.68 \pm 2.3	10.3 \pm 4.0
GANet-J	96.67 \pm 5.16	51.61 \pm 8.89	73.66 \pm 7.62	76.11 \pm 3.22	85.45 \pm 7.82	93.44 \pm 1.92	7.7 \pm 5.7
GANet-K	98.67 \pm 1.63	47.74 \pm 8.51	74.15 \pm 7.65	80.83 \pm 4.43	84.55 \pm 8.43	93.28 \pm 1.2	5.5 \pm 5.0
GANet-L	96.67 \pm 5.16	50.32 \pm 7.8	71.22 \pm 8.22	80.83 \pm 3.97	83.64 \pm 8.43	92.32 \pm 1.87	10.3 \pm 4.6
GANet-M	98.0 \pm 2.67	44.52 \pm 8.01	69.76 \pm 6.83	76.94 \pm 2.58	82.73 \pm 10.52	91.52 \pm 0.64	14.2 \pm 3.8
GANet-N	98.67 \pm 1.63	53.55 \pm 6.95	70.73 \pm 7.71	72.5 \pm 6.3	83.64 \pm 10.98	92.32 \pm 2.3	9.9 \pm 6.1
GANet-O	98.0 \pm 1.63	46.45 \pm 5.24	73.17 \pm 8.02	79.17 \pm 4.48	83.64 \pm 11.71	91.84 \pm 1.06	11.1 \pm 3.4
GANet-P	98.0 \pm 2.67	48.39 \pm 8.16	71.22 \pm 4.97	79.44 \pm 2.55	83.64 \pm 10.98	93.28 \pm 2.24	9.4 \pm 4.1

Table 9 presents the predictive performance achieved by the classification method using the k NN network and the proposed GA settings. With the exception of the Teaching base,

in all the other databases there was a considerable improvement in the result for several GA configurations.

The Friedman statistical test (DEMŠAR, 2006) was conducted to analyze the results considering a significance level $\alpha = 0.05$. The null hypothesis states that the predictive performance of the network construction methods are equivalent. Such a hypothesis is rejected by the test. Then, we adopted the Nemenyi post-test to identify which methods have a statistical difference. The post-test result is shown in Fig. 15, which presents the critical Nemenyi diagram. According to the figure, it is possible to observe that the configurations GANet-C, GANet-K, GANet-E and GANet-G have the best average rankings. In common, these four GA configurations use tournament selection and two-point crossover and were able to statistically outperform the k NN network and the GANet-M configuration. This result reveals the potential of GA-based network optimization architectures for data classification.

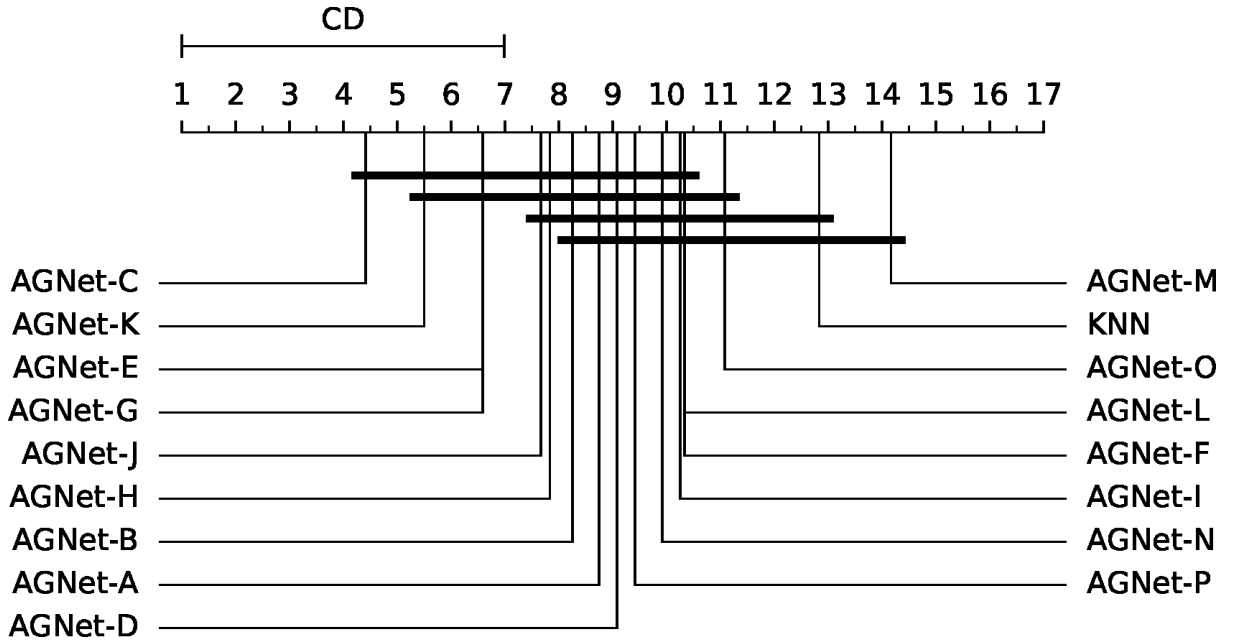


Figure 15 – Critical Nemenyi diagram comparing the average predictive performance ranking of analyzed models.

Then, we conducted another analysis related to the characterization of GANet networks. Fig. 16 shows the networks obtained by k NNG (left) and GA (right) respectively for the Iris and Teaching databases. In the figure, it is possible to observe that in addition to improving the predictive performance for most databases, the optimization process is also responsible for representing the relationships between vertices with fewer edges, which contributes to reducing the complexity of the process of classification. On the other hand, it also helps us to better explain the difficulty of the GA configurations for the Teaching base, which seems to be related to *overfitting* during the network optimization process.

In the Table 10 we can see the characterization of network obtained by k NNG; the two best GA configurations: GANet-C and GANet-K and by the worst configuration: the

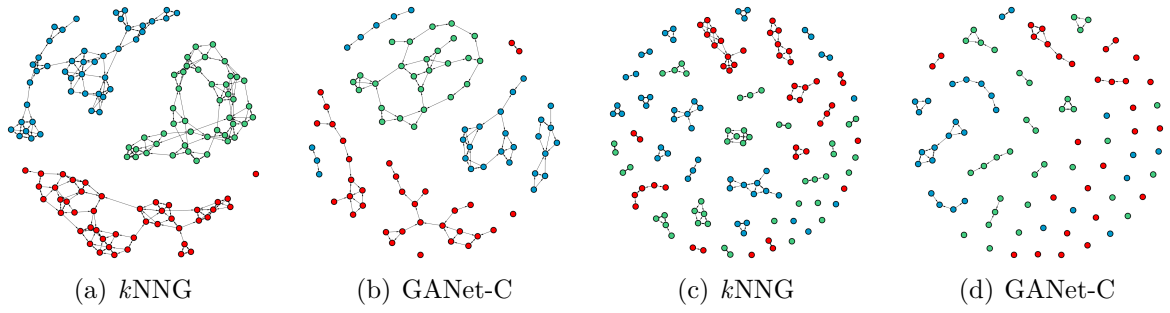


Figure 16 – Visual comparison between the networks obtained by k NNG and GANet-C, using the bases Iris (first two on the left) and Teaching (last two on the right).

GANet-M. All configurations in function of the assortativity measures (ASM), closeness (CLO), shortest mean path (SMP) and clustering coefficient (CLUS). In summary, the best GA configurations have higher CLO values and lower SMP values in common, while the k NN network provides networks with higher ASM and CLUS values. On the other hand, compared to GANet-C and GANet-K, the worst GA configuration (GANet-M) presented higher values of SMP and lower values of CLO. These results demonstrate the potential of the network optimization process to transform the structure and topology of the k NN network, in order to adapt the resulting network to the classification problem considered, contributing to a better predictive performance in most of the cases.

Table 10 – Measures summarizing the topological characteristics in terms of Mean Assortativity (ASM), Closeness (CLO), Shortest Mean Path (SMP), Clustering Coefficient (CLUS) in different network configurations. Techniques marked with “*” achieved the best result in the Table 9.

Dataset	Algs.	ASM	CLO	SMP	CLUS
Iris	k NNG	0.16	0.26	3.51	0.44
	GANet-C	0.01	0.47	1.33	0.27
	GANet-K*	-0.01	0.51	0.97	0.26
	GANet-M	0.03	0.33	3.00	0.36
Appendicitis	k NNG	0.34	0.28	5.33	0.43
	GANet-C*	-0.08	0.43	1.24	0.24
	GANet-K	0.11	0.53	0.94	0.24
	GANet-M	0.04	0.33	2.19	0.34
Teaching	k NNG*	0.42	0.82	0.84	0.63
	GANet-C	0.00	0.84	0.41	0.30
	GANet-K	0.12	0.85	0.49	0.45
	GANet-M	0.05	0.74	0.84	0.39

The Table 11 provides a comparison between the best GA configuration obtained by the present study (GANet-C) and PSOnet, the state-of-the-art method for network structural optimization. As can be seen, PSOnet achieves the best results for the Iris, Teaching and Balance databases, being surpassed by GANet-C for Glass, Libras and

Appendicitis. To statistically analyze the performance of both methods, the Wilcoxon test was performed. Considering a significance level $\alpha = 0.05$, the test shows that the performance of the models is equivalent. This is an interesting result, as it indicates that the networks obtained by the GA, in addition to presenting significantly better performance than those obtained by the kNN method, are competitive in relation to the networks obtained by the PSONet. Since PSONet is based on a sophisticated optimization technique for large-scale problems, the binary representation of the GA to manipulate the network configurations contributes to this result, which reduces the (finite) search space and facilitates the optimization process.

Table 11 – Comparison of the best configuration of GANet versus PSONet, a state-of-the-art method for structural optimization of networks.

Base de dados	GANet-C	PSONet (CARNEIRO et al., 2019)
Iris	98.00 \pm 2.67	100.0 \pm 0.00
Teaching	50.32 \pm 4.38	62.58 \pm 5.24
Glass	73.66 \pm 6.62	67.80 \pm 3.58
Libras	80.56 \pm 4.56	77.50 \pm 2.97
Appendicitis	87.27 \pm 7.27	82.72 \pm 1.81
Balance	93.12 \pm 1.65	95.36 \pm 0.93

5.4 Chapter summary

This study evaluates a GA for structural optimization of networks in the context of high-level data classification. Sixteen different configurations were explored in order to obtain a good network configuration that could be applied within the data classification technique via importance characterization. Statistical investigation conducted in six real datasets revealed that such a method can improve the extraction of underlying information in the data.

The results showed that networks optimized by our method were able to statistically outperform networks generated by the kNN network, the most adopted technique in the literature. The characterization of the generated networks also highlighted the ability of the proposed GA to explore network configurations with structure and topology considerably different from those obtained by the kNN network.

GANet for detection of autism spectrum disorder from vibrational spectroscopy salivary samples

The lack of auxiliary approaches that can support the clinical examination of autism spectrum disorder (ASD), such as biological diagnostic indicators makes the detection of ASD very difficult. Physicians often support medical decision on their comprehension of children's developmental history and behavior. ASD can be diagnosed as early as 18 months of age. Many children, however, do not obtain a proper diagnosis until they are adolescents or adults. This delay may disrupt early treatment of this developmental disorder, which may affect the maximum possible neurophysiological development provided by appropriate early therapy. In this particular research project, analytical models for the auxiliary diagnosis of ASD were developed. In this context, the GANet framework is used to detect ASD in a person by analyzing processed salivary data through the attenuated total reflectance Fourier-transform infrared (ATR-FTIR).

6.1 Overview

Autism spectrum disorder (ASD) is a group of neurodevelopmental conditions characterized by problems in social communication and extremely restricted and repetitive behavior or interests that manifest at an early age and its severity can range from mild to severe. The spectrum is vast, encompassing classic Kanner's syndrome and Asperger's syndrome. ASD affects, approximately, one in 100 children worldwide (ZEIDAN et al., 2022). Typically, ASD can be diagnosed as early as 18 to 24 months, when the specific ASD characteristics can be recognized (MALIK-SONI et al., 2022).

ASD individuals have atypical cognitive profiles, including impaired social cognition and perception, executive dysfunction, and atypical perceptual and information processing. Early developmental environmental influences, mutations and genetics are classical

components in the etiology of ASD (PAULSEN et al., 2022). Early, comprehensive, and targeted behavioral therapies can increase social communication while also reducing anxiety and aggressiveness, besides drugs can help with comorbid symptoms. It is essential to create a supportive environment that acknowledges and respects the individual's differences. The assessment must be multidisciplinary and developmental, also early ASD detection associated with properly intervention are crucial to support their development (SHAW et al., 2020).

The clinical assessment of behavioral indications and specific ASD characteristics is the sole basis for ASD diagnosis. In this way, it is very common that the diagnosis of ASD occurs late, especially in low- and middle-income countries where the population's access to doctors and other health professionals is restricted. However, research has indicated that early intervention helps children in overcoming some of the challenges they may face in life (ROGERS; VISMARA, 2008).

6.2 Attenuated total reflection-Fourier transform infrared spectroscopy

Attenuated total reflection-Fourier transform infrared (ATR-FTIR) spectroscopy is a strong and versatile analytical tool that has the capability to offer both qualitative and quantitative data with a high degree of chemical specificity. ATR-FTIR spectroscopy has been used frequently in both industry and research to characterize a wide range of materials, including complex mixtures because it is a relatively cheap and efficient analytical method. It is a flexible method for analyzing a wide variety of materials, including gasses, liquids, and solids, with minimal to no sample preparation, at a low cost per analysis, and in a short amount of time. The kinetics of chemical processes in solutions, especially liquid settings, can also be studied by ATR-FTIR spectroscopy.

The cost of collecting and analyzing body fluids in ATR-FTIR platforms is much lower when compared to neuroimaging and genomics research. Saliva samples have emerged as one of the most promising targets in the quest for distinctive biomarkers for ASD (QIAO et al., 2018; BOROUMAND et al., 2021), due to its accessibility and the unique biological information related to the ASD condition. However, the resulting spectra are complex and challenging to interpret because of the high dimensionality and spatial and temporal distribution data characteristics.

6.3 Data processing

The database was obtained with the approval of the Research Ethics Committee of the Federal University of Uberlândia, under protocol 249.200.9. A total of 53 people

(19 confirmed ASD cases and 34 controls) participated in this study. The control group matched the ASD group in terms of age, gender, and race, as well as male and female subjects were included in both groups. In addition, five subjects (26.31%) have attention deficit/hyperactivity disorder besides ASD. Clinical characteristics of subjects are described in Table 12. The data are in triplicate, resulting in three processed saliva samples per person.

Table 12 – Main demographic characteristics of the individuals, according to group: ASD (19) or control (34).

Feature	ASD	Control
Age (years) range	2 – 16	3 – 15
Male	73%	61%
Black	10.5%	2.9%
Pardo	15.8%	20.6%
White	73.7%	76.5%
Weight in birth (grams) range	800 – 4135	1400 – 3950
Attention-Deficit/ Hyperactivity Disorder (%)	26.31%	0%

A FTIR spectrum is a high-dimensional “fingerprint” of the chemical structure of a sample’s biological distinctive signals. Therefore, FTIR spectral fingerprints in conjunction with multivariate spectral analysis are commonly utilized for the characterization and identification of saliva biofluid. Typically, the multivariate data analysis of infrared spectra consists of spectral pre-processing and model design, such as clustering or classification.

Typically, in pre-processing, spectra with low signal-to-noise ratio and high noise are discarded, also spectra with low intensity of a relevant peak, such as the amide I peak, or spectra with high intensity of an undesirable or irrelevant peak(s), such as water vapor in samples are discarded.

In this work, the FTIR spectra data were pre-processed by performing normalization, smoothing, differentiation and truncation. The Fig. 17 illustrates the data change after the application of each pre-processing technique. Bellow are descriptions of each of these pre-processing methods:

1. **Normalization by the peak of amide (amide I):** Thinness or concentration discrepancies might sometimes be the most conspicuous source of spectrum variance between samples, concealing the biochemical differences of relevance. To reduce these impacts, the spectra are scaled to satisfy a specific criterion. Normalization to a particular peak can be applied. In this work, spectra normalization of each spectrum by the peak of amide 1 (highest value in the infrared band region between 1660 cm^{-1} and 1630 cm^{-1}), a region with potential for protein characterization is adopted.

2. **Smoothing with Savitzky-Golay filter (smoot):** In signal processing, the Savitzky-Golay filter is utilized to reduce noise in a signal and increase the trend smoothness of a signal. Based on polynomial degree and window size, the filter computes a polynomial fit for each window. One must be aware that smoothing might “smooth out” crucial information in spectra while simultaneously introducing distortions into the spectral data.
3. **Differentiation (diff):** Utilized both for baseline correction and to separate overlapping bands. In most cases, the gradients of the relevant spectral bands are far bigger than existing baselines, and differentiation serves to accentuate these contrasts.
4. **Spectrum truncation (trunc):** To avoid noise and outliers that can prevent the trained models from generalizing effectively the spectra are reduced to the range between 900cm^{-1} and 1800cm^{-1}

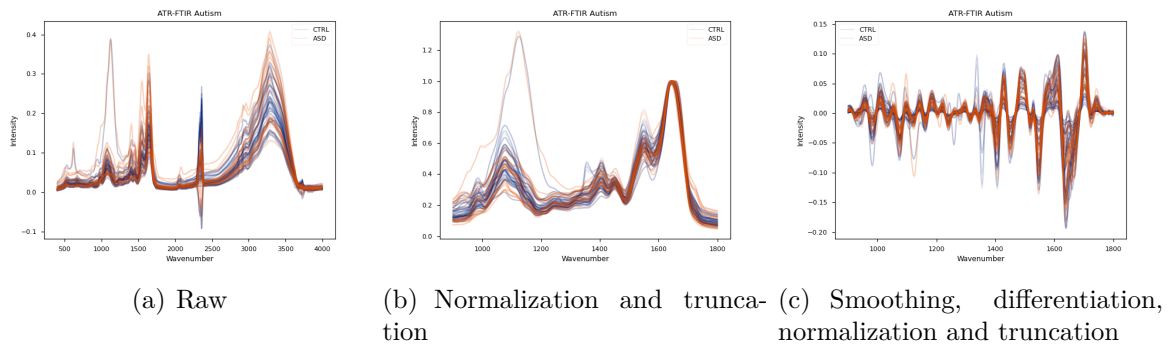


Figure 17 – Data pre-processing techniques plot of saliva samples where blue indicates control data and orange ASD data.

6.4 GA structural optimization for ASD detection

GA is a consolidated framework for solving optimization problems when their biological mechanism components are specifically constructed based on the problem to be solved. We meticulously designed the encoding strategy and the genetic operators in the proposed algorithm to ensure its accuracy in building network architectures within spectra data.

In the proposed encoding strategy, each vertex represents a spectra of the saliva ATR-FTIR sample and each individual represents the connections between the vertices. In other words, after a saliva sample is processed through ATR-FTIR, it produces a spectra result and this spectra is represented in a network as a vertex. When two vertices are bond, then they are linked by an edge, and this configuration is represented as an individual

in the GA. The Fig.18 illustrates the specific individual encoding. The 159 samples were splitted into 93 for training, 33 for validation and 33 for test, keeping in mind that, as there are 3 samples per person, they must all belong to the same group. The validation data are preserved to evaluate the model on different data than it was trained with. As there are 93 training samples, then there are 93 vertices in the network. Also, the genome or individual has 93 positions. Each position has a map representing potential links with other vertices or neighbors. The map size is predetermined by the q value.

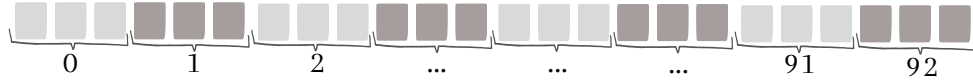


Figure 18 – Individual encoding for ASD detection on salivary FTIR data through GAs. In this example, the genome size is 93×3 . There are 93 positions for 93 salivary spectra samples. And there are three positions for possible connections with other vertices. The number of possible connections is pre-defined by the q parameter.

We used a population size of 100 individuals, therefore there are 100 candidate solutions. The initial population is created with random values, with 1 denoting the presence of an edge between the neighbors and 0 indicating the absence of such an edge. The Fig.19 shows an example of a population of possible solutions randomly initialized.

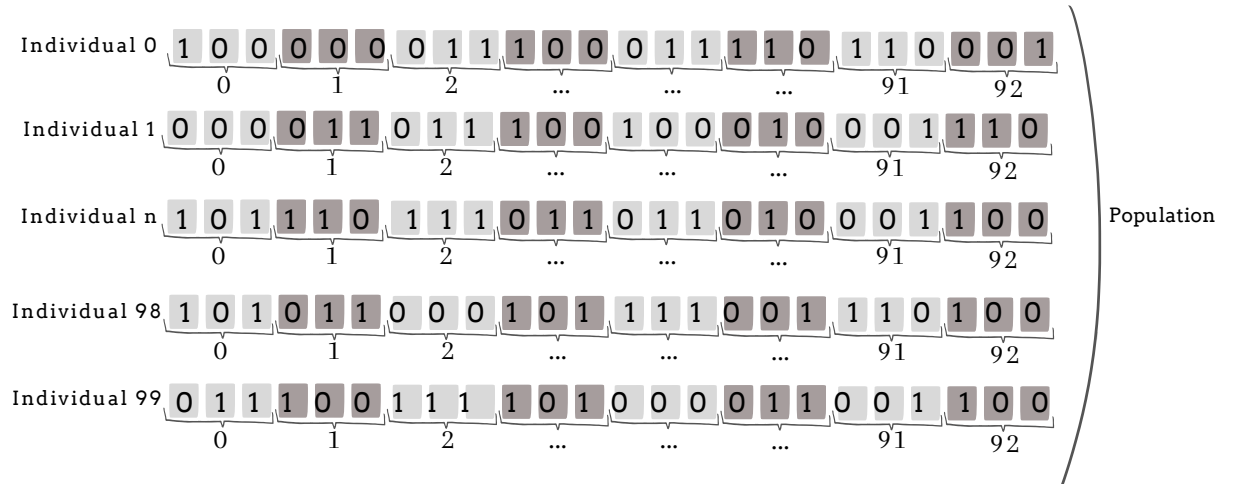


Figure 19 – The algorithm begins by initializing a random population of encodings that forms the initial generation of individuals. Example with a population of size 100. Individuals with genome size 93×3 .

The Fig. 20 shows an example of the mapping process. In the example q is equal to 3. A similarity metric is used to determine which vertices are more likely to be neighbors. We tested both Euclidean and Cosine metrics. In this line, each individual contains a dictionary to build up a network. The figure shows that vertex 0 has two connections to its nearest vertices, vertex 1 has one connection, and so on. When vertices belong to different classes, they do not connect.

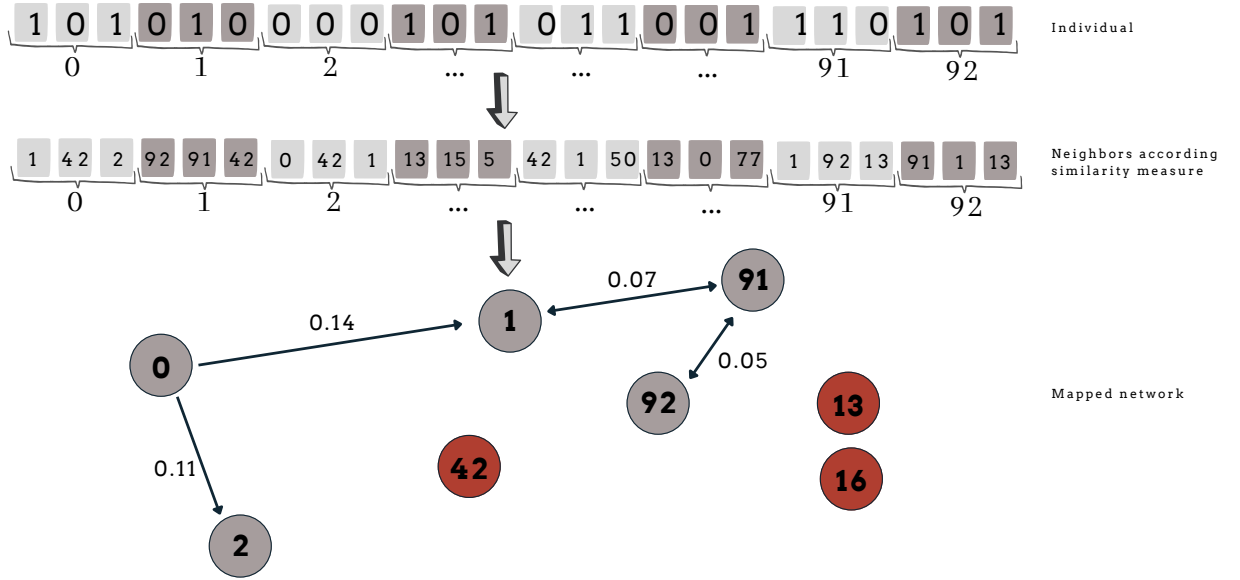


Figure 20 – Individual encoding representing a network. The neighbors are chosen based on a similarity metric. Whether the connection will occur or not is a stochastic process. Please note that the vertices must have the same target classes for the connections to occur. A practical example is presented above; vertex 92 has vertex 13 as a neighbor; despite the fact that in the individual there is a representation of a connection between them, they do not connect because they belong to different classes.

After a population is randomly initialized with the predetermined population size and the proposed encoding approach, during evolution, the fitness of each individual is evaluated. The method was designed to maximize the importance based classifier's accuracy as the fitness function to evaluate the best individuals to form the next generation. The parent individuals are then chosen based on their fitness, and the genetic operators, including the crossover and mutation operators, produce new offspring. Next, the current population is used to choose a population of individuals who will survive into the next generation through environmental selection. The current population consists of the parent population and the generated offspring population. Over the course of many evolutionary processes the maximum number of generations is reached and the best individual is generated. We employed tournament selection; ordered and pure reinsertion; and two points as crossover operators. As our experiments presented in section 5.3 showed these operators were statistically superior to k NNG and the other tested configurations. Please refer to section 2.3.2 for a description of the operator's behavior.

The Fig. 21 illustrates the proposed method's overall concept. First, (1) the saliva sample is collected, then (2) it is processed through the spectrometer, (3) the saliva spectra is pre-processed (refer to 6.3), then (4) the proposed GANet starts to evolve better-suited individuals over generations until its maximal number is reached. The optimized network (5) is used to classify the test data. And finally, (6) the GANet serves as complementary ASD detection.

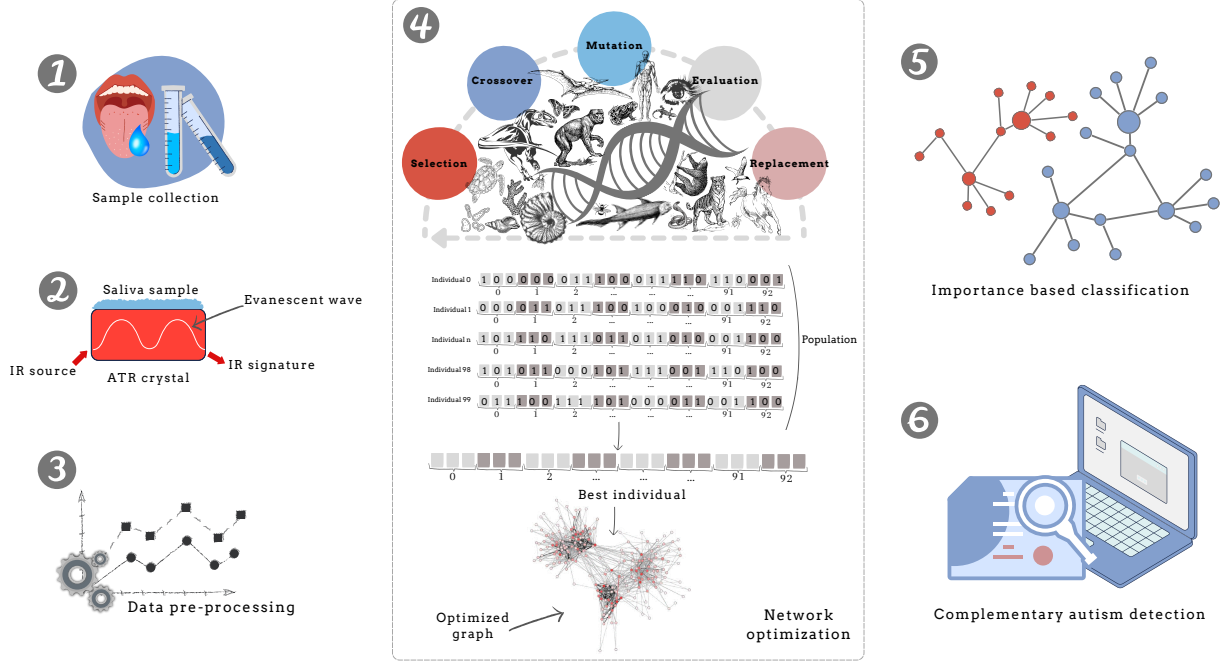


Figure 21 – GANet overall concept. (1) Saliva collection is obtained. (2) The samples are processed through the spectrometer. (3) Spectra data is pre-processed. (4) GANet evolves better-suited individuals. (5) High level data classification occurs. And, (6) GANet complements ASD detection. In step (4) there is an example of the proposed encoding strategy which represent the network. The final individual represents the evolved network, the best candidate solution. The encoding of this network consists of the connections between each salivary spectra. For each vertex, its codes are the map of the q value of possible neighbors. We use 3 and 5 to represent the neighbors' possible connections.

6.5 Results and Discussion

The four best GANet configurations obtained in the previous chapter (Fig. 15) were employed here: GANet-C, GANet-E, GANet-G and GANet-K. Please refer to Table 8 for information regarding their configuration and operators. In addition, besides the Euclidean metric, we also analyzed the Cosine metric in these simulations. Each experiment was conducted over a training, validation and test sets in a total of five executions. The training set is used to learn the models, the validation set is used to determine the predictive parameters and the test set is used to evaluate their performance.

Table 13 presents the results obtained by GANet configurations considering the degree network measure as well as Euclidean and cosine proximity metrics. The results shows that GANet-E equipped with Euclidean metric and Amide I normalization obtained the best results. The same simulations performed with degree (\mathcal{DC} measure were also conducted with pagerank (\mathcal{PG}), although its performance was clearly inferior, as shown by Table 14.

Besides the comparison between the results obtained by GANet configurations with degree (GANet-E+ \mathcal{DC}) and pagerank (GANet-E+ \mathcal{PG}), we also compared both network measures against the results obtained by LDA and SVM classifiers. LDA has been widely

Table 13 – Results using *degree* in terms of accuracy, sensitivity, specificity and the harmonic mean of sensitivity and specificity. The best results are boldfaced. Results of the overall best performance are underlined.

Metric	Config.	Pre-processing	Accuracy	Sensitivity	Specificity	H. Mean
Cosine	GANet-C	Amide I	0.65	0.13	1.00	0.29
		Smoot., diff., norm.	0.65	0.63	0.70	0.65
	GANet-E	Amide I	0.68	0.51	0.83	0.64
		Smoot., diff., norm.	0.63	0.57	0.70	0.62
	GANet-G	Amide I	0.63	0.13	0.96	0.29
		Smoot., diff., norm.	0.65	0.63	0.70	0.65
	GANet-K	Amide I	0.63	0.13	0.97	0.29
		Smoot., diff., norm.	0.63	0.59	0.70	0.63
Euclidean	GANet-C	Amide I	0.65	0.13	1.00	0.29
		Smoot., diff., norm.	0.64	0.65	0.66	0.64
	<u>GANet-E</u>	Amide I	<u>0.78</u>	<u>0.61</u>	<u>0.90</u>	<u>0.74</u>
		Smoot., diff., norm.	0.64	0.65	0.66	0.64
	GANet-G	Amide I	0.67	0.16	1.00	0.34
		Smoot., diff., norm.	0.64	0.65	0.66	0.64
	GANet-K	Amide I	0.64	0.16	0.97	0.33
		Smoot., diff., norm.	0.66	0.65	0.68	0.66

Table 14 – Classification results in terms of accuracy, sensitivity, specificity and the harmonic mean of sensitivity and specificity. The best results are boldfaced. Results of the overall best performance are underlined.

Algorithm	Pre-processing	Accuracy	Sensitivity	Specificity	H. Mean
GANet-E+ \mathcal{DC} (Euclidean)	Amide I	<u>0.78</u>	<u>0.61</u>	<u>0.90</u>	<u>0.74</u>
	Smoot., diff., norm.	0.64	0.65	0.66	0.64
GANet-E+ \mathcal{PG} (Cosine)	Amide I	0.61	0.18	0.86	0.35
	Smoot., diff., norm.	0.58	0.30	0.74	0.46
SVM	Amide I	0.68	0.11	0.99	0.25
	Smoot., diff., norm.	0.56	0.20	0.88	0.37
LDA	Amide I	0.58	0.40	0.68	0.52
	Smoot., diff., norm.	0.61	0.50	0.64	0.57

used in the context of FTIR data classification (MORAIS et al., 2020), while SVM has been considered a state-of-the-art classifier by the literature of ATR-FTIR spectroscopy (DEISS et al., 2020). According to Table 14, GANet-E+ \mathcal{DC} achieved the best predictive performance with 0.74 of harmonic mean between sensitivity and specificity, which is a quite attractive result. One can see no other technique was even close to that result. LDA achieved reasonable results, with 0.57 of harmonic mean, and GANet-E+ \mathcal{PG} and SVM presented the worse results, even after tuned.

Both pagerank and degree centrality are measures of vertex importance in a network; however, they differ in their definitions and applications. Degree centrality is a local measure that calculates the number of direct connections a vertex has in the network. In contrast, pagerank is a global measure that considers the distribution of connections between vertices when assigning scores. In the ASD detection via ATR-FTIR samples, it seems that the local aspects of degree contributes to achieve the best results. We believe such a local analysis is less affected by the overlapping between the samples of different groups. Indeed, SVM also had troubles to fit such data, even with a gaussian kernel.

6.6 Chapter summary

This study evaluates the proposed GANet for structural network optimization in the context of ASD spectra data classification. Four different GANet configurations, GANet-C, GANet-E, GANet-G and GANet-K, were compared to two low-level approaches, SVM and LDA. The outcomes demonstrated that the suggested method outperforms both low-level techniques, proving the GANet’s power to build up a network that can detect the underlying data relationship. In addition, some data preprocessing methods and network measures were tested.

Conclusion

In this paper, we propose an optimization network method for network based classification via the importance concept using GA algorithms. This chapter gives an overview of some of the hypotheses investigated in this work as well as discusses possible future directions to extend the work in this dissertation and concludes.

7.1 Main findings

In this dissertation, we developed two cooperative approaches toward bio-inspired optimization techniques with heuristics based on complex networks for high-level data classification based on importance characterization and there are many fascinating opportunities for future work that follow from the research presented.

The three hypotheses from Chapter 1 are summarized to guide the primary findings discussion of this research work:

H_1 : *Degree, betweenness, closeness and shortest path length are network measures able to perform the characterization of importance at the same level of pagerank.*

H_2 : *GA for graph construction provides optimized networks competitive with traditional and state-of-the-art methods of graph construction.*

H_3 : *The proposed method, GANet, has the potential to improve the detection of autism using recorded spectra by Fourier Transform Infrared Spectroscopy.*

H_1 : The use of network measures as heuristics for characterizing the relevance of a vertex in a network component for high-level classification via importance characterization technique is investigated by applying the a variety of network measures, including pagerank, degree, betweenness, closeness and shortest path length. In Chapter 4, an experimental setup with several real-world data sets was provided to evaluate the potential of each network measure. Predictive capability and robustness were considered when evaluating the results. Statistical study validated pagerank as one of the better measures, and, curiously, also identified the degree as performing at a close level. Therefore, both

centrality measures were the most suggested by our study, as they demonstrated the best performance in terms of both predictive capability and predictive robustness.

H₂: Structural network optimization is an important problem which needs to be solved in order to achieve higher classification results. The forward problem, on the other hand, is challenging. The traditional network construction methods rely on simple criteria such as using the same number of connections for all vertices. Over billions of years, nature has optimized complex bio-structures to have incredible durability and strength. Thus, bio-inspired techniques have been used for network construction, in an effort to discover more suitable connections while maximizing a fitness function. In this sense, in Chapter 5, we investigated different genetic algorithm configurations and extensive experimental results validated the effectiveness of our proposed method. Further, the Wilcoxon test demonstrated that the characterization of the generated networks through GANet approach offered substantially superior performance than the k NNG, the most adopted technique in the literature, and that the GANet method was comparable to the PSO-Net, the state-of-the-art. Thus, the discrete space representation through GAs proved to be an efficient mechanism.

H₃: The potential of GANet was studied to classify whether or not the person has ASD. Attenuated total reflectance (ATR) enabled the salivary samples to be evaluated as a simple and cheap technology. However, the characterization of the ASD spectra data is a complex task that requires exhaustive experiments. The spectral data were subjected to a preliminary elaboration based on normalization by the peak of amide I, smoothing with Savitzky-Golay filter, differentiation and Spectrum truncation to reduce the noise and extract larger analytical information. In Chapter 6, the four best GANet configurations were tested against the LDA method, the most used method in FTIR literature and against the SVM, the state-of-the-art. The set was built by using the IR data from 159 samples scanned in reflectance mode, and the range 1800–900 was selected because furnishing the most useful analytical information. The experiments examined in this dissertation revealed GA’s ability to classify a sample as ASD or not.

In spite of the progress that has been made, there are still some areas that require additional research. In this sense there are some future directions:

- ❑ Investigation of alternative GA configurations, such as rank selection and cycle crossover (KATOCH; CHAUHAN; KUMAR, 2021) and multi-objective genetic algorithms (AGMO) (KONAK; COIT; SMITH, 2006).
- ❑ Analysis of other bio-inspired algorithms, such as the ant colony optimization (DORIGO; BIRATTARI; STUTZLE, 2006) and frog leaping algorithm (EUSUFF; LANSEY; PASHA, 2006).
- ❑ Due to the FTIR data characteristics, a more adequate technique for the network construction may be needed, for example one aimed at a time series/sequential

pattern, such as visibility graph (LACASA et al., 2008).

7.2 Bibliographic contributions

This work resulted in the following bibliographic production:

- ❑ FERNANDES, J. M.; OLIVEIRA, G. M. B. de; CARNEIRO, M. G. Network optimization based on genetic algorithm for high-level data classification. **IEEE Latin America Transactions**, IEEE, v. 21, n. 2, p. 295–301, 2023. Disponível em: <<https://doi.org/10.1109/TLA.2023.10015222>>.
- ❑ FERNANDES, J. M.; SUZUKI, G.; ZHAO, L.; CARNEIRO, M. G. Data classification via centrality measures of complex networks (accepted). **2023 International Joint Conference on Neural Networks (IJCNN)**, IEEE, 2023.

Also, the following related papers were submitted in collaboration with colleagues:

- ❑ LIMA, R.; FERNANDES, J. M.; SILVA, R. S.; CARNEIRO, M. G. Molecular detection of oral cancer using complex networks. 2023.
- ❑ M, F. C.; FERNANDES, J. M.; SILVA, R. S.; CARNEIRO, M. G. Ocanspectra: an oral cancer detection system from salivary atr-ftir spectroscopy. 2023.

Bibliography

ALYASSERI, Z. A. A.; AL-BETAR, M. A.; DOUSH, I. A.; AWADALLAH, M. A.; ABASI, A. K.; MAKHADMEH, S. N.; ALOMARI, O. A.; ABDULKAREEM, K. H.; ADAM, A.; DAMASEVICIUS, R. et al. Review on covid-19 diagnosis models based on machine learning and deep learning approaches. **Expert systems**, Wiley Online Library, v. 39, n. 3, p. e12759, 2022. Disponível em: <<https://doi.org/10.1111/exsy.12759>>.

ASUNCION, A.; NEWMAN, D. **UCI machine learning repository**. 2007.

BANSAL, J. C. Particle swarm optimization. In: **Evolutionary and swarm intelligence algorithms**. Springer, 2019. p. 11–23. Disponível em: <https://doi.org/10.1007/978-3-319-91341-4_2>.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. Disponível em: <<https://doi.org/10.1126/science.286.5439.509>>.

BARLOW, H. B. Unsupervised learning. **Neural computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 1, n. 3, p. 295–311, 1989. Disponível em: <<https://doi.org/10.1162/neco.1989.1.3.295>>.

Bertini Jr, J. R.; ZHAO, L.; MOTTA, R.; LOPES, A. de A. A nonparametric classification method based on k-associated graphs. **Information Sciences**, Elsevier, v. 181, n. 24, p. 5435–5456, 2011. Disponível em: <<https://doi.org/10.1016/j.ins.2011.07.043>>.

BIGGS, N.; LLOYD, E. K.; WILSON, R. J. **Graph Theory, 1736-1936**. [S.l.]: Oxford University Press, 1986.

BOROUMAND, M.; OLIANAS, A.; CABRAS, T.; MANCONI, B.; FANNI, D.; FAA, G.; DESIDERIO, C.; MESSANA, I.; CASTAGNOLA, M. Saliva, a bodily fluid with recognized and potential diagnostic applications. **Journal of Separation Science**, Wiley Online Library, v. 44, n. 19, p. 3677–3690, 2021. Disponível em: <<https://doi.org/10.1002/jssc.202100384>>.

BRANDES, U. A faster algorithm for betweenness centrality. **Journal of mathematical sociology**, Taylor & Francis, v. 25, n. 2, p. 163–177, 2001. Disponível em: <<https://doi.org/10.1080/0022250X.2001.9990249>>.

- BRANDES, U. On variants of shortest-path betweenness centrality and their generic computation. **Social Networks**, Elsevier, v. 30, n. 2, p. 136–145, 2008. Disponível em: <<https://doi.org/10.1016/j.socnet.2007.11.001>>.
- CAI, H.; ZHENG, V. W.; CHANG, K. C.-C. A comprehensive survey of graph embedding: Problems, techniques, and applications. **IEEE transactions on knowledge and data engineering**, IEEE, v. 30, n. 9, p. 1616–1637, 2018. Disponível em: <<https://doi.org/10.1109/TKDE.2018.2807452>>.
- CARBONELL, J. G.; MICHALSKI, R. S.; MITCHELL, T. M. An overview of machine learning. **Machine learning**, Elsevier, p. 3–23, 1983. Disponível em: <<https://doi.org/10.1016/B978-0-08-051054-5.50005-4>>.
- CARNEIRO, M. G. **Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural**. Tese (Doutorado) — Universidade de São Paulo, 2017.
- CARNEIRO, M. G.; CHENG, R.; ZHAO, L.; JIN, Y. Particle swarm optimization for network-based data classification. **Neural Networks**, Elsevier, v. 110, p. 243–255, 2019. Disponível em: <<https://doi.org/10.1016/j.neunet.2018.12.003>>.
- CARNEIRO, M. G.; CUPERTINO, T. H.; CHENG, R.; JIN, Y.; ZHAO, L. Nature-inspired graph optimization for dimensionality reduction. In: IEEE. **2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)**. 2017. p. 1113–1119. Disponível em: <<https://doi.org/10.1109/ICTAI.2017.00170>>.
- CARNEIRO, M. G.; GAMA, B. C.; RIBEIRO, O. S. Complex network measures for data classification. In: IEEE. **2021 International Joint Conference on Neural Networks (IJCNN)**. 2021. p. 1–8. Disponível em: <<https://doi.org/10.1109/IJCNN52387.2021.9533608>>.
- CARNEIRO, M. G.; ZHAO, L. Organizational data classification based on the importance concept of complex networks. **IEEE transactions on neural networks and learning systems**, IEEE, v. 29, n. 8, p. 3361–3373, 2017. Disponível em: <<https://doi.org/10.1109/TNNLS.2017.2726082>>.
- CARNEIRO, M. G.; ZHAO, L. Analysis of graph construction methods in supervised data classification. In: IEEE. **2018 7th Brazilian Conference on Intelligent Systems (BRACIS)**. [S.l.], 2018. p. 390–395.
- CARNEIRO, M. G.; ZHAO, L.; CHENG, R.; JIN, Y. Network structural optimization based on swarm intelligence for highlevel classification. In: IEEE. **2016 International Joint Conference on Neural Networks (IJCNN)**. 2016. p. 3737–3744. Disponível em: <<https://doi.org/10.1109/IJCNN.2016.7727681>>.
- CHENG, R.; JIN, Y. A social learning particle swarm optimization algorithm for scalable optimization. **Information Sciences**, Elsevier, v. 291, p. 43–60, 2015. Disponível em: <<https://doi.org/10.1016/j.ins.2014.08.039>>.
- COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE transactions on information theory**, IEEE, v. 13, n. 1, p. 21–27, 1967. Disponível em: <<https://doi.org/10.1109/TIT.1967.1053964>>.

- CUPERTINO, T. H.; CARNEIRO, M. G.; ZHENG, Q.; ZHANG, J.; ZHAO, L. A scheme for high level data classification using random walk and network measures. **Expert Systems with Applications**, Elsevier, v. 92, p. 289–303, 2018. Disponível em: <<https://doi.org/10.1016/j.eswa.2017.09.014>>.
- DARWISH, A. Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications. **Future Computing and Informatics Journal**, Elsevier, v. 3, n. 2, p. 231–246, 2018. Disponível em: <<https://doi.org/10.1016/j.fcij.2018.06.001>>.
- DEISS, L.; MARGENOT, A. J.; CULMAN, S. W.; DEMYAN, M. S. Tuning support vector machines regression models improves prediction accuracy of soil properties in mir spectroscopy. **Geoderma**, Elsevier, v. 365, p. 114227, 2020. Disponível em: <<https://doi.org/10.1016/j.geoderma.2020.114227>>.
- DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine learning research**, v. 7, n. Jan, p. 1–30, 2006.
- DHIMAN, G. Esa: a hybrid bio-inspired metaheuristic optimization approach for engineering problems. **Engineering with Computers**, Springer, v. 37, n. 1, p. 323–353, 2021. Disponível em: <<https://doi.org/10.1007/s00366-019-00826-w>>.
- DOGAN, A.; BIRANT, D. Machine learning and data mining in manufacturing. **Expert Systems with Applications**, Elsevier, v. 166, p. 114060, 2021. Disponível em: <<https://doi.org/10.1016/j.eswa.2020.114060>>.
- DORIGO, M.; BIRATTARI, M.; STUTZLE, T. Ant colony optimization. **IEEE computational intelligence magazine**, IEEE, v. 1, n. 4, p. 28–39, 2006. Disponível em: <<https://doi.org/10.1109/CI-M.2006.248054>>.
- ELSISI, M.; MAHMOUD, K.; LEHTONEN, M.; DARWISH, M. M. Reliable industry 4.0 based on machine learning and iot for analyzing, monitoring, and securing smart meters. **Sensors**, MDPI, v. 21, n. 2, p. 487, 2021. Disponível em: <<https://doi.org/10.3390/s21020487>>.
- ENGELN, J. E. V.; HOOS, H. H. A survey on semi-supervised learning. **Machine Learning**, Springer, v. 109, n. 2, p. 373–440, 2020. Disponível em: <<https://doi.org/10.1007/s10994-019-05855-6>>.
- EUSUFF, M.; LANSEY, K.; PASHA, F. Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization. **Engineering optimization**, Taylor & Francis, v. 38, n. 2, p. 129–154, 2006. Disponível em: <<https://doi.org/10.1080/03052150500384759>>.
- FERNANDES, J.; BISSARO, L.; SANTOS, F.; CARNEIRO, M. Deep learning in risk assessment. In: SBC. **Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional**. 2019. p. 1068–1079. Disponível em: <<https://doi.org/10.5753/eniac.2019.9358>>.
- FERNANDES, J. M.; OLIVEIRA, G. M. B. de; CARNEIRO, M. G. Network optimization based on genetic algorithm for high-level data classification. **IEEE Latin America Transactions**, IEEE, v. 21, n. 2, p. 295–301, 2023. Disponível em: <<https://doi.org/10.1109/TLA.2023.10015222>>.

FERNANDES, J. M.; SUZUKI, G.; ZHAO, L.; CARNEIRO, M. G. Data classification via centrality measures of complex networks (accepted). **2023 International Joint Conference on Neural Networks (IJCNN)**, IEEE, 2023.

GABOW, H. N.; GALIL, Z.; SPENCER, T.; TARJAN, R. E. Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. **Combinatorica**, Springer, v. 6, n. 2, p. 109–122, 1986. Disponível em: <<https://doi.org/10.1007/BF02579168>>.

GREENER, J. G.; KANDATHIL, S. M.; MOFFAT, L.; JONES, D. T. A guide to machine learning for biologists. **Nature Reviews Molecular Cell Biology**, Nature Publishing Group, v. 23, n. 1, p. 40–55, 2022. Disponível em: <<https://doi.org/10.1038/s41580-021-00407-0>>.

HOLLAND, J. H. Genetic algorithms. **Scientific american**, JSTOR, v. 267, n. 1, p. 66–73, 1992. Disponível em: <<https://doi.org/10.1038/scientificamerican0792-66>>.

HUANG, H.-Y.; BROUGHTON, M.; MOHSENI, M.; BABBUS, R.; BOIXO, S.; NEVEN, H.; MCCLEAN, J. R. Power of data in quantum machine learning. **Nature communications**, Nature Publishing Group, v. 12, n. 1, p. 1–9, 2021. Disponível em: <<https://doi.org/10.1038/s41467-021-22539-9>>.

KATOCH, S.; CHAUHAN, S. S.; KUMAR, V. A review on genetic algorithm: past, present, and future. **Multimedia Tools and Applications**, Springer, v. 80, p. 8091–8126, 2021. Disponível em: <<https://doi.org/10.1007/s11042-020-10139-6>>.

KENNEDY, J.; EBERHART, R. Particle swarm optimization. In: IEEE. **Proceedings of ICNN'95-international conference on neural networks**. [S.l.], 1995. v. 4, p. 1942–1948.

KONAK, A.; COIT, D. W.; SMITH, A. E. Multi-objective optimization using genetic algorithms: A tutorial. **Reliability engineering & system safety**, Elsevier, v. 91, n. 9, p. 992–1007, 2006. Disponível em: <<https://doi.org/10.1016/j.res.2005.11.018>>.

KORA, P.; YADLAPALLI, P. Crossover operators in genetic algorithms: A review. **International Journal of Computer Applications**, Foundation of Computer Science, v. 162, n. 10, 2017. Disponível em: <<https://doi.org/10.5120/ijca2017913370>>.

KOTARY, D. K.; NANDA, S. J.; GUPTA, R. A many-objective whale optimization algorithm to perform robust distributed clustering in wireless sensor network. **Applied Soft Computing**, Elsevier, v. 110, p. 107650, 2021. Disponível em: <<https://doi.org/10.1016/j.asoc.2021.107650>>.

LACASA, L.; LUQUE, B.; BALLESTEROS, F.; LUQUE, J.; NUNO, J. C. From time series to complex networks: The visibility graph. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 105, n. 13, p. 4972–4975, 2008. Disponível em: <<https://doi.org/10.1073/pnas.0709247105>>.

LANGVILLE, A. N.; MEYER, C. D. Deeper inside pagerank. **Internet Mathematics**, Taylor & Francis, v. 1, n. 3, p. 335–380, 2004. Disponível em: <<https://doi.org/10.1080/15427951.2004.10129091>>.

LANGVILLE, A. N.; MEYER, C. D. **Google's PageRank and beyond: The science of search engine rankings**. [S.l.]: Princeton university press, 2011.

LAWSON, C. E.; MARTÍ, J. M.; RADIVOJEVIC, T.; JONNALAGADDA, S. V. R.; GENTZ, R.; HILLSON, N. J.; PEISERT, S.; KIM, J.; SIMMONS, B. A.; PETZOLD, C. J. et al. Machine learning for metabolic engineering: A review. **Metabolic Engineering**, Elsevier, v. 63, p. 34–60, 2021. Disponível em: <<https://doi.org/10.1016/j.ymben.2020.10.005>>.

LEI, Y.; YANG, B.; JIANG, X.; JIA, F.; LI, N.; NANDI, A. K. Applications of machine learning to machine fault diagnosis: A review and roadmap. **Mechanical Systems and Signal Processing**, Elsevier, v. 138, p. 106587, 2020. Disponível em: <<https://doi.org/10.1016/j.ymssp.2019.106587>>.

LI, M.; LIU, R.-R.; LÜ, L.; HU, M.-B.; XU, S.; ZHANG, Y.-C. Percolation on complex networks: Theory and application. **Physics Reports**, Elsevier, v. 907, p. 1–68, 2021. Disponível em: <<https://doi.org/10.1016/j.physrep.2020.12.003>>.

LIMA, R.; FERNANDES, J. M.; SILVA, R. S.; CARNEIRO, M. G. Molecular detection of oral cancer using complex networks. 2023.

LIU, Y.; ARUNACHALAM, S.; TEMME, K. A rigorous and robust quantum speed-up in supervised machine learning. **Nature Physics**, Nature Publishing Group, v. 17, n. 9, p. 1013–1017, 2021. Disponível em: <<https://doi.org/10.1038/s41567-021-01287-z>>.

M, F. C.; FERNANDES, J. M.; SILVA, R. S.; CARNEIRO, M. G. Ocanspectra: an oral cancer detection system from salivary atr-ftir spectroscopy. 2023.

MAKAROV, I.; KISELEV, D.; NIKITINSKY, N.; SUBELJ, L. Survey on graph embeddings and their applications to machine learning problems on graphs. **PeerJ Computer Science**, PeerJ Inc., v. 7, p. e357, 2021. Disponível em: <<https://doi.org/10.7717/peerj-cs.357>>.

MALIK-SONI, N.; SHAKER, A.; LUCK, H.; MULLIN, A. E.; WILEY, R. E.; LEWIS, M. S.; FUENTES, J.; FRAZIER, T. W. Tackling healthcare access barriers for individuals with autism from diagnosis to adulthood. **Pediatric Research**, Nature Publishing Group US New York, v. 91, n. 5, p. 1028–1035, 2022. Disponível em: <<https://doi.org/10.1038/s41390-021-01465-y>>.

MIRJALILI, S. Genetic algorithm. In: **Evolutionary algorithms and neural networks**. Springer, 2019. p. 43–55. Disponível em: <https://doi.org/10.1007/978-3-319-93025-1_4>.

MITCHELL, T. M. **Machine learning**. [S.l.]: McGraw-hill New York, 1997. v. 1.

MORAIS, C. L.; LIMA, K. M.; SINGH, M.; MARTIN, F. L. Tutorial: multivariate classification for vibrational spectroscopy in biological samples. **Nature Protocols**, Nature Publishing Group UK London, v. 15, n. 7, p. 2143–2162, 2020. Disponível em: <<https://doi.org/10.1038/s41596-020-0322-8>>.

NI, J.; TANG, M.; LIU, Y.; GORDON, O.; LUO, C. An improved immune-inspired self-healing approach based on swarm aggregation algorithm for multi-robot system. **Int. J. Robot. Autom.**, v. 34, n. 5, p. 503–517, 2019. Disponível em: <<https://doi.org/10.2316/J.2019.206-0117>>.

- OKAMOTO, K.; CHEN, W.; LI, X.-Y. Ranking of closeness centrality for large-scale social networks. In: SPRINGER. **International workshop on frontiers in algorithmics**. 2008. p. 186–195. Disponível em: <https://doi.org/10.1007/978-3-540-69311-6_21>.
- PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. **The PageRank citation ranking: Bringing order to the web**. [S.l.], 1999.
- PAULSEN, B.; VELASCO, S.; KEDAIGLE, A. J.; PIGONI, M.; QUADRATO, G.; DEO, A. J.; ADICONIS, X.; UZQUIANO, A.; SARTORE, R.; YANG, S. M. et al. Autism genes converge on asynchronous development of shared neuron classes. **Nature**, Nature Publishing Group UK London, v. 602, n. 7896, p. 268–273, 2022. Disponível em: <<https://doi.org/10.1038/s41586-021-04358-6>>.
- POLI, R.; KENNEDY, J.; BLACKWELL, T. Particle swarm optimization. **Swarm intelligence**, Springer, v. 1, n. 1, p. 33–57, 2007. Disponível em: <<https://doi.org/10.1007/s11721-007-0002-0>>.
- PUSTOKHINA, I. V.; PUSTOKHIN, D. A.; NGUYEN, P. T.; ELHOSENY, M.; SHANKAR, K. Multi-objective rain optimization algorithm with welm model for customer churn prediction in telecommunication sector. **Complex & Intelligent Systems**, Springer, p. 1–13, 2021. Disponível em: <<https://doi.org/10.1007/s40747-021-00353-6>>.
- QIAO, Y.; WU, M.; FENG, Y.; ZHOU, Z.; CHEN, L.; CHEN, F. Alterations of oral microbiota distinguish children with autism spectrum disorders from healthy controls. **Scientific reports**, Nature Publishing Group UK London, v. 8, n. 1, p. 1597, 2018. Disponível em: <<https://doi.org/10.1038/s41598-018-19982-y>>.
- RAO, C. R.; RAO, C. R.; STATISTIKER, M.; RAO, C. R.; RAO, C. R. **Linear statistical inference and its applications**. Wiley New York, 1973. v. 2. Disponível em: <<https://doi.org/10.1002/9780470316436>>.
- RESENDE, V. H.; CARNEIRO, M. G. Analysis of complex network measures for multi-label classification. **International Journal on Artificial Intelligence Tools**, World Scientific, v. 30, n. 04, p. 2150023, 2021. Disponível em: <<https://doi.org/10.1142/S0218213021500238>>.
- ROGERS, S. J.; VISMARA, L. A. Evidence-based comprehensive treatments for early autism. **Journal of Clinical Child & Adolescent Psychology**, Taylor & Francis, v. 37, n. 1, p. 8–38, 2008. Disponível em: <<https://doi.org/10.1080/15374410701817808>>.
- SALIH, S. Q.; ALSEWARI, A. A.; AL-KHATEEB, B.; ZOLKIPLI, M. F. Novel multi-swarm approach for balancing exploration and exploitation in particle swarm optimization. In: SPRINGER. **International Conference of Reliable Information and Communication Technology**. 2018. p. 196–206. Disponível em: <https://doi.org/10.1007/978-3-319-99007-1_19>.
- SAMMUT, S.-J.; CRISPIN-ORTUZAR, M.; CHIN, S.-F.; PROVENZANO, E.; BARDWELL, H. A.; MA, W.; COPE, W.; DARIUSH, A.; DAWSON, S.-J.; ABRAHAM, J. E. et al. Multi-omic machine learning predictor of breast cancer therapy response. **Nature**, Nature Publishing Group, v. 601, n. 7894, p. 623–629, 2022. Disponível em: <<https://doi.org/10.1038/s41586-021-04278-5>>.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of research and development**, IBM, v. 3, n. 3, p. 210–229, 1959. Disponível em: <<https://doi.org/10.1147/rd.33.0210>>.

SARKER, I. H. Machine learning: Algorithms, real-world applications and research directions. **SN Computer Science**, Springer, v. 2, n. 3, p. 1–21, 2021. Disponível em: <<https://doi.org/10.1007/s42979-021-00592-x>>.

SHAW, K. A.; MAENNER, M. J.; BAIO, J.; WASHINGTON, A.; CHRISTENSEN, D. L.; WIGGINS, L. D.; PETTYGROVE, S.; ANDREWS, J. G.; WHITE, T.; ROSENBERG, C. R. et al. Early identification of autism spectrum disorder among children aged 4 years—early autism and developmental disabilities monitoring network, six sites, united states, 2016. **MMWR Surveillance Summaries**, Centers for Disease Control and Prevention, v. 69, n. 3, p. 1, 2020. Disponível em: <<https://doi.org/10.15585/mmwr.ss6903a1>>.

SILVA, T. C.; ZHAO, L. Network-based high level data classification. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 23, n. 6, p. 954–970, 2012. Disponível em: <<https://doi.org/10.1109/TNNLS.2012.2195027>>.

SNIJDERS, T. A. The degree variance: an index of graph heterogeneity. **Social networks**, Elsevier, v. 3, n. 3, p. 163–174, 1981. Disponível em: <[https://doi.org/10.1016/0378-8733\(81\)90014-9](https://doi.org/10.1016/0378-8733(81)90014-9)>.

SPORNS, O. Graph theory methods: applications in brain networks. **Dialogues in clinical neuroscience**, Taylor & Francis, 2022.

STAM, C.; TEWARIE, P.; DELLEN, E. V.; STRAATEN, E. V.; HILLEBRAND, A.; MIEGHEM, P. V. The trees and the forest: characterization of complex brain networks with minimum spanning trees. **International Journal of Psychophysiology**, Elsevier, v. 92, n. 3, p. 129–138, 2014. Disponível em: <<https://doi.org/10.1016/j.ijpsycho.2014.04.001>>.

STRUMBERGER, I.; TUBA, E.; BACANIN, N.; ZIVKOVIC, M.; BEKO, M.; TUBA, M. Designing convolutional neural network architecture by the firefly algorithm. In: **IEEE. 2019 International Young Engineers Forum (YEF-ECE)**. 2019. p. 59–65. Disponível em: <<https://doi.org/10.1109/YEF-ECE.2019.8740818>>.

SUN, S.; CAO, Z.; ZHU, H.; ZHAO, J. A survey of optimization methods from a machine learning perspective. **IEEE transactions on cybernetics**, IEEE, v. 50, n. 8, p. 3668–3681, 2019. Disponível em: <<https://doi.org/10.1109/TCYB.2019.2950779>>.

SUTTON, R. S.; BARTO, A. G. **Reinforcement learning: An introduction**. [S.l.]: MIT press, 2018.

TANG, J.; LIU, G.; PAN, Q. A review on representative swarm intelligence algorithms for solving optimization problems: Applications and trends. **IEEE/CAA Journal of Automatica Sinica**, IEEE, v. 8, n. 10, p. 1627–1643, 2021. Disponível em: <<https://doi.org/10.1109/JAS.2021.1004129>>.

VAPNIK, V. **The nature of statistical learning theory**. Springer science & business media, 1999. Disponível em: <<https://doi.org/10.1007/978-1-4757-3264-1>>.

- WANG, Q.; JIANG, H.; JIANG, Y.; YI, S.; LI, L.; XING, C.-C.; HUANG, J. On searching multiple disjoint shortest paths in scale-free networks with hyperbolic geometry. **IEEE Transactions on Network Science and Engineering**, IEEE, 2022. Disponível em: <<https://doi.org/10.1109/TNSE.2022.3169691>>.
- XIAO, Y. **Bio-inspired computing and networking**. CRC Press, 2016. Disponível em: <<https://doi.org/10.1201/b10781>>.
- YADAV, S. L.; SOHAL, A. Comparative study of different selection techniques in genetic algorithm. **International Journal of Engineering, Science and Mathematics**, International Journals of Multidisciplinary Research Academy, v. 6, n. 3, p. 174–180, 2017.
- YANG, X.-S. Nature-inspired optimization algorithms: Challenges and open problems. **Journal of Computational Science**, Elsevier, v. 46, p. 101104, 2020. Disponível em: <<https://doi.org/10.1016/j.jocs.2020.101104>>.
- ZEIDAN, J.; FOMBONNE, E.; SCORAH, J.; IBRAHIM, A.; DURKIN, M. S.; SAXENA, S.; YUSUF, A.; SHIH, A.; ELSABBAGH, M. Global prevalence of autism: a systematic review update. **Autism Research**, Wiley Online Library, v. 15, n. 5, p. 778–790, 2022. Disponível em: <<https://doi.org/10.1002/aur.2696>>.
- ZEMEL, R.; CARREIRA-PERPIÑÁN, M. Proximity graphs for clustering and manifold learning. **Advances in neural information processing systems**, v. 17, 2004.
- ZHANG, J.; LUO, Y. Degree centrality, betweenness centrality, and closeness centrality in social network. In: ATLANTIS PRESS. **2017 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017)**. 2017. p. 300–303. Disponível em: <<https://doi.org/10.2991/msam-17.2017.68>>.