PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

GUSTAVO SAVI FRAINER

UM MODELO DE *DEEP LEARNING* PARA CLASSIFICAÇÃO DE ORGANISMOS VIVOS UTILIZANDO O SEGMENTO 18S rRNA DA SEQUÊNCIA GENÉTICA

Porto Alegre
2025

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

UM MODELO DE *DEEP LEARNING* PARA CLASSIFICAÇÃO DE ORGANISMOS VIVOS UTILIZANDO O SEGMENTO 18S rRNA DA SEQUÊNCIA GENÉTICA

GUSTAVO SAVI FRAINER

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz

Porto Alegre
2025

Ficha Catalográfica

F812m Frainer, Gustavo Savi

Um modelo de deep learning para classificação de organismos vivos utilizando o segmento 18S rRNA da sequência genética / Gustavo Savi Frainer. – 2025.

102 p.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz.

1. Classificação Taxonômica. 2. Metabarcoding. 3. Deep Neural Network. 4. Machine Learning. I. Ruiz, Duncan Dubugras Alcoba. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS com os dados fornecidos pelo(a) autor(a).

Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

GUSTAVO SAVI FRAINER

UM MODELO DE *DEEP LEARNING* PARA CLASSIFICAÇÃO DE ORGANISMOS VIVOS UTILIZANDO O SEGMENTO 18S rRNA DA SEQUÊNCIA GENÉTICA

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado(a) em 26 de Março de 2025.

BANCA EXAMINADORA:

Profª. Drª. Laura Roberta Pinto Utz (PPGEEB/PUCRS)

Prof. Dr. Dalvan Jair Griebler (PPGCC/PUCRS)

Prof. Dr. Duncan Dubugras Alcoba Ruiz (PPGCC/PUCRS - Orientador)

AGRADECIMENTOS

Primeiramente, dedico este trabalho à minha família e amigos. Sem seu apoio e companheirismo não teria sido possível alcançar este resultado.

Agradeço aos membros, colegas e amigos, dos grupos LIS e Pró-Mata. O conhecimento compartilhado e todo o suporte recebido tornaram essa jornada mais valiosa e gratificante.

Por fim, um agradecimento especial ao meu orientador, Prof. Duncan Ruiz, que tornou o processo, quando possível, muito mais tranquilo e assertivo. Com seus conselhos, ajuda e encorajamento foi possível partir de uma hipótese e chegar a esta conclusão.

O presente resultado foi alcançado em cooperação com a HP Brasil Indústria e Comércio de Equipamentos Eletrônicos LTDA. e com recursos provenientes da Lei de Informática (Lei nº 8.248, de 1991).

UM MODELO DE *DEEP LEARNING* PARA CLASSIFICAÇÃO DE ORGANISMOS VIVOS UTILIZANDO O SEGMENTO 18S rRNA DA SEQUÊNCIA GENÉTICA

RESUMO

A taxonomia, no campo da biologia, é a ciência que classifica os seres vivos hierarquicamente de acordo com características em comum. Atualmente, existem diversas técnicas para a classificação taxonômica de organismos por meio sequenciamento genético, comumente usando análise metagenômica. No entanto, essas técnicas tendem a ser computacionalmente custosas e, em cenários em que há mutações ou variações genéticas dentro de um mesmo subgrupo, falhas ou inconclusivas.

Com o recente avanço do campo da inteligência artificial e *machine learning*, novas potenciais formas de classificação estão sendo estudadas e desenvolvidas. Neste trabalho, é apresentada uma solução de classificador taxonômico baseado em um modelo de *deep learning* e, também, é feita uma análise comparativa entre os resultados da solução apresentada e os resultados obtidos com o classificador q2-feature-classifier da plataforma QIIME2. Os resultados obtidos mostram que a solução desenvolvida alcança acurácias maiores, especialmente no nível de Species.

Palavras-Chave: Classificação Taxonômica, Metabarcoding, Rede Neural Profunda, Aprendizado de Máquina.

A DEEP LEARNING MODEL FOR THE CLASSIFICATION OF LIVING ORGANISMS USING THE 18S rRNA SEGMENT OF THE GENETIC SEQUENCE

ABSTRACT

Taxonomy, in the field of biology, is the science that classifies living beings hierarchically according to common characteristics. Currently, there are several techniques for the taxonomic classification of organisms through genetic sequencing, commonly using metagenomic analysis. However, these techniques tend to be computationally expensive and, in scenarios where there are mutations or genetic variations within the same subgroup, they may fail or be inconclusive.

With the recent advances in the field of artificial intelligence and machine learning, new potential forms of classification are being studied and developed. In this work, a taxonomic classifier solution based on a deep learning model is presented and a comparative analysis is also made between the results of the presented solution and the results obtained with the q2-feature-classifier of the QIIME2 platform. The results obtained show that the developed solution achieves greater accuracies, especially at the Species level.

Keywords: Taxonomic Classification, Metabarcoding, Deep Neural Network, Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

CNN – Rede Neural Convolucional (*Convolutional Neural Network*)

CPU – Unidade Central de Processamento (*Central Processing Unit*)

DBN – Rede de Crenças Profundas (*Deep Belief Network*)

DL – Aprendizado Profundo (*Deep Learning*)

DNA – Ácido Desoxirribonucleico (*Deoxyribonucleic Acid*)

eDNA – Ácido Desoxirribonucleico Ambiental (*Environmental Deoxyribonucleic Acid*)

GPU – Unidade de Processamento Gráfico (*Graphics Processing Unit*)

IA – Inteligência Artificial

IUPAC – União Internacional de Química Pura e Aplicada (*International Union of Pure and Applied Chemistry*)

MOTU – Unidade Taxonômica Operacional Molecular (*Molecular Operational Taxonomic Unit)*

OTU – Unidade Taxonômica Operacional (*Operational Taxonomic Unit*)

PCR – Reação em Cadeia da Polimerase (*Polymerase Chain Reaction*)

RAI – Índice de Abundância Relativa (*Relative Abundance Index*)

RAM – Memória de Acesso Aleatório (*Random-Access Memory*)

ReLU – Unidade Linear Retificada (*Rectified Linear Unit*)

RNA – Ácido Ribonucleico *(Ribonucleic Acid*)

rRNA – Ácido Ribonucleico Ribossómico (*Ribosomal Ribonucleic Acid*)

VRAM – Memória de Acesso Aleatório de Vídeo (*Video Random-Access Memory*)

SUMÁRIO

1. INTRODUÇÃO

No campo da biologia, taxonomia é a ciência de nomear, descrever e classificar organismos de acordo com características em comum. Existem diversas classificações taxonômicas, com diferentes critérios e estruturas, das quais a mais conhecida atualmente teve origem no trabalho Systema Naturae [17] publicado em 1735, do botânico sueco Carl Linnæus, e foi sendo aprimorada resultando no que é conhecido como Taxonomia Moderna.

Atualmente, as principais técnicas de classificação são baseadas na análise de sequencias genéticas por meio de algoritmos, permitindo a aplicação em contextos mais complexos e com maior volume de dados. Além disso, novos métodos de sequenciamento genético que foram desenvolvidos ao longo dos últimos anos tornaram as pesquisas na área mais acessíveis e escaláveis como, por exemplo, biomonitoramento baseado em eDNA [8].

Environmental DNA, ou eDNA, refere-se ao material genético encontrado em um ambiente [30]. O conjunto de amostras de um local representam as espécies presentes no mesmo, servindo para identificar o perfil do ecossistema. O estudo de eDNA tem se mostrado com grande aplicabilidade em diversos cenários, tais como na conservação de ecossistemas, identificação e mapeamento da biodiversidade. [3][4][23]

Por vezes, estudos que analisam amostras de conjuntos de organismos necessitam de meios de sequenciamento genético mais eficientes e que possam ser aplicados diretamente na amostra, visto que a separação dos organismos em geral não é viável. Nesses casos, é utilizado o *Metabarcoding*, normalmente nos segmentos genéticos 16S rRNA e 18S rRNA, para identificar as espécies presentes nas amostras.

A classificação taxonômica e identificação de espécies por algoritmos são feitas a partir da comparação da sequência genética da amostra com as sequencias de referências, sendo que os critérios de comparação variam de acordo com o algoritmo escolhido. Atualmente, o algoritmo mais utilizado é o Naive Bayes Classifier[33], um algoritmo de *machine learning* cujo método busca avaliar a semelhança entre sequencias pela observação de fragmentos em comum.

Considerando modelos recentes de *deep learning* e o potencial desses em abstrair e aprender padrões, neste trabalho foram realizados experimentos aplicando essas técnicas na classificação taxonômica de organismos vivos. Como solução proposta, este trabalho apresenta um modelo CNN para classificação de organismos vivos com base no segmento 18S rRNA. Além disso, são apresentados os resultados dos testes e uma análise comparativa com um classificador baseado em Naive Bayes: o *plug-in* q2-feature-classifier da plataforma QIIME2.

1.1 Contribuição

Como principal contribuição, este trabalho provê uma solução de inteligência artificial para classificação de organismos vivos com base em sequências genéticas. Essa solução é composta por um modelo de rede neural profunda que pode ser treinado com sequências genéticas e adaptado para diferentes cenários, de acordo com os dados utilizados. Além disso, é compartilhada uma análise do modelo demonstrando a sua aplicabilidade e potenciais aprimoramentos.

1.2 Organização da Dissertação

O conteúdo deste trabalho está estruturado nos capítulos subsequentes da seguinte forma:

Cap. 2. Conceitos básicos, onde são introduzidos os principais conceitos e técnicas utilizadas neste trabalho,

Cap. 3. Classificação taxonômica, o qual descreve como é realizada atualmente e as principais limitações,

Cap. 4. Trabalhos relacionados, no qual é abordada a literatura, apresentando os pontos centrais dos trabalhos relacionados que foram encontrados durante a realização da revisão sistemática da literatura,

Cap 5. Solução desenvolvida, o qual apresenta a solução desenvolvida, descrevendo a respectiva estrutura e funcionamento,

Cap. 6. Experimentação, onde são apresentados os experimentos feitos, bem como a metodologia, dados e recursos utilizados, além das análises realizadas; e,

Cap. 7. Conclusões, no qual são apresentadas as principais conclusões decorrentes do desenvolvimento deste trabalho, além de citados potenciais trabalhos futuros e melhorias.

Além disso, informações complementares são apresentadas na seção de apêndices. Os apêndices de A a E são referentes aos resultados dos experimentos, enquanto o apêndice F apresenta a revisão da literatura realizada sobre a utilização *deep learning* para a classificação de organismos vivos, visando identificar quais as técnicas atuais, resultados obtidos e a viabilidade.

2. CONCEITOS BÁSICOS

2.1 Taxonomia Moderna

Linnæus, o qual ficou conhecido como pai da taxonomia moderna, descreveu em seu trabalho o sistema hierárquico que havia desenvolvido, cuja composição possuía três grupos primários denominados reinos: *Regnum Animale* (reino animal), *Regnum Vegetabile* (reino vegetal) e *Regnum Lapideum* (reino mineral). Durante seus anos de pesquisa, Linnæus foi aprimorando o sistema proposto de diversas formas, tendo publicado ao todo 12 edições do Systema Naturae sob seu nome.

O sistema hierárquico de Linnæus, apesar de ter sido muito importante e já abordar níveis como Classes, Ordens, Gêneros e Espécies, ainda continha diversas inconsistências e lacunas. Um dos exemplos de problema recorrente tinha-se quando diferentes indivíduos eram agrupados em um mesmo *taxon* (como são chamados os grupos individuais; plural *taxa*) por não ter uma divisão adequada para os indivíduos em questão, criando uma falsa associação entre os mesmos. Com o passar dos anos, outros pesquisadores usaram o sistema para criar variações, com melhor entendimento dos organismos, novos conhecimentos e critérios mais precisos e consistentes. Dentre os diversos estudos que contribuíram para o desenvolvimento de versões aprimoradas do sistema hierárquico, um dos que teve maior impacto na forma que a hierarquia é estruturada e nos critérios para classificação foi o "On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life"[7], de Charles Darwin.

O trabalho de Darwin teve grande impacto na classificação taxonômica pois foi um dos primeiros estudos a abordar a relação entre espécies de acordo com um ancestral em comum. Tal mudança permitiu agrupar organismos semelhantes de forma mais consistente e precisa, uma vez que organismos que possuem mesma origem tendem a compartilhar características. Além disso, ainda que não tenha sido abordado o código genético em si na época, o fato de indivíduos de um mesmo grupo possuírem genomas semelhantes possibilitou que novas técnicas fossem desenvolvidas para a utilização de sequencias genéticas na identificação e classificação de organismos baseados na sua semelhança com sequências de referência para cada *taxon*.

Atualmente, os principais níveis taxonômicos utilizados são, em ordem hierárquica: Domínio, Reino, Filo, Classe, Ordem, Família, Gênero e Espécie. Dentre esses, o Domínio o mais recente a ser adotado, tendo sua inclusão na estrutura atribuída a Royall T. Moore[19]. Além dos principais níveis, subníveis também são frequentemente utilizados, variam de acordo com cada ramificação onde se encontram. No entanto, devido às novas descobertas sobre os organismos e de novas variáveis que impactam nos critérios de classificação, nem sempre as estruturas hierárquicas tradicionais mostram-se adequadas

para a classificação de um grupo de organismos. De forma semelhante ao problema de organismos distintos em um mesmo grupo que ocorria no sistema proposto por Linnæus, a gama de características associadas a cada organismo varia tanto que torna-se difícil definir critérios de classificação que funcionem em todos os casos.

2.2 Taxonomia Baseada em OTUs

Uma OTU, do inglês *Operational Taxonomic Unit*, é a unidade básica presente em um sistema de taxonomia numérica. A abordagem da taxonomia numérica é baseada no agrupamento de organismos semelhantes entre si, sendo cada grupo uma OTU diferente.

As técnicas de taxonomia por meio de OTUs são amplamente utilizadas, tendo recebido destaque quando aplicadas no agrupamento de organismos com base na similaridade dos códigos genéticos sequenciados, cenário em que recebe o nome de Molecular Operational Taxonomic Unit (MOTU)[9]. Além disso, um dos benefícios dessa técnica é a possibilidade de agrupar sem definir características em comum para cada grupo, apenas usando a similaridade entre os indivíduos, tornando-a uma alternativa interessante em casos onde as estruturas hierárquicas convencionais não conseguem representar adequadamente grupos distintos.

2.3 Sequenciamento Genético

Todo organismo vivo possui um genoma, um conjunto de ácidos nucleicos, que contém informações sobre as características desse organismo e seu funcionamento. O sequenciamento genético, por sua vez, é o processo de identificar a sequência de bases nitrogenadas que compõem uma cadeia de nucleotídeos de um ácido nucleico.

O sequenciamento do genoma completo de um organismo é um processo complexo e difícil, sendo normalmente feito a partir do agrupamento do sequenciamento de subsegmentos menores do genoma (figura 2.1). Além disso, dependendo da finalidade, pode não ser necessário o sequenciamento do código genético por completo. Nesse caso, são sequenciadas apenas regiões específicas do material.

Existem diversas regiões de códigos genéticos que são amplamente estudadas de forma isolada com objetivos diversos. Algumas delas estão presentes em organismos de diferentes grupos taxonômicos, podendo ter variações em suas sequências de bases e sendo usadas para comparar e analisar semelhança entre organismos.

As regiões são identificadas de acordo com a sua posição na sequência genética e, comumente, possuem um trecho com baixa variabilidade que é usado como identifica-

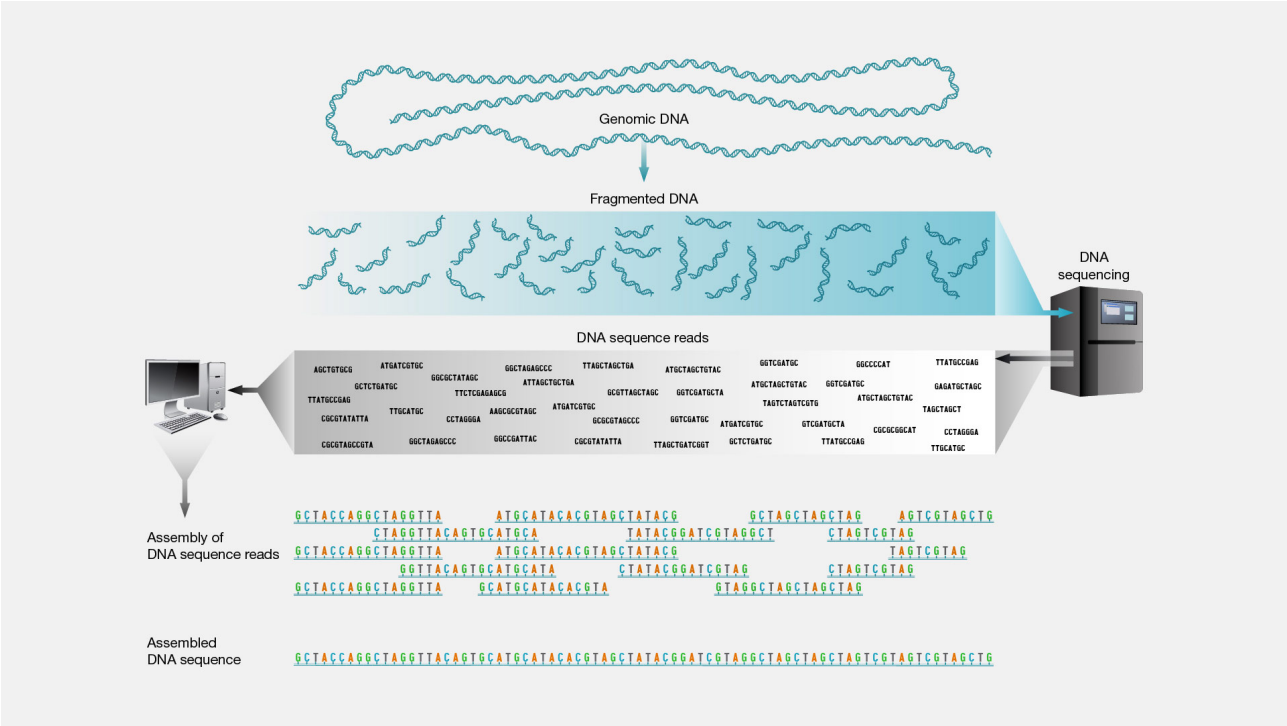

Figura 2.1: Sequenciamento genético usando a técnica *Shotgun*.

Fonte: National Human Genome Research Institute [21]

dor. A existência desse trecho identificador é especialmente importante para métodos de sequenciamento genético que utilizam um primer sintético.

2.3.1 Primer

Um primer é um segmento curto de ácido nucleico. Primers também são conhecidos como iniciadores por serem os responsáveis pela iniciação do processo de síntese do material genético.

Alguns métodos envolvidos em processos de sequenciamento utilizam um primer sintético para iniciar uma reação. Um exemplo —ilustrado na figura 2.2— é o método de reação em cadeia da polimerase (em inglês *polymerase chain reaction* - PCR), em que o primer se conecta ao identificador presente no material genético e inicia uma reação em cadeia, onde as bases subsequentes também são conectadas, permitindo a replicação da sequência.

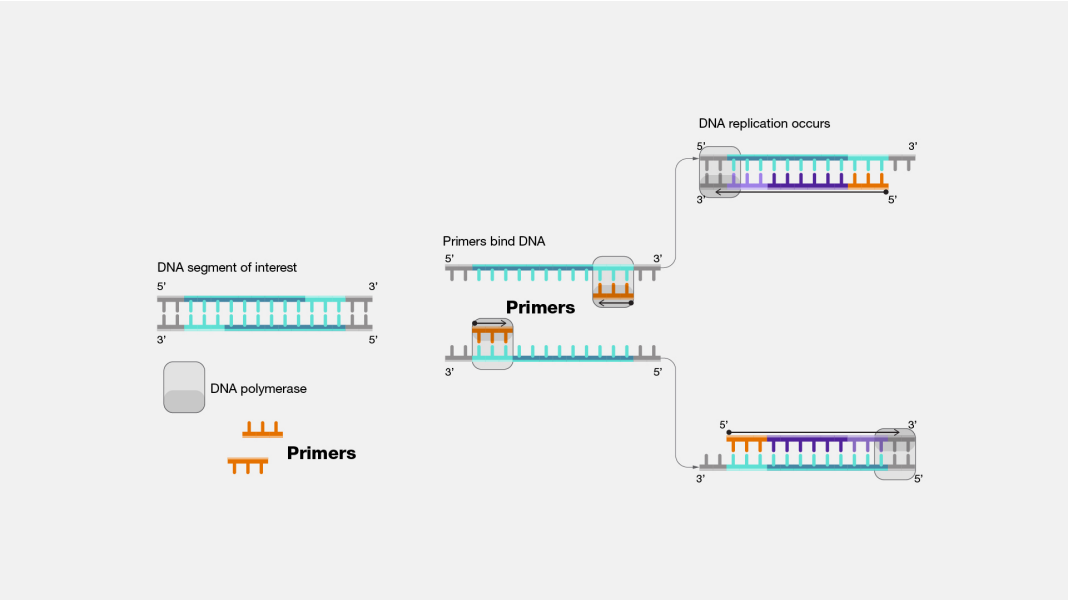

Figura 2.2: Replicação de sequência genética com primer.

Fonte: National Human Genome Research Institute [20]

2.3.2 Notação IUPAC para bases

A International Union of Pure and Applied Chemistry (IUPAC) propôs, inicialmente em 1970, a criação de uma notação das bases [13], a fim de facilitar a representação das sequências genéticas. Esse padrão recebeu diversas adaptações e refinamentos, como o apresentado no artigo "Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984"[6]. Além da representação básica das quatro bases principais (A, T, G e C), a notação sugerida ainda contempla a incerteza entre duas ou mais bases (tabela 2.1).

2.4 Barcoding

Barcoding [11] é um método de classificação a nível de espécie de organismos a partir de pequenos fragmentos específicos do código genético. O processo de *barcoding* é fundamentado em quatro etapas: extração do DNA, amplificação, sequenciamento e análise.

A extração do DNA refere-se a coleta da amostra de material genético que será usado para a classificação. Após extraída a amostra, a mesma é submetida a um processo de amplificação, onde réplicas dos fragmentos desejados são feitos usando técnicas como o PCR. Os conjuntos de fragmentos genéticos gerados durante a amplificação são então sequenciados e analisados, sendo comparados com uma base de *barcodes* previamente

Tabela 2.1: Resumo das recomendações da IUPAC de nomenclatura de ácidos nucleicos com códigos de uma letra.

Símbolo	**Significado**	**Origem da Nomeação**
G	G	Guanina (**G**uanine)
A	A	Adenina (**A**denine)
T	T	Timina (**T**hymine)
C	C	Citosina (**C**ytosine)
R	G ou A	Purina (pu**R**ine)
Y	T ou C	Pirimidina (p**Y**rimidine)
M	A ou C	Amina (a**M**ino)
K	G ou T	Cetona (**K**etone)
S	G ou C	Interação forte (**S**trong interaction)
W	A ou T	Interação fraca (**W**eak interaction)
H	A ou C ou T	não é G, **H** sucede G no alfabeto
B	G ou T ou C	não é A, **B** sucede A no alfabeto
V	G ou C ou A	não é T (não é U), **V** sucede U no alfabeto
D	G ou A ou T	não é C, **D** sucede C no alfabeto
N	G ou A ou T ou C	qualquer base (a**N**y)

Fonte: [6]

classificados. Essa comparação permite fazer a classificação da amostra de acordo com a similaridade da mesma e das amostras de referência.

2.5 Metabarcoding

O processo de *metabarcoding* é semelhante ao *barcoding*, tendo como principal diferença que o primeiro é aplicado em uma amostra composta por múltiplos organismos, enquanto o segundo é aplicado em um único indivíduo. Ambos os processos seguem as mesmas quatro etapas fundamentais: extração do DNA, amplificação, sequenciamento e análise. [23] [18]

Estudos que precisam analisar amostras de conjuntos de organismos, sem analisar os mesmos individualmente, necessitam de meios de sequenciamento genético mais eficientes e que possam ser aplicados diretamente na amostra, visto que a separação dos organismos em geral não é viável. Nesses casos, o *metabarcoding*, aplicado em segmentos genéticos como o 16S rRNA e o 18S rRNA, é uma alternativa interessante para identificar as espécies presentes nas amostras.

2.6 Plug-in q2-feature-classifier

O q2-feature-classifier[25] é um *plug-in* da plataforma *open source* de bioinformática e ciência de dados chamada QIIME2. Esse *plug-in* permite executar tarefas de classificação taxonômica a partir de conjuntos de sequências genéticas. Além disso, para facilitar o processo, a ferramenta já possui diversos algoritmos classificadores pré-definidos baseados em diferentes técnicas, incluindo técnicas atuais como o Naïve Bayes.

2.7 Machine Learning

Machine learning, ou aprendizado de máquina em português, é um subcampo da inteligência artificial, com forte base em métodos e modelos estatísticos. O objetivo principal é o estudo e desenvolvimento de algoritmos para identificar comportamentos por meio do reconhecimento de padrões e a generalização desses comportamentos, inclusive para cenários não previamente conhecidos.

Existem duas etapas principais na utilização de modelos de aprendizado de máquina: treino e predição. Na etapa de treino, o modelo realiza o processo de aprendizado, no qual o mesmo identifica os padrões com base nos dados recebidos. Já na etapa de predição, o modelo prediz um valor com base em dados de entrada e nos padrões aprendidos na etapa de treino e a qualidade da predição pode ser aferida.

2.7.1 Arquitetura de Rede Neural

A arquitetura de uma rede neural é a definição da estrutura da rede neural, a qual descreve as camadas que a constituem, como a camada de entrada, as camadas ocultas e a camada de saída. Essa definição abrange a especificação dos componentes que fazem parte da rede, a organização interna desses elementos e as interconexões entre eles, determinando, assim, o fluxo e a interação das informações dentro do sistema.

2.7.2 Deep Learning

Deep learning (DL), ou aprendizado profundo, é um subconjunto da área de aprendizado de máquina em que os modelos são compostos por múltiplas camadas de processamento [16]. Essas camadas são agrupadas de forma sequencial, com as camadas internas sendo chamadas de *hidden layers*, ou camadas ocultas (figura 2.3).

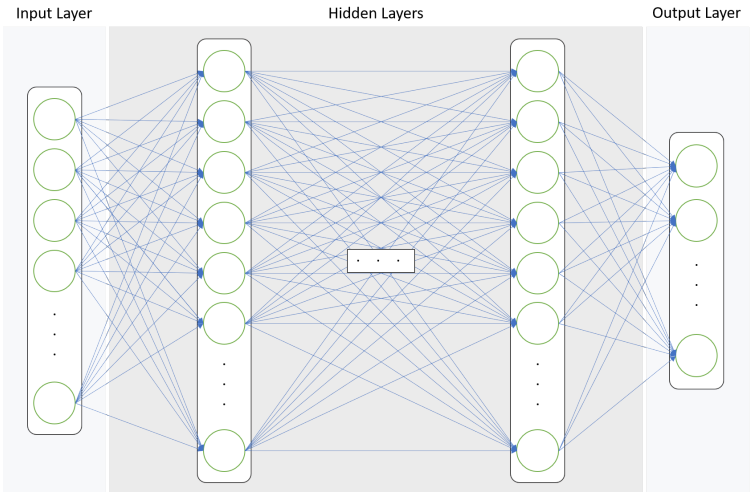

Figura 2.3: Camadas de um modelo de *Deep Learning*.

Cada uma das camadas do modelo aplica transformações nos dados recebidos da camada anterior, gerando abstrações e representações novas a cada nível. Com isso, as diversas camadas do modelo, juntas, formam a composição de representações das informações em vários níveis de abstração, permitindo que funções muito mais complexas sejam aprendidas. As camadas também podem ter funções específicas, como, por exemplo, amplificar características ou suprimir ruídos dos dados.

2.7.2.1 Rede Neural Convolucional

Uma Rede Neural Convolucional, ou CNN, é um tipo de modelo de aprendizado profundo projetado principalmente para processar dados estruturados em vetores [27]. A CNN processa dados aplicando filtros convolucionais para extrair características, camadas de *pooling* para reduzir a dimensionalidade e camadas totalmente conectadas para fazer previsões finais.

A implementação dessas etapas de processamento permite que a rede neural consiga identificar e generalizar padrões complexos presentes nos dados. Essa capacidade de identificar os padrões, ainda que com variações, como de posição, faz com que esse tipo de modelo seja amplamente aplicado para identificação de objetos, segmentação e classificação de imagens.

3. CLASSIFICAÇÃO TAXONÔMICA

Tradicionalmente, as técnicas de classificação são baseadas em características físicas observadas nos organismos. Esse tipo de técnica possui algumas vantagens, como a observação de novas características, mas tende a ser um processo demorado, sendo aplicado a cada organismo individualmente. Por isso, meios tradicionais de classificação podem ser inviáveis em casos como, por exemplo, quando necessita-se realizar a classificação de forma rápida de um grupo de organismos que, muitas vezes, não podem ser isolados para estudo.

Com isso, métodos de classificação baseados nas sequências genéticas foram desenvolvidas. Essas técnicas são aplicadas por meio da comparação entre a sequência genética do organismo a ser classificado e um conjunto de sequências genéticas de organismos previamente classificados por técnicas tradicionais. Apesar de requerer que seja feito o sequenciamento genético dos organismos, tanto os que precisam ser classificados quanto dos previamente classificados que são usados para comparação, esses novos meios permitem a automatização de diversas etapas.

3.1 Atuais Técnicas de Classificação

Atualmente, existem várias formas de realizar a classificação taxonômica usando sequências genéticas. Muitas dessas técnicas receberam adaptações e geraram novas formas, mas duas delas se destacaram, seja pelos resultados ou por fundamentar o desenvolvimento de outras. Essas técnicas são a de comparação de sequências alinhadas e a baseada em Naïve Bayes.

3.1.1 Alinhamento de Sequências Genéticas

Uma das formas mais básicas de realizar a classificação é pela comparação entre sequências genéticas alinhadas, onde é feito o alinhamento das amostras de referência com a amostra em estudo e as bases nas posições respectivas são comparadas. Após isso, é calculada a similaridade entre a amostra em estudo e cada referência, para então atribuir a mesma classificação da amostra de referência com maior similaridade ao organismo da amostra em estudo.

3.1.2 Classificação com Naïve Bayes

A técnica de classificação Naïve Bayes assume que as variáveis dos dados a serem classificados são independentes, tendo, assim, a probabilidade do dado ser de uma determinada classe definida de acordo com as probabilidades de cada variável calculadas separadamente. Essa técnica é chamada assim pois aplica a técnica de classificação bayesiana assumindo de uma forma ingênua, do inglês *Naive*, que as variáveis não possuem relação entre si.

Na classificação taxonômica por meio da técnica de Naïve Bayes, a sequência genética é quebrada em várias subsequências que são consideradas as variáveis dos dados. Para cada subsequência é calculada a probabilidade dela pertencer a cada uma das classes. Após todas as probabilidades das subsequências terem sido calculadas, elas, então, são utilizadas para calcular a probabilidade do dado como um todo pertencer a cada classe taxonômica.[32]

3.2 Limitações

Apesar dos avanços no desenvolvimento e aprimoramento de técnicas de classificação, ainda existem limitações significativas. Uma das principais limitações das técnicas baseadas na comparação direta entre sequências alinhadas é o custo computacional. Dado que a comparação é feita entre a sequência em estudo com cada uma das sequências de referência individualmente e que a qualidade dos resultados depende do uso de um conjunto de referência representativo e amplamente abrangente, esse tipo de aplicação torna-se em muitos casos inviável por ser mais lenta ou por possuir um custo computacional proibitivo.

Já técnicas que utilizam Naive Bayes, ainda que mais rápidas em relação as comparações diretas de sequências, também possui limitações relevantes. Esse algoritmo é baseado na presença e ausência de segmentos menores, não considerando as respectivas posições e, consequentemente, nem o impacto de partes específicas da sequência no processo.

É sabido que alguns trechos de sequências genéticas possuem maior variabilidade que outros, fato que pode ser abordado para aprimorar a atribuição dos pesos para cada segmento de acordo com o respectivo impacto na classificação. Sendo assim, por exemplo, trechos com menor variabilidade poderiam ter maior impacto na classificação em níveis hierárquicos mais abrangentes, enquanto partes com maior variabilidade teriam maior relevância em níveis mais específicos da taxonomia.

4. TRABALHOS RELACIONADOS

Existem na literatura trabalhos que abordam a utilização de *deep learning* na classificação taxonômica, como os listados na tabela 4.1. Esses trabalhos foram estudados durante a realização da revisão sistemática da literatura, a qual está presente no apêndice F.

Tabela 4.1: Publicações selecionadas.

Título	**Data**
AmpliconNet: Sequence Based Multi-layer Perceptron for Amplicon Read Classification Using Real-time Data Augmentation [15]	2018
Beating Naive Bayes at Taxonomic Classification of 16S rRNA Gene Sequences. [33]	2021
BERT contextual embeddings for taxonomic classification of bacterial DNA sequences [12]	2022
Convolutional neural networks improve fungal classification [31]	2020
Investigation of machine learning algorithms for taxonomic classification of marine metagenomes [22]	2023
Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. [5]	2018
Taxonomic classification of metagenomic sequences from Relative Abundance Index profiles using deep learning [14]	2021
Taxonomy metagenomic analysis for microbial sequences in three domains system via machine learning approaches [1]	2018

Nesses trabalhos, foi identificado um maior interesse em abordar a classificação de organismos por meio do segmento 16S rRNA, devido algumas características como, por exemplo, os tipos de organismos que podem ser classificados. Apesar dessa predileção, nenhum dos estudos apresentou limitações quanto a isso, sendo as estruturas dos modelos agnósticas em relação a quais segmentos foram utilizados.

Por outro lado, a pouca quantidade de dados de referência disponíveis é um problema comum entre os estudos. Considerando a necessidade de um grande volume de dados que o treinamento de modelos de *deep learning* demanda, os conjuntos de dados públicos são muitas vezes pequenos demais.

Além disso, a qualidade dos dados também possui grande impacto nos estudos, como visto em "Investigation of machine learning algorithms for taxo- nomic classification of marine metagenomes"[22]. Ainda que os dados sejam curados, sendo removidos ruídos e inconsistências, os conjuntos disponíveis são desbalanceados, com poucas classes sendo responsáveis pela maior parte dos registros. A utilização de conjuntos des-

balanceados tende a ocasionar problemas no treinamento de modelos, como *overfitting*, demandando melhor pré-processamento e arquiteturas mais tolerantes a distribuição heterogênea das classes.

Para a poder utilizar as sequências em modelos de inteligência artificial, essas precisam estar estruturadas em um formato de dados compatível. Os estudos em sua maioria utilizavam técnicas como *bag of words* e K-mer para a representação das sequências, não tendo nenhum trabalho que abordasse a representação das sequências pelo mapeamento individual das bases, como *one-hot encoding*[28], por exemplo.

No que se refere aos tipos de arquiteturas dos modelos, 6 dos 8 trabalhos estudados utilizaram arquiteturas com camadas convolucionais, alcançando acurácia superior a 90% em diversos testes. No entanto, muitos desses casos dependiam também de informações externas, como por exemplo a utilização de Relative Abundance Index, no trabalho "Taxonomic classification of metagenomic sequences from Relative Abundance Index profiles using deep learning"[14], para considerar a frequência de ocorrência de cada classe no ambiente amostrado.

A importância da escolha dos parâmetros também foi abordada em alguns estudos, como em "Optimizing taxonomic classification of markergene amplicon sequences with QIIME 2's q2-feature-classifier plugin"[5]. Porém, os estudos não apresentam aprofundamento sobre como foram definidas as arquiteturas dos modelos e escolha dos parâmetros, nem avaliaram quantitativamente as melhorias obtidas. De fora geral, as arquiteturas apresentadas ou eram muito simples e genéricas, ou eram importadas de outros contextos, como a ResNET em "AmpliconNet: Sequence Based Multi-layer Perceptron for Amplicon Read Classification Using Real-time Data Augmentation "[15].

Outro problema recorrente nos estudos é quanto a necessidade de recursos computacionais em aplicações de *deep learning*, a qual muitas vezes torna impraticável o uso dessa tecnologia em cenários reais. A escolha de arquiteturas adequadas, com tamanho e com camadas adequadas, tem grande impacto na demanda de capacidade computacional, sendo fator crucial para viabilizar a utilização desse tipo de solução.

5. SOLUÇÃO DESENVOLVIDA

Este trabalho utilizou técnicas de *deep learning* para desenvolver um modelo de inteligência artificial para a classificação de organismos vivos. A arquitetura definida para o modelo foi baseada em CNN, devido ao processo de *feature extraction* que ocorre nas camadas convolucionais. Nesse processo as sequências são percorridas gerando os *features maps*, os quais são utilizados para relacionar características e padrões das sequências com classes de organismos vivos.

Buscando definir uma arquitetura otimizada para o contexto de classificação taxonômica, foram adicionadas conexões residuais junto as camadas residuais. Com isso, tanto padrões mais simples quanto mais complexos presentes nas sequências genéticas podem ser aprendidos pelo modelo.

A solução proposta trabalha com sequências com 900 bases de comprimento. A escolha desse tamanho deve-se ao equilíbrio entre a quantidade de dados disponíveis com esse comprimento e a capacidade de relacionar as sequências com as respectivas classes. Além disso, para a utilização das sequências como entradas do modelo, elas precisam ser pré-processadas, gerando as respectivas representações vetoriais de dimensões (900, 4).

5.1 Representação Vetorial das Sequências

As sequências presentes no conjunto de dados são do tipo textual, sendo uma cadeia de caracteres em que cada um representa uma base da sequência. Essa representação segue o padrão de notação IUPAC, descrito na seção 2.3.2.

Visto que os modelos de IA não são capazes de trabalhar com dados textuais diretamente, foram criadas funções de codificação das sequências. Como resultado das funções, são geradas representações vetoriais numéricas, compatíveis com os tipos de dados esperados por modelos de IA.

A representação vetorial de cada sequência é composta por um vetor principal com 900 posições, representando as 900 bases da sequência. Cada uma das posições é composta por um vetor com 4 campos, que representam, respectivamente, a probabilidade da base ser do tipo A, T, G ou C.

5.2 Arquitetura do Modelo

O modelo é composto, conforme o mostrado na figura 5.1, por oito blocos de camadas convolucionais com conexões residuais, sendo cada um deles seguido por uma

Tabela 5.1: Codificação vetorial das bases IUPAC.

Base IUPAC	**Probabilidade**			
	A	**T**	**G**	**C**
A	1,0	0,0	0,0	0,0
T	0,0	1,0	0,0	0,0
G	0,0	0,0	1,0	0,0
C	0,0	0,0	0,0	1,0
W	0,5	0,5	0,0	0,0
S	0,0	0,0	0,5	0,5
M	0,5	0,0	0,0	0,5
K	0,0	0,5	0,5	0,0
R	0,5	0,0	0,5	0,0
Y	0,0	0,5	0,0	0,5
B	0,0	0,3	0,3	0,3
D	0,3	0,3	0,3	0,0
H	0,3	0,3	0,0	0,3
V	0,3	0,0	0,3	0,3
N	0,25	0,25	0,25	0,25

camada de *pooling*, e por duas camadas totalmente conectadas. Ao todo, são 52.470.645 parâmetros treináveis que compõem o modelo.

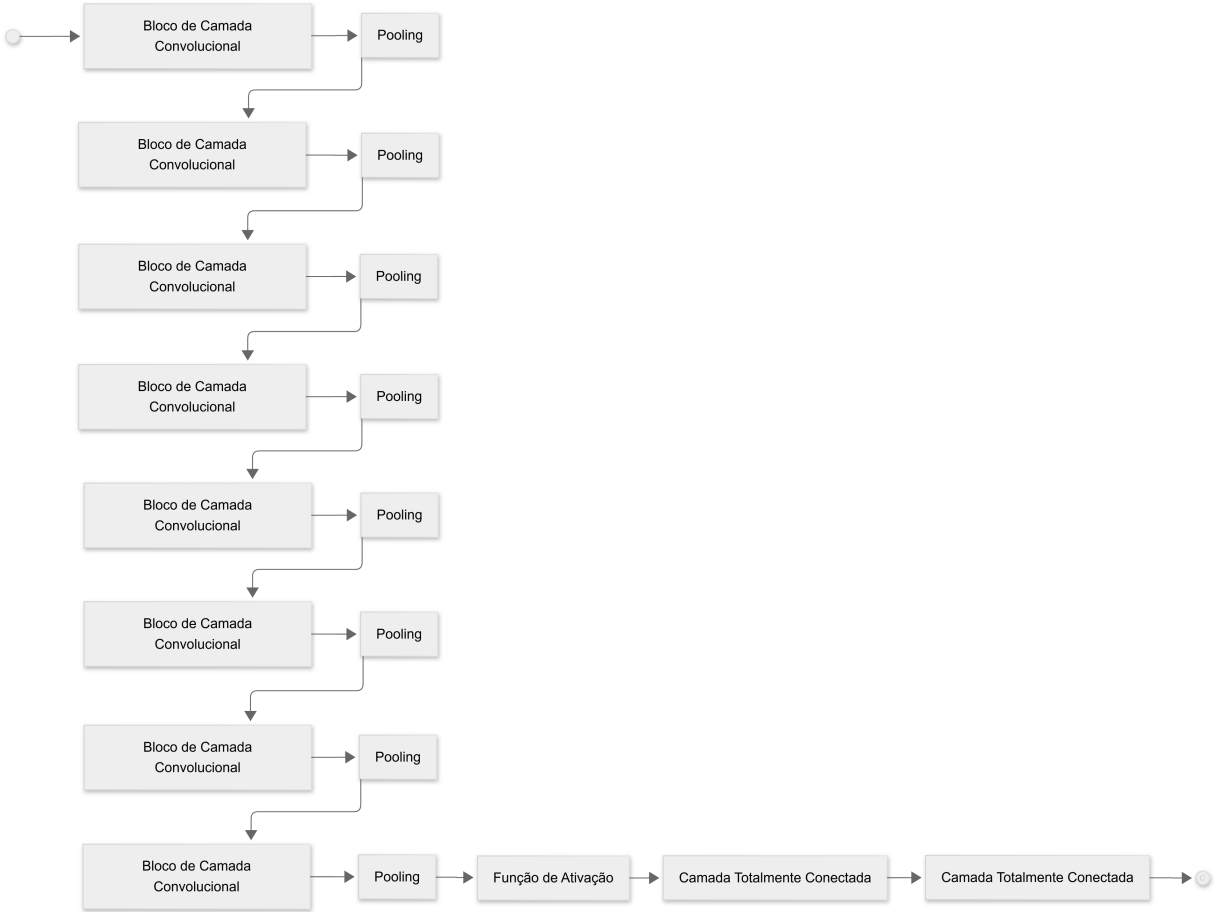

Figura 5.1: Diagrama da arquitetura do modelo.

5.2.1 Blocos de Camadas Convolucionais

Os blocos de camadas convolucionais, como ilustrado na figura 5.2, são compostos por quatro partes principais: camada de convolução, normalização, conexão residual e ativação. Os blocos variam entre si apenas no número de canais e tamanho das sequências.

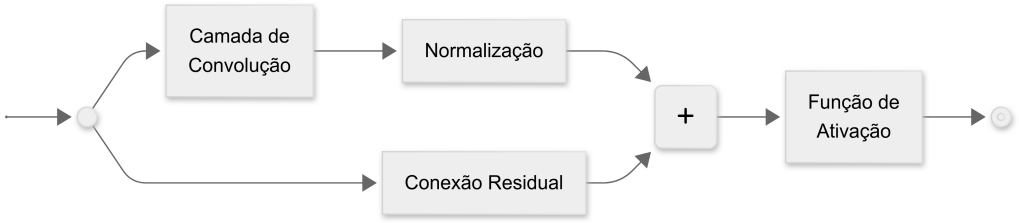

Figura 5.2: Diagrama do bloco de camada convolucional.

As operações convolucionais são do tipo unidimensional, aplicadas em múltiplos canais e com *kernels* de tamanho 4. Para a aplicação adequada das conexões residuais, a técnica de normalização de *batch* foi aplicada nas saídas das camadas convolucionais.

Após a camada convolucional e a normalização, é feita a conexão residual, descrita na seção 5.2.1.1. Por fim, uma função de ativação não-linear é aplicada nos vetores resultantes das etapas anteriores. Em todos os blocos a função de ativação aplicada é a *Rectified Linear Unit* (ou ReLU).

5.2.1.1 Conexões Residuais

A utilização de múltiplas camadas ocultas em um modelo de *deep learning* visa permitir a generalização e identificação de padrões mais complexos nos dados. No entanto, a cada transformação aplicada nos dados, é normal que um pouco das informações das sequências originais e de padrões mais simples sejam perdidos. A fim de evitar que informações relevantes sejam perdidas durante o processo, conexões residuais foram adicionadas aos blocos.

As conexões residuais funcionam concatenando cópias adaptadas dos valores de entrada junto aos vetores resultantes das camadas convolucionais. Como as dimensões dos dados são diferentes, é aplicada uma operação de convolução unidimensional com *kernel* de tamanho 1 na entrada antes da concatenação. Com isso, informações das sequências originais ou de blocos anteriores tendem a ser propagadas pela rede neural, permitindo que tanto padrões complexos sejam aprendidos, quanto padrões mais simples.

5.2.2 Camadas de Pooling

Após cada bloco com camada convolucional, é aplicada uma camada de *pooling* para reduzir a dimensionalidade dos dados. No modelo desenvolvido, a função de *pooling* utilizada é a *adaptive average pooling*, a qual aplica uma função adaptável de média.

5.2.3 Camadas Totalmente Conectadas

As últimas camadas que compõem o modelo são duas do tipo totalmente conectadas. Essas camadas são responsáveis pela classificação, relacionando as *features* extraídas nas camadas convolucionais com as classes taxonômicas.

6. EXPERIMENTAÇÃO

Para avaliar a solução proposta e validar a sua potencial aplicabilidade, foram realizados experimentos utilizando um conjunto de dados de sequências 18S rRNA. Os experimentos tiveram como cerne demonstrar a capacidade da solução proposta em classificar sequências reais de organismos. Também objetivaram comparar os resultados da solução desenvolvida com resultados obtidos por meio de técnicas amplamente utilizadas na bioinformática, representadas pela solução implementada com Naive Bayes na ferramenta q2-feature-classifier.

As classificações foram realizadas em diferentes níveis hierárquicos separadamente, sendo eles: Class, Order, Family, Genus e Species. Apesar da ferramenta q2-feature-classifier realizar a classificação em múltiplos níveis simultaneamente, cada etapa de testes considerou apenas o respectivo nível. Já nos testes com a solução desenvolvida, apenas as classes do nível em questão eram utilizadas. Dado que não há ambiguidade entre as classes, sendo cada classe pertencente a apenas um único nível e a uma única ramificação hierárquica, os níveis superiores aos classificados podem ser deduzidos de forma determinística.

6.1 Metodologia

O protocolo de experimentação aplicado foi composto por quatro fases principais. São elas: pré-processamento dos dados, experimentos de referência, experimentos da solução desenvolvida e, por fim, a comparação e análise dos resultados.

Os recursos computacionais disponíveis em todos os experimentos foram exatamente os mesmo e são descritos na seção 6.2. Apesar das implementações utilizarem os recursos de forma distinta, inclusive o *plug-in* q2-feature-classifier não utilizando a placa-gráfica, o compartilhamento, quando adequados para ambos os casos, reduz o risco de discrepâncias nos resultados por fatores externos.

Ainda com o intuito garantir a qualidade da comparação entre os experimentos com a implementação proposta e com a de referência, foram utilizados os mesmos conjuntos de dados pré-processados —descritos na seção 6.3— em ambos cenários. Além disso, para observar a consistência e replicabilidade das implementações, foram executados múltiplos testes com variações nos conjuntos de dados, sendo as mesmas variações aplicadas em ambas implementações.

As execuções dos experimentos de referência e dos experimentos da solução desenvolvida, ocorreram em momentos diferentes, por vezes intercaladas, devido ao com-

partilhamento dos recursos computacionais. No entanto, nenhuma atualização ou modificação foi realizada nos recursos durante todo o processo.

Para a avaliação da qualidade dos resultados, foi utilizada a métrica de acurácia. A métrica aplicada segue a fórmula 6.1, considerando o número de classificações corretas em relação ao total de sequências classificadas, considerando classificações inconclusivas como incorretas.

$$\text{Acurácia} = \frac{\text{Classificações Corretas}}{\text{Total de Classificações}} \tag{6.1}$$

6.2 Recursos Computacionais

Todos os experimentos foram realizados usando o mesmo equipamento. O computador utilizado tem em sua configuração, descrita na tabela 6.1, duas unidades centrais de processamento Intel(R) Xeon(R) Silver 4114, 48GB de memória RAM e uma unidade de processamento gráfico NVIDIA RTX A6000. O sistema operacional usado foi o Linux Ubuntu.

Tabela 6.1: Configuração do computador usado.

Hardware	
CPU's	2x Intel(R) Xeon(R) Silver 4114 CPU
Memória	46GB DDR4 @ 2666MHz
GPU	NVIDIA RTX A6000 com 48GB GDDR6
Software	
Sistema Operacional	Linux Ubuntu 22.04.4 LTS
Python	Python 3.10
PyTorch	PyTorch 2.4.0
QIIME2	2024.10.1

6.3 Dados

O conjunto de dados usado no estudo é proveniente da base de dados PR2[24], disponibilizados na versão 5.0.0 do *dataset* de sequências anotadas. A escolha desse conjunto de dados deu-se por ele ser curado e apresentar menor quantidade de ruídos e ambiguidades que outras bases comumente usadas, como SILVA[29] e GreenGenes[10]. Somado a isso, o *dataset* possui um bom volume de dados de sequências 18S e com tamanhos adequados ao estudo. Ao todo, 221.085 registros compõem o conjunto de dados original, contendo as respectivas sequências, classificações taxonômicas e metadados complementares.

Diferentemente das classificações taxonômicas mais tradicionais, a classificação das sequências da PR2 é dividida em 9 grupos. São esses, em ordem hierárquica: Domain, Supergroup, Division, Subdivision, Class, Order, Family, Genus e Species. Contudo, esse estudo abordou apenas os cinco últimos, os quais são mais frequentemente utilizados na literatura e apresentam maior aplicabilidade.

Ainda que selecionado por ter dados com poucos ruídos e ambiguidades, nem todos os registros presentes são relevantes ou compatíveis com as aplicações em estudo, seja pelo segmento sequenciado não ser o 18S rRNA, por ser uma sequência muito curta ou pela amostra não ser proveniente do núcleo. Por esse motivo, primeiramente foi realizado o pré-processamento do *dataset*.

6.3.1 Pré-processamento dos dados

A preparação dos dados foi dividida em duas etapas principais: filtragem e tratamento. Enquanto a etapa de filtragem removeu dados incompatíveis com o estudo, o tratamento manipulou e criou novos campos adequados às técnicas abordadas.

Ao final do processo de filtragem, restaram 156.681 registros dos 221.085 originais. A filtragem foi composta por:

- Seleção apenas das sequências referentes aos genes do tipo 18S_rRNA
- Seleção das sequências com início na posição 1
- Seleção das sequências com término a partir da posição 900
- Remoção dos registros de amostras genéticas não proveniente do núcleo

O processo de tratamento iniciou com a substituição de valores de controle, usados quando uma classificação é desconhecida, por valores nulos padronizados. Em seguida, foi criado um novo campo chamado "truncated_sequence", o qual era composto pelas 900 bases iniciais das respectivas sequências.

Ao final da preparação, obteve-se um conjunto de dados com os seguintes campos: Domain, Supergroup, Division, Subdivision, Class, Order, Family, Genus, Species e Truncated_Sequence. Ao todo, esse conjunto era composto por 156.681 registros relevantes ao escopo.

Em adição a preparação do conjunto de dados, outras operações foram realizadas durante a geração dos conjuntos de treino e teste. Estas operações envolvem a remoção de registros ambíguos e a limitação de uma quantidade mínima de registros para as classes. Dado que essas operações dependem do nível hierárquico que será trabalhado e da forma de amostragem dos dados, elas foram aplicadas durante a etapa de divisão do conjunto, descrita na sessão 6.3.3.

6.3.2 Distribuição dos Dados

A característica de desbalanceamento presente no *dataset* original manteve-se após a etapa de pré-processamento. Por mais que essa característica seja indesejada no desenvolvimento de modelos de redes neurais profundas, a necessidade de um grande volume de dados inviabiliza a simples eliminação de amostras das classes mais frequentes.

Além disso, técnicas de *data augmentation* conhecidas, que geralmente adicionam variantes dos dados com ruídos, não garantem consistência das informações. Isso acontece pois não se sabe a relevância de cada subsegmento da sequência na identificação da sua classe, logo a adição de ruídos aleatórios pode gerar sequências não representativas ou até que pertençam a outras classes.

Na tabela 6.2 é possível ver algumas informações referentes à distribuição dos dados após o pré-processamento, mas antes da divisão em conjuntos de treino e teste. Na tabela é mostrado, por nível hierárquico, a quantidade total de amostras, o tamanho da maior classe, o desvio padrão dos tamanhos das classes e o tamanho médio e mediano das classes.

Tabela 6.2: Estatísticas da distribuição dos dados.

Nível	**Amostras**	**Número de Amostras da Maior Classe**	**Número de Amostras por Classe**		
			Desvio Padrão	**Média**	**Mediana**
Class	154.209	25.206	2.439,3352	710,6405	63,0
Order	127.850	21.888	1.310,1005	291,2300	30,0
Family	116.458	11.558	517,9292	98,0286	12,0
Genus	121.262	1.803	28,9425	5,6706	1,0
Species	86.618	1.112	8,2965	2,2475	1,0

Para reduzir o impacto do desbalanceamento dos dados nesse estudo, diferentes formas de amostragem e proporções entre conjuntos de treino e teste foram utilizadas, as quais são descritas na sessão 6.3.3. Ainda que não seja uma solução definitiva, a execução de testes com diferentes configurações permite ter a percepção da consistência das soluções, da potencial aplicabilidade e a identificação de pontos passíveis de melhora.

6.3.3 Divisão dos dados de treino e teste

O processo de divisão do conjunto de dados é responsável por definir quais dados do conjunto original serão usados no treinamento do modelo e quais serão usados para testar e avaliar o modelo. A escolha dos tamanhos dos subconjuntos e da forma de amostragem influenciam fortemente nos resultados. Sendo assim, para garantir a consistência

e significância dos resultados, pra cada nível hierárquico foram realizados múltiplos testes com diferentes divisões dos dados. Essas divisões variam quanto a forma de amostragem, proporção dos tamanhos dos subconjuntos e restrição de tamanho mínimo de cada classe.

A amostragem dos dados foi realizada com duas técnicas, a amostragem aleatória simples e a estratificada. Enquanto a amostragem aleatória simples seleciona a quantidade de dados desejada de forma aleatória e sem restrições, a técnica estratificada seleciona separadamente os dados de cada classe. Ainda que ambas sejam aleatórias, na estratificada a quantidade de registros selecionados de cada classe varia de acordo com a frequência da classe dentro do conjunto original. Assim, a amostragem estratificada gera subconjuntos representativos e garante que todas as classes tenham registros presentes tanto no conjunto de treino quanto no de teste.

Considerando o tamanho do conjunto de dados, a distribuição dos registros e proporções comumente utilizados na literatura, foram escolhidas as seguintes proporções para, respectivamente, treino e teste: 80% e 20%, 90% e 10%, 95% e 05%.

As proporções das divisões foram escolhidas buscando observar o comportamento na capacidade de aprendizado das soluções. O equilíbrio entre os dados de teste e treino é de suma importância para o desempenho adequado do modelo, pois usar um conjunto de dados muito grande pra treino permite maior aprendizagem pelo modelo, mas resulta em menos dados para avaliação dos resultados. Por outro lado, um conjunto maior de teste permite alcançar maior precisão na avaliação do modelo, mas consequentemente restariam menos dados para treinar o modelo, resultando em piores resultados.

Também foi adicionada uma limitação de quantidade mínima de registros para cada classe, sendo os valores limites utilizados de 5 e 10 registros. Essa restrição visa não só reduzir ruídos e *outliers* presentes no conjunto, como garantir que existam registros suficientes para gerar amostras representativas.

As diferentes formas de amostragem, proporções dos tamanho dos subconjuntos e limites dos tamanhos das classe foram combinadas, gerando diferentes configurações de divisão dos dados. Essas configurações são apresentadas na tabela 6.3.

Ainda buscando aumentar o número de experimentos para obter resultados mais significativos, para cada configuração gerada foram executadas divisões com diferentes *seeds*. Os algoritmos utilizados para as amostragens são, na verdade, pseudo-aleatórios. Essa característica faz com que diferentes execuções com a mesma configuração gerem os mesmos resultados. Para obter resultados distintos, é necessário adicionar valores distintos que são usados internamente como modificadores, chamados de *seeds*. Os valores de *seed* utilizados foram 0, 14, 56, 92, 84, 101, 105 e 227, resultando em oito variantes de cada configuração válida.

Ao final, cada uma das variantes das configurações de divisão foi aplicada, quando possível, a cada um dos cinco níveis hierárquicos definidos, removendo também informa-

Tabela 6.3: Configurações das divisões dos dados.

Amostragem	**Proporção**	**Limite Mínimo**
Aleatória Simples	80:20	5
Aleatória Simples	80:20	10
Aleatória Simples	90:10	5
Aleatória Simples	90:10	10
Aleatória Simples	95:05	5
Aleatória Simples	95:05	10
Estratificada	80:20	5
Estratificada	80:20	10
Estratificada	90:10	5
Estratificada	90:10	10
Estratificada	95:05	10

ções dos níveis subsequentes ao aplicado. Algumas configurações não puderam ser aplicadas no nível de Species por limitações quanto a quantidade mínima de dados. Como resultado, foram geradas 400 divisões de dados, sendo cada uma delas usada em um experimento de referência e um experimento da solução desenvolvida. A relação de todas as configurações geradas é apresentada no apêndice A.

6.4 Experimentos de Referência

A execução dos experimentos de classificação com Naïve Bayes foi realizada com o *plug-in* q2-feature-classifier. O protocolo e configurações usadas seguem o apresentado no tutorial oficial da ferramenta [25]. Para cada par de conjuntos de treino e teste gerados na divisão de dados, um experimento é realizado individualmente.

O fluxo de execução de cada experimento segue o apresentado no diagrama 6.1. Na primeira parte da execução do q2-feature-classifier, são realizadas as operações de importação dos dados e treinamento do modelo, com as configurações padrões presentes no tutorial da ferramenta. Esse é o momento em que o modelo tem acesso ás informações de referência e gera um classificador pré-treinado.

O *plug-in* trabalha com dois formatos de arquivos, o FASTA com as sequências e um arquivo de texto com as classificações. Como os conjuntos de treino e teste gerados na etapa de pré-processamento são armazenados em um arquivo CSV, eles precisaram ser formatados e exportados nos formatos compatíveis com o *plug-in*, mas sem alterações nas informações originais.

Após gerar o classificador pré-treinado, ocorre a etapa de testes. Nessa etapa, o conjunto de teste é inserido no classificador, o qual gera as classificações para cada registro. A ferramenta permite também a exportação dos resultados em formato de texto,

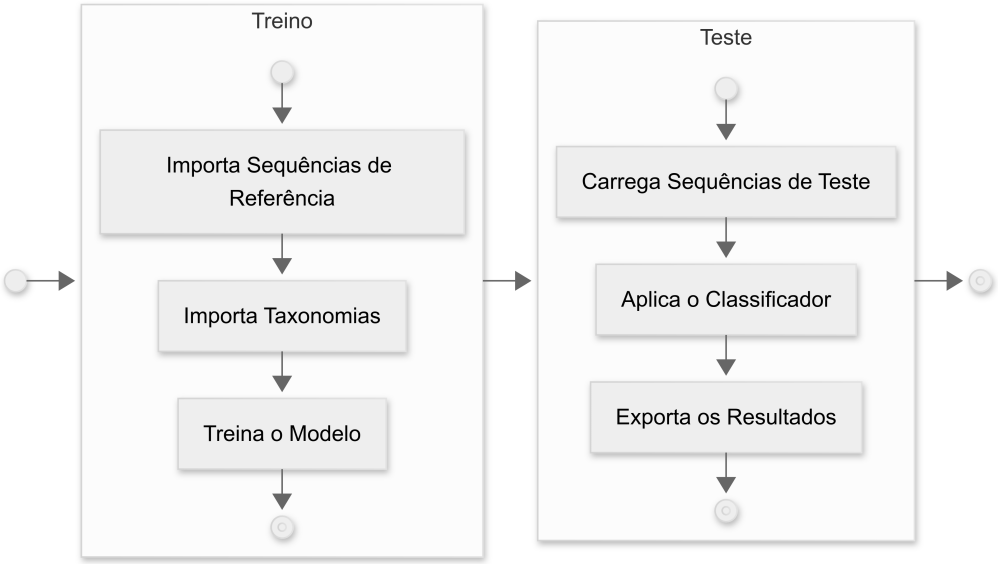

Figura 6.1: Diagrama da execução do *plug-in* q2-feature-classifier.

facilitando a utilização em ferramentas externas que não reconheçam os formatos usados internamente.

A duração das etapas foi avaliada e os tempos médios, agrupados por cada dupla de nível e proporção, estão presentes na tabela 6.4. Mais detalhes da duração das etapas podem ser encontrados no apêndice D.1.

Nota-se que os tempos de treino são consideravelmente baixos se comparados com a duração normal do processo de treinamento de modelos de *deep learning*. Já no tempo médio de teste é possível perceber um aumento não linear em relação ao tamanho do conjunto de dados. Observa-se, por exemplo, em nível de Species o conjunto com 5% do total de sequências demorou 20 segundos para ser classificado, enquanto o conjunto com 20% dos dados demorou 24 segundos.

A classificação das sequências de teste seguiram as configurações pré-definidas pela própria ferramenta. Por padrão, a ferramenta tenta classificar todos os níveis taxonômicos disponíveis, então, por exemplo, num experimento em nível de Ordem a ferramenta tenta classificar em Ordem e em todos os níveis antecessores. O mesmo ocorre em todos os demais níveis, onde o nível desejado é classificado junto com todos os níveis superiores.

No entanto, existe um limite de confiança mínima pré-definido no classificador para que a atribuição de uma classe em um dado nível seja feita. Caso a classificação de um registro não alcance a confiança mínima em algum nível, a classificação é interrompida naquele ponto e é considerada só a classificação dos níveis anteriores. Considerando que os experimentos buscam avaliar a classificação nos níveis separadamente, os resultados inconclusivos, onde nenhuma classe foi atribuída no nível em estudo, foram considerados como incorretos.

Tabela 6.4: Tempos médios de duração das etapas de treino e teste dos experimentos usando o q2-feature-classifier.

Nível	**Proporção**	**Tempo Médio de Treino**	**Tempo Médio de Teste**
Class	80:20	0:01:43	0:00:38
Class	90:10	0:01:52	0:00:29
Class	95:05	0:01:56	0:00:24
Order	80:20	0:01:37	0:00:37
Order	90:10	0:01:45	0:00:28
Order	95:05	0:01:47	0:00:24
Family	80:20	0:01:37	0:00:37
Family	90:10	0:01:46	0:00:28
Family	95:05	0:01:46	0:00:24
Genus	80:20	0:01:41	0:00:37
Genus	90:10	0:01:49	0:00:29
Genus	95:05	0:01:26	0:00:23
Species	80:20	0:00:46	0:00:24
Species	90:10	0:00:48	0:00:22
Species	95:05	0:00:40	0:00:20

Após a exportação das classificações obtidas nos testes de todos os experimentos com o *plug-in*, foi realizada uma análise para avaliar o desempenho da ferramenta. Na tabela 6.5 é apresentada uma sumarização das acurácias dos resultados, sendo possível encontrar a apresentação das métricas individuais de cada experimento no apêndice B.

Em relação aos resultados obtidos, observa-se que a forma de amostragem, salvo em nível de Species, não teve grande impacto na acurácia, com as acurácias das amostragens aleatórias simples e das estratificadas sendo muito próximas. Nota-se também que o classificador demonstrou consistência, com pequenas variações e bons resultados nos três níveis superiores (Class, Order e Family). Já nos níveis de Genus e Species o classificador apresentou grande inconsistência e maus resultados, chegando a errar praticamente todas as classificações em alguns testes.

6.5 Experimentos da Solução Desenvolvida

Para a avaliação da solução desenvolvida, foram realizados, ao todo, 400 experimentos envolvendo o treinamento e teste do modelo definido. Esses experimentos utilizaram os conjuntos de dados de treino e teste descritos na sessão 6.3.3, sendo eles os mesmos utilizados nos experimentos de referência.

Tabela 6.5: Acurácia dos experimentos usando o q2-feature-classifier.

Nível	**Amostragem**	**Quantidade de Experimentos**	**Acurácia**		
			Mínima	**Média**	**Máxima**
Class	Aleatória Simples	40	0,9409	0,9467	0,9500
Class	Estratificada	40	0,9452	0,9477	0,9514
Order	Aleatória Simples	40	0,9446	0,9492	0,9544
Order	Estratificada	40	0,9439	0,9481	0,9547
Family	Aleatória Simples	40	0,9368	0,9419	0,9487
Family	Estratificada	40	0,9372	0,9423	0,9475
Genus	Aleatória Simples	40	0,7851	0,8235	0,8512
Genus	Estratificada	40	0,7930	0,8255	0,8580
Species	Aleatória Simples	40	0,0000	0,3563	0,8971
Species	Estratificada	40	0,0009	0,3037	0,8946

6.5.1 Configurações e Parâmetros

Para o treinamento do modelo, foram utilizados os parâmetros conforme descritos na tabela 6.6.

Tabela 6.6: Principais hiperparâmetros utilizados no treinamento.

Hiperparâmetros	
Número de épocas	700
Taxa de aprendizado inicial	0,005
Ajuste da taxa de aprendizado	CosineAnnealingWarmRestarts
Função de custo	CrossEntropyLoss
Otimizador	AdamW com AMSGrad
Tamanho dos lotes	Variável

6.5.1.1 Batch Sizes

O tamanho dos lotes, *batch size*, varia de acordo com a época em execução. Essa escolha deu-se visando aproveitar a rapidez do aprendizado com *batches* maiores em alguns momentos e adicionar pequenos ruídos com o uso de *batches* menores em outros momentos.

O tamanho dos *batches* utilizados de acordo com a época segue o apresentado na tabela 6.7. Os valores 128, 256 e 1.000 foram escolhidos para os *batches* menores por serem recorrentes na literatura. Já o tamanho de 15.000 registros para os *batches* maiores foi escolhido após testes iniciais, nos quais esse foi o maior tamanho utilizado sem causar falhas por falta de recursos de memória de vídeo.

Tabela 6.7: Tamanho dos *batches* de acordo com as épocas.

Épocas	**Tamanho dos Lotes**
[0, 45]	15.000
[46, 91]	15.000
[92, 137]	15.000
[138, 183]	1.000
[184, 229]	15.000
[230, 275]	15.000
[276, 321]	15.000
[322, 367]	15.000
[368, 413]	256
[414, 459]	15.000
[460, 505]	15.000
[506, 551]	15.000
[552, 597]	15.000
[598, 643]	128
[644, 699]	15.000

6.5.1.2 Número de Épocas

A fim de alcançar melhores resultados e considerando o desconhecimento prévio do comportamento da convergência do aprendizado, não foi utilizada nenhuma técnica de parada, ou em inglês *“early stopping”*. Com isso, todos os experimentos foram realizados com o mesmo número de épocas.

Apesar dessa escolha aumentar significativamente o tempo médio de treino, ela também permite uma avaliação mais aprofundada do comportamento do modelo, potenciais melhorias e até auxiliar identificação de problemas, como *overfittinig*.

O número de épocas foi definido como sendo 700, visto que em testes preliminares foi observado que o aprendizado convergia antes da 700ª época em todos os casos até então executados. Na tabela 6.8 são apresentadas as métricas referentes as melhores épocas dos experimentos realizados.

Nas métricas, a mínima de melhor época representa quantas épocas foram computadas até alcançar a acurácia mais alta no melhor caso, ou seja, no treinamento que alcançou o melhor resultado mais rapidamente. A máxima de melhor época por sua vez representa o número de épocas que o pior treinamento precisou para alcançar o melhor resultado. Por fim, as métricas de média e mediana são referentes a quantidade média e mediana de épocas que os treinamentos precisaram para alcançar as suas melhores acurácias.

Tabela 6.8: Métricas das melhores épocas agrupadas por nível e tipo de amostragem.

Nível	**Amostragem**	**Quantidade de Experimentos**	**Melhor Época**			
			Mínima	**Média**	**Mediana**	**Máxima**
Class	Aleatória Simples	40	144	292	303	310
Class	Estratificada	40	145	295	303	310
Order	Aleatória Simples	40	295	304	305	310
Order	Estratificada	40	292	313	306	623
Family	Aleatória Simples	40	298	313	305	629
Family	Estratificada	40	294	329	306	630
Genus	Aleatória Simples	40	303	601	626	630
Genus	Estratificada	40	299	586	626	630
Species	Aleatória Simples	40	572	615	616	629
Species	Estratificada	40	585	614	614	628

6.5.2 Fluxo de Execução

O fluxo de execução dos experimentos da solução desenvolvida é composto, sequencialmente, por uma etapa de carregamento dos dados, uma de inicialização do modelo e, para cada época, uma etapa de treino e uma de teste, como apresentado no diagrama 6.2. Após a execução de todas as épocas, a melhor época é selecionada junto com os pesos do respectivo modelo.

A primeira etapa carrega os conjuntos de dados de treino e teste referentes ao experimento em questão. Nessa fase, os dados dos arquivos são carregados em uma estrutura pré-definida compatível com os padrões indicados pela biblioteca PyTorch. Além disso, as sequências são manipuladas para gerar as respectivas representações vetoriais, como descrito na sessão 5.1.

Na segunda etapa, o modelo criado é inicializado com os parâmetros pré-definidos nos testes. Usando ferramentas disponibilizadas pela própria biblioteca PyTorch, o modelo inicializado é também compilado para melhorar o seu desempenho e utilização dos recursos de *hardware* disponíveis.

Por fim, uma etapa de treino e uma etapa de teste são executadas sequencialmente em cada época do experimento. Após todas as épocas terem sido computadas, a com melhor acurácia nos testes é selecionada e o respectivo modelo pré-treinado é exportado, permitindo utilizações futuras sem a necessidade de novos treinamentos.

6.5.3 Resultados

Após a realização dos 400 testes, abrangendo os cinco níveis taxonômicos em estudo, foi possível analisar as métricas em relação as acurácias obtidas. Os valores

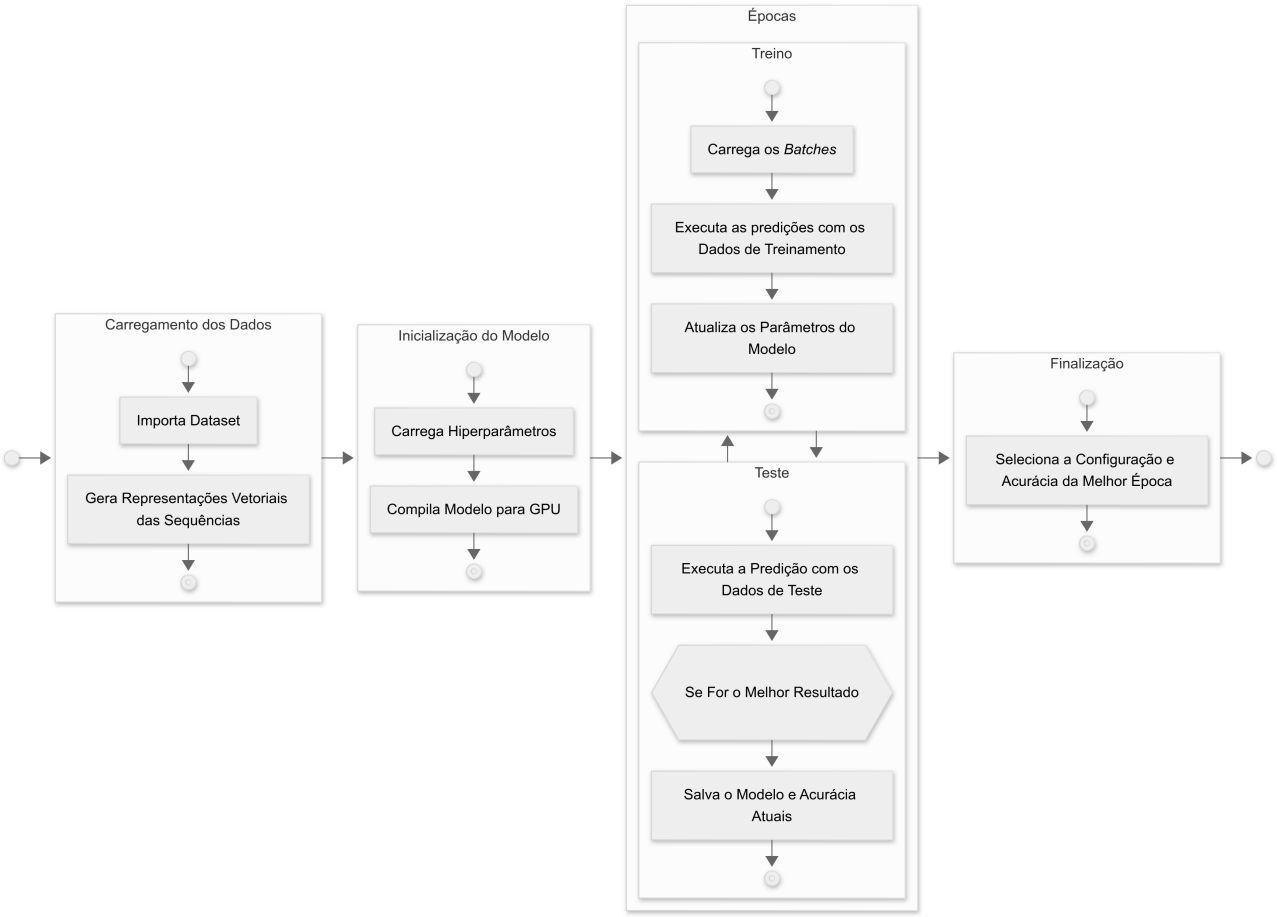

Figura 6.2: Diagrama da execução dos experimentos com a solução desenvolvida.

agregados são apresentados na tabela 6.9, já os valores individuais estão presentes no apêndice C.

Com base nas acurácias observadas, é possível perceber que, em cada nível, a diferença absoluta dos valores mínimos, médios e máximos entre as técnicas de amostragem são sempre inferiores a 0.017, 0.009 e 0.004, respectivamente. Além disso, a pequena diferença entre os valores mínimos e máximos em cada grupo nos níveis de Class, Order e Family, evidencia a consistência da solução em níveis superiores.

6.5.4 Tempos de Execução

Durante os experimentos, foram coletados os tempos de treinamento e de predição. Os valores agregados são apresentados na tabela 6.10, já os valores individuais estão presentes no apêndice E.

É sabido que o treinamento de modelos tende a ser demorado pela complexidade e pelo grande volume de dados que precisam ser processados. Apesar de não terem sido

Tabela 6.9: Acurácia dos experimentos usando a solução desenvolvida.

Nível	Amostragem	Quantidade de Experimentos	Acurácia		
			Mínima	Média	Máxima
Class	Aleatória Simples	40	0,9876	0,9900	0,9921
Class	Estratificada	40	0,9884	0,9905	0,9922
Order	Aleatória Simples	40	0,9818	0,9852	0,9880
Order	Estratificada	40	0,9776	0,9854	0,9907
Family	Aleatória Simples	40	0,9708	0,9756	0,9810
Family	Estratificada	40	0,9710	0,9758	0,9802
Genus	Aleatória Simples	40	0,8422	0,8880	0,9203
Genus	Estratificada	40	0,8585	0,8952	0,9223
Species	Aleatória Simples	40	0,8515	0,9088	0,9537
Species	Estratificada	40	0,8648	0,9171	0,9503

Tabela 6.10: Tempos das etapas de treinamento e predição agrupados por nível e proporção.

Nível	Proporção	Tempo de Treinamento			Tempo de Predição		
		Mínimo	Médio	Máximo	Mínimo	Médio	Máximo
Class	80:20	1:32:02	1:33:02	1:35:19	0:00:37	0:00:38	0:00:43
Class	90:10	1:33:20	1:37:24	1:40:18	0:00:17	0:00:19	0:00:23
Class	95:05	1:41:30	1:42:12	1:43:30	0:00:12	0:00:13	0:00:17
Order	80:20	1:15:54	1:16:39	1:18:25	0:00:34	0:00:34	0:00:35
Order	90:10	1:20:53	1:22:28	1:25:55	0:00:17	0:00:17	0:00:22
Order	95:05	1:22:54	1:23:40	1:26:08	0:00:12	0:00:13	0:00:16
Family	80:20	1:02:43	1:03:31	1:05:03	0:00:30	0:00:31	0:00:34
Family	90:10	1:06:08	1:07:16	1:09:21	0:00:15	0:00:16	0:00:20
Family	95:05	1:08:58	1:11:24	1:13:02	0:00:11	0:00:12	0:00:15
Genus	80:20	0:34:33	0:38:04	0:42:09	0:00:16	0:00:17	0:00:22
Genus	90:10	0:35:24	0:39:26	0:44:05	0:00:12	0:00:13	0:00:17
Genus	95:05	0:36:03	0:36:44	0:38:42	0:00:10	0:00:11	0:00:14
Species	80:20	0:08:51	0:11:58	0:15:21	0:00:10	0:00:11	0:00:15
Species	90:10	0:09:02	0:12:25	0:15:47	0:00:09	0:00:10	0:00:14
Species	95:05	0:09:14	0:09:35	0:10:05	0:00:08	0:00:09	0:00:11

utilizadas nesse estudo, existem técnicas que visam melhorar esse aspecto e que podem ser exploradas na otimização da solução de acordo com o cenário de implementação.

Outro ponto importante é referente ao processo de predição, o qual é muito mais rápido se comparado com o treinamento. A paralelização do processamento de sequências pelo modelo pré-treinado também possibilita a classificação de conjuntos de dados maiores sem ter um aumento linear no tempo de predição.

Um exemplo desse fato pode ser observado ao comparar os tempos de predição das proporções 95:05 com 80:20 no nível de Genus. A relação entre os tamanhos dos conjuntos de teste é de 4 vezes, mas a relação entre os tempos é inferior a 1,6 vezes.

6.5.5 Utilização dos Recursos

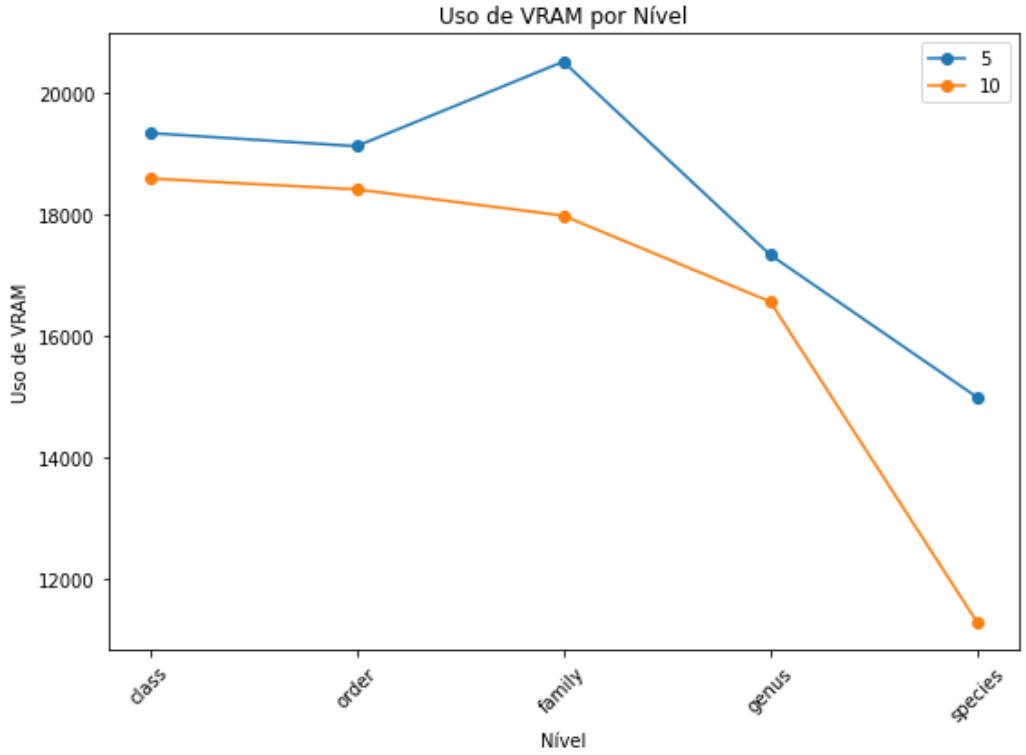

Figura 6.3: Uso de VRAM de cada limite mínimo, para cada nível.

Tabela 6.11: Uso de VRAM na etapa de treino.

Nível	**Limite Mínimo**	**Total de Sequências**	**Uso de VRAM(GB) no Treino**		
			Mínimo	**Médio**	**Máxima**
Class	5	123.607	18.446,00	19.338,68	20.428,00
Class	10	123.487	16.782,00	18.591,00	20.408,00
Order	5	100.974	17.538,00	19.121,93	20.156,00
Order	10	100.563	17.020,00	18.411,54	20.258,00
Family	5	82.965	17.706,00	20.514,31	24.274,00
Family	10	81.939	16.848,00	17.978,91	19.906,00
Genus	5	48.838	16.742,00	17.334,50	18.294,00
Genus	10	41.683	15.848,00	16.562,95	17.264,00
Species	5	16.901	14.454,00	14.990,00	16.090,00
Species	10	11.667	10.150,00	11.288,29	12.754,00

Devido ao tempo de desenvolvimento, as implementações de modelos de *deep learning* em cenários reais dependem de *hardware* especializado para serem consideradas hábeis, sendo essa dependência um importante fator limitante. Para grandes volumes de dados, um dos recursos mais demandados é a memória, sendo a VRAM quando utilizado o processamento em GPU. No gráfico 6.3 é representado o uso médio de VRAM no

Tabela 6.12: Uso de VRAM na etapa de predição.

Nível	Limite Mínimo	Total de Sequências	Uso de VRAM(GB) na Predição		
			Mínimo	Médio	Máxima
Class	5	123.607	6.848,00	7.952,37	9.680,00
Class	10	123.487	3.168,00	6.467,29	8.916,00
Order	5	100.974	5.604,00	7.173,12	8.760,00
Order	10	100.563	2.034,00	5.535,87	8.762,00
Family	5	82.965	4.716,00	6.708,06	8.578,00
Family	10	81.939	1.832,00	5.091,16	8.428,00
Genus	5	48.838	2.644,00	4.394,50	6.072,00
Genus	10	41.683	1.106,00	2.806,66	5.572,00
Species	5	16.901	928,00	1.398,81	2.630,00
Species	10	11.667	506,00	843,66	1.862,00

treinamento do modelo em cada nível hierárquico, podendo-se notar uma tendencia a redução em níveis mais específicos, os quais costumam ter um menor volume de dados.

Para mitigar potenciais problemas decorrentes da alta demanda de VRAM, é possível reduzir o tamanho dos *batches*, dividindo os conjuntos de dados em lotes menores para serem carregados separadamente. No entanto, com uma quantidade maior de *batches*, o processo de carregamento e de atualização dos valores da rede neural serão mais complexos, acarretando em um maior tempo de treinamento.

Além disso, diversas GPU's mais recentes possuem estruturas dedicadas para trabalhar com operações de IA, como os chamados *tensor cores* da NVIDIA. Isso permite a implementação de novas técnicas de otimização, como a compilação específica do modelo para essas estruturas.

Também é importante observar que, da mesma forma que ocorre em relação aos tempos de treinamento e de teste, a demanda por recursos de hardware é muito menor nas etapas de teste ou predição do que na etapa de treinamento. A tabela 6.11 e a tabela 6.12 apresentam, respectivamente, métricas do uso de VRAM nas etapas de treino e de predição, onde em nível de Species com limite mínimo de 10 registros, por exemplo, a demanda média de VRAM nas predições foi inferior a 8% em relação aos treinamentos. Isso permite treinar o modelo uma única vez em um equipamento mais potente e usar o modelo pré-treinado resultante em equipamentos mais limitados.

6.6 Análise

A grande quantidade de experimentos, 400 de referência e 400 da solução desenvolvida, junto com as variações nos conjuntos de treino e teste utilizados, permitiram analisar a consistência dos resultados.

As acurácias alcançadas nos experimentos com a solução desenvolvida são superiores as alcançadas nos experimentos de referências, como apresentado na tabela 6.13. Nota-se também que discrepância entre as acurácias dos experimentos da solução e dos de referência é ainda maior no nível de Species. Nesse nível, enquanto o plug-in não conseguiu resultados relevantes e consistentes, a solução obteve acurácia média superior a 90% e mínima superior a 85%.

Tabela 6.13: Comparação das acurácias dos experimentos de referência e dos experimentos da solução desenvolvida.

Nível	**Amostragem**	**Acurácia Média**	
		Referência	**Solução**
Class	Aleatória Simples	0,9467	0,9900
Class	Estratificada	0,9477	0,9905
Order	Aleatória Simples	0,9492	0,9852
Order	Estratificada	0,9481	0,9854
Family	Aleatória Simples	0,9419	0,9756
Family	Estratificada	0,9423	0,9758
Genus	Aleatória Simples	0,8235	0,8880
Genus	Estratificada	0,8255	0,8952
Species	Aleatória Simples	0,3563	0,9088
Species	Estratificada	0,3037	0,9171

Em contra-partida, ainda que solução desenvolvida possua tempos de teste e predição menores, os tempos de treinamento são muito maiores quando comparados com os experimentos de referência. A aplicação de novas técnicas e o uso de equipamentos otimizados com tecnologias mais recentes para processamento de IA pode ajudar a reduzir os tempos, mas, devido a complexidade dos modelos CNN, é esperado que o processo de treinamento seja consideravelmente maior que o processo de predição.

7. CONCLUSÕES

Esse trabalho teve como objetivo central explorar a aplicabilidade de *deep learning* para a classificação taxonômica por meio da criação e análise de um modelo de rede neural profunda. O estudo considerou um cenário mais abrangente, não limitando sub-ramos da hierarquia taxonômica e demonstrando o potencial de aplicação tanto em contextos genéricos quanto específicos.

Como principal fruto desse trabalho, obteve-se um modelo de *deep learning* composto por camadas de convolução, *pooling* e camadas totalmente conectadas. O modelo foi calibrado após uma série de experimentos, aliando alta acurácia, baixo custo relativo para treinamento e rapidez na classificação.

Para a análise, foi feita uma comparação entre o modelo desenvolvido e uma técnica de classificação amplamente utilizada atualmente, a qual é baseada em Naïve Bayes e implementada pelo *plug-in* q2-feature-classifier da plataforma QIIME2. Após analisar os resultados dos 400 experimentos realizados com cada ferramenta, o *plug-in* e a solução desenvolvida, foi possível identificar que a solução desenvolvida obteve melhores resultados, alcançando acurácias melhores e de forma mais consistente.

A menor diferença entre as acurácias médias da solução desenvolvida e da referência foi no nível Family, no qual a solução desenvolvida teve resultado aproximadamente 3,5% superior. Já no nível de Species foi onde obteve-se maior discrepância entre as acurácias médias, com a solução desenvolvida alcançando o triplo da acurácia média do classificador de referência.

A solução proposta também obteve tempos de classificação menores. No entanto, há uma limitação considerável quanto aos recursos utilizados e ao tempo de treinamento significativamente maior.

7.1 Considerações Finais e Trabalhos Futuros

As limitações decorrentes da alta demanda de capacidade computacional e dos longos tempos de treinamento são pontos negativos recorrentes na utilização de *deep learning*. O crescente interesse em inteligência artificial está resultando em um rápido desenvolvimento de novas tecnologias por pesquisadores e pela indústria. Sendo assim, a continuidade deste trabalho por meio da exploração de técnicas mais recentes e da utilização de novos equipamentos especializados pode resultar em melhorias consideráveis no que diz respeito, principalmente, aos tempos de treinamento e aos custos, impactando na aplicabilidade e replicabilidade da solução.

Também foi observado que os resultados variam de acordo com o volume e qualidade dos dados, especialmente em níveis mais específicos da hierarquia taxonômica. Ainda que a solução desenvolvida tenha obtido resultados positivos, com acurácia significativa e demonstrando grande potencial, um conjunto de dados melhor e maior tende a alcançar resultados ainda melhores.

Além disso, a utilização de um modelo baseado em CNN também permite *incremental learning*, onde novos dados podem ser adicionadas a um modelo pré-treinado. Isso permite que os modelos sejam atualizados com novos dados ou refinados com dados específicos.

Por fim, esse trabalho considerou a classificação de organismos vivos apenas por meio do segmento 18S rRNA da sequência genética. Sabendo-se da existência de outros segmentos, como o 16S rRNA, que são amplamente usados em diferentes estudos e contextos, mostra-se interessante a avaliação da solução desenvolvida com conjuntos de dados de outros segmentos genéticos.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Afify, H. M.; Al-Masni, M. A. “Taxonomy metagenomic analysis for microbial sequences in three domains system via machine learning approaches”, *Informatics in Medicine Unlocked*, vol. 13, 2018, pp. 151–157.

[2] Alberdi, A.; Aizpurua, O.; Gilbert, M. T. P.; Bohmann, K. “Scrutinizing key steps for reliable metabarcoding of environmental samples”, *Methods in Ecology and Evolution*, vol. 9–1, Jan 2018, pp. 134–147.

[3] Barnes, M. A.; Turner, C. R. “The ecology of environmental DNA and implications for conservation genetics”, *Conservation Genetics*, vol. 17–1, Fev 2016, pp. 1–17.

[4] Bohmann, K.; Evans, A.; Gilbert, M. T. P.; Carvalho, G. R.; Creer, S.; Knapp, M.; Yu, D. W.; De Bruyn, M. “Environmental DNA for wildlife biology and biodiversity monitoring”, *Trends in Ecology & Evolution*, vol. 29–6, Jun 2014, pp. 358–367.

[5] Bokulich, N.; Kaehler, B.; Rideout, J. R.; Dillon, M.; Bolyen, E.; Knight, R.; Huttley, G.; Caporaso, J. “Optimizing taxonomic classification of marker-gene amplicon sequences with qiime 2’s q2-feature-classifier plugin”, *Microbiome*, vol. 6, Article 90, Maio 2018, DOI 10.1186/s40168-018-0470-z. Disponível em: https://doi.org/10.1186/s40168-018-0470-z. Acesso em: Jan 2024.

[6] Cornish-Bowden, A. “Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984.”, *Nucleic acids research*, vol. 13–9, 1985, pp. 3021–3030.

[7] Darwin, C. “On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life”. Londres: John Murray, 1859, 502p.

[8] Dully, V.; Balliet, H.; Frühe, L.; Däumer, M.; Thielen, A.; Gallie, S.; Berrill, I.; Stoeck, T. “Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture – An inter-laboratory study”, *Ecological Indicators*, vol. 121, Article ID 107049, Fev 2021, DOI 10.1016/j.ecolind.2020.107049. Disponível em: https://doi.org/10.1016/j.ecolind.2020.107049. Acesso em: Jan 2024.

[9] Floyd, R.; Abebe, E.; Papert, A.; Blaxter, M. “Molecular barcodes for soil nematode identification”, *Molecular ecology*, vol. 11–4, 2002, pp. 839–850.

[10] GreenGenes. Capturado em: https://greengenes.secondgenome.com/, Jan 2024.

[11] Hebert, P. D. N.; Cywinska, A.; Ball, S. L.; deWaard, J. R. “Biological identifications through dna barcodes”, *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270–1512, 2003, pp. 313–321.

[12] Helaly, M. A.; Rady, S.; Aref, M. M. “Bert contextual embeddings for taxonomic classification of bacterial dna sequences”, *Expert Systems with Applications*, vol. 208, Article ID 117972, Dez 2022, DOI 10.1016/j.eswa.2022.117972. Disponível em: https://doi.org/10.1016/j.eswa.2022.117972. Acesso em: Jan 2024.

[13] IUPAC-IUB Comm. on Biochem. Nomenclature (CBN). “Abbreviations and symbols for nucleic acids, polynucleotides, and their constituents”, *Biochemistry*, vol. 9–20, Set 1970, pp. 4022–4027, publisher: American Chemical Society.

[14] Karagöz, M. A.; Nalbantoglu, O. U. “Taxonomic classification of metagenomic sequences from relative abundance index profiles using deep learning”, *Biomedical Signal Processing and Control*, vol. 67, Article ID 102539, Maio 2021, DOI 10.1016/j.bspc.2021.102539. Disponível em: https://doi.org/10.1016/j.bspc.2021.102539. Acesso em: Jan 2024.

[15] Kishk, A.; El-Hadidi, M. “Ampliconnet: Sequence based multi-layer perceptron for amplicon read classification using real-time data augmentation”. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 2413–2418.

[16] LeCun, Y.; Bengio, Y.; Hinton, G. “Deep learning”, *Nature*, vol. 521–7553, Maio 2015, pp. 436–444.

[17] Linnæus, C. “Systema naturæ, sive regna tria naturæ systematice proposita per classes, ordines, genera, & species”. Lugduni Batavorum. (Haak), 1735, vol. 1.

[18] Liu, M.; Clarke, L. J.; Baker, S. C.; Jordan, G. J.; Burridge, C. P. “A practical guide to DNA metabarcoding for entomological ecologists”, *Ecological Entomology*, vol. 45–3, Jun 2020, pp. 373–385.

[19] Moore, R. T. “Proposal for the recognition of super ranks”, *Taxon*, vol. 23–4, 1974, pp. 650–652.

[20] National Human Genome Research Institute. “Primer”. Capturado em: https://www.genome.gov/genetics-glossary/Primer, Fev 2024.

[21] National Human Genome Research Institute. “Shotgun sequencing”. Capturado em: https://www.genome.gov/genetics-glossary/Shotgun-Sequencing, Fev 2024.

[22] Park, H.; Lim, S. J.; Cosme, J.; O’Connell, K.; Sandeep, J.; Gayanilo, F.; Jr, G. R. C.; Montes, E.; Nitikitpaiboon, C.; Fisher, S.; Moustahfid, H.;

Thompson, L. R. “Investigation of machine learning algorithms for taxonomic classification of marine metagenomes”, *Microbiology Spectrum*, vol. 11-5, e05237-22, Set 2023, DOI 10.1128/spectrum.05237-22. Disponível em: https://doi.org/10.1128/spectrum.05237-22. Acesso em: Jan 2024.

[23] Porter, T. M.; Hajibabaei, M. “Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis”, *Molecular Ecology*, vol. 27–2, Jan 2018, pp. 313–338.

[24] PR2. Capturado em: https://pr2-database.org/, Set 2024.

[25] QIIME2. “Training feature classifiers with q2-feature-classifier”. Capturado em: https://docs.qiime2.org/2024.5/tutorials/feature-classifier/, Set 2024.

[26] Rishan, S. T.; Kline, R. J.; Rahman, M. S. “Applications of environmental DNA (eDNA) to detect subterranean and aquatic invasive species: A critical review on the challenges and limitations of eDNA metabarcoding”, *Environmental Advances*, vol. 12, Article ID 100370, Abr 2023, DOI 10.1016/j.envadv.2023.100370. Disponível em: https://www.sciencedirect.com/science/article/pii/S2666765723000303. Acesso em: Jan 2024.

[27] Sarker, I. H. “Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions”, *SN Computer Science*, vol. 2-6, Article 420, Ago 2021, DOI 10.1007/s42979-021-00815-1. Disponível em: https://doi.org/10.1007/s42979-021-00815-1. Acesso em: Jan 2024.

[28] Scikit-Learn. “Encoding categorical features”. Capturado em: https://scikit-learn.org/stable/modules/preprocessing.html#encoding-categorical-features, Mar 2025.

[29] SILVA. Capturado em: http://www.arb-silva.de/, Jan 2024.

[30] Taberlet, P.; Coissac, E.; Hajibabaei, M.; Rieseberg, L. H. “Environmental DNA”, *Molecular Ecology*, vol. 21–8, 2012, pp. 1789–1793.

[31] Vu, D.; Groenewald, M.; Verkley, G. “Convolutional neural networks improve fungal classification”, *Scientific Reports*, vol. 10, Article ID 12628, Jul 2020, DOI 10.1038/s41598-020-69245-y. Disponível em: https://doi.org/10.1038/s41598-020-69245-y. Acesso em: Jan 2024.

[32] Wang, Q.; Garrity, G. M.; Tiedje, J. M.; Cole, J. R. “Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy”, *Appl Environ Microbiol*, vol. 73–16, Ago 2007, pp. 5261–5267.

[33] Ziemski, M.; Wisanwanichthan, T.; Bokulich, N. A.; Kaehler, B. D. “Beating Naive Bayes at Taxonomic Classification of 16S rRNA

Gene Sequences.”, *Frontiers in microbiology*, vol. 12, Article ID 644487, Jun 2021, DOI 10.3389/fmicb.2021.644487. Disponível em: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2021.644487. Acesso em: Jan 2024.

APÊNDICE A – TABELA DE DIVISÕES

Tabela A.1: Divisão do conjunto em treino e teste para o nível de Class com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Sequências de Treino	Sequências de Teste
Class	Aleatória Simples	80:20	5	0	98.886	24.721
Class	Aleatória Simples	80:20	5	14	98.886	24.721
Class	Aleatória Simples	80:20	5	56	98.886	24.721
Class	Aleatória Simples	80:20	5	84	98.886	24.721
Class	Aleatória Simples	80:20	5	92	98.886	24.721
Class	Aleatória Simples	80:20	5	101	98.886	24.721
Class	Aleatória Simples	80:20	5	105	98.886	24.721
Class	Aleatória Simples	80:20	5	227	98.886	24.721
Class	Aleatória Simples	80:20	10	0	98.790	24.697
Class	Aleatória Simples	80:20	10	14	98.790	24.697
Class	Aleatória Simples	80:20	10	56	98.790	24.697
Class	Aleatória Simples	80:20	10	84	98.790	24.697
Class	Aleatória Simples	80:20	10	92	98.790	24.697
Class	Aleatória Simples	80:20	10	101	98.790	24.697
Class	Aleatória Simples	80:20	10	105	98.790	24.697
Class	Aleatória Simples	80:20	10	227	98.790	24.697
Class	Aleatória Simples	90:10	5	0	111.246	12.361
Class	Aleatória Simples	90:10	5	14	111.246	12.361
Class	Aleatória Simples	90:10	5	56	111.246	12.361
Class	Aleatória Simples	90:10	5	84	111.246	12.361
Class	Aleatória Simples	90:10	5	92	111.246	12.361
Class	Aleatória Simples	90:10	5	101	111.246	12.361
Class	Aleatória Simples	90:10	5	105	111.246	12.361
Class	Aleatória Simples	90:10	5	227	111.246	12.361
Class	Aleatória Simples	90:10	10	0	111.138	12.349
Class	Aleatória Simples	90:10	10	14	111.138	12.349
Class	Aleatória Simples	90:10	10	56	111.138	12.349
Class	Aleatória Simples	90:10	10	84	111.138	12.349
Class	Aleatória Simples	90:10	10	92	111.138	12.349
Class	Aleatória Simples	90:10	10	101	111.138	12.349
Class	Aleatória Simples	90:10	10	105	111.138	12.349
Class	Aleatória Simples	90:10	10	227	111.138	12.349
Class	Aleatória Simples	95:05	10	0	117.313	6.174
Class	Aleatória Simples	95:05	10	14	117.313	6.174
Class	Aleatória Simples	95:05	10	56	117.313	6.174
Class	Aleatória Simples	95:05	10	84	117.313	6.174
Class	Aleatória Simples	95:05	10	92	117.313	6.174
Class	Aleatória Simples	95:05	10	101	117.313	6.174
Class	Aleatória Simples	95:05	10	105	117.313	6.174
Class	Aleatória Simples	95:05	10	227	117.313	6.174

Tabela A.2: Divisão do conjunto em treino e teste para o nível de Class com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Sequências de Treino	Sequências de Teste
Class	Estratificada	80:20	5	0	98.885	24.722
Class	Estratificada	80:20	5	14	98.885	24.722
Class	Estratificada	80:20	5	56	98.885	24.722
Class	Estratificada	80:20	5	84	98.885	24.722
Class	Estratificada	80:20	5	92	98.885	24.722
Class	Estratificada	80:20	5	101	98.885	24.722
Class	Estratificada	80:20	5	105	98.885	24.722
Class	Estratificada	80:20	5	227	98.885	24.722
Class	Estratificada	80:20	10	0	98.789	24.698
Class	Estratificada	80:20	10	14	98.789	24.698
Class	Estratificada	80:20	10	56	98.789	24.698
Class	Estratificada	80:20	10	84	98.789	24.698
Class	Estratificada	80:20	10	92	98.789	24.698
Class	Estratificada	80:20	10	101	98.789	24.698
Class	Estratificada	80:20	10	105	98.789	24.698
Class	Estratificada	80:20	10	227	98.789	24.698
Class	Estratificada	90:10	5	0	111.246	12.361
Class	Estratificada	90:10	5	14	111.246	12.361
Class	Estratificada	90:10	5	56	111.246	12.361
Class	Estratificada	90:10	5	84	111.246	12.361
Class	Estratificada	90:10	5	92	111.246	12.361
Class	Estratificada	90:10	5	101	111.246	12.361
Class	Estratificada	90:10	5	105	111.246	12.361
Class	Estratificada	90:10	5	227	111.246	12.361
Class	Estratificada	90:10	10	0	111.138	12.349
Class	Estratificada	90:10	10	14	111.138	12.349
Class	Estratificada	90:10	10	56	111.138	12.349
Class	Estratificada	90:10	10	84	111.138	12.349
Class	Estratificada	90:10	10	92	111.138	12.349
Class	Estratificada	90:10	10	101	111.138	12.349
Class	Estratificada	90:10	10	105	111.138	12.349
Class	Estratificada	90:10	10	227	111.138	12.349
Class	Estratificada	95:05	10	0	117.312	6.175
Class	Estratificada	95:05	10	14	117.312	6.175
Class	Estratificada	95:05	10	56	117.312	6.175
Class	Estratificada	95:05	10	84	117.312	6.175
Class	Estratificada	95:05	10	92	117.312	6.175
Class	Estratificada	95:05	10	101	117.312	6.175
Class	Estratificada	95:05	10	105	117.312	6.175
Class	Estratificada	95:05	10	227	117.312	6.175

Tabela A.3: Divisão do conjunto em treino e teste para o nível de Order com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Sequências de Treino	Sequências de Teste
Order	Aleatória Simples	80:20	5	0	80.779	20.195
Order	Aleatória Simples	80:20	5	14	80.779	20.195
Order	Aleatória Simples	80:20	5	56	80.779	20.195
Order	Aleatória Simples	80:20	5	84	80.779	20.195
Order	Aleatória Simples	80:20	5	92	80.779	20.195
Order	Aleatória Simples	80:20	5	101	80.779	20.195
Order	Aleatória Simples	80:20	5	105	80.779	20.195
Order	Aleatória Simples	80:20	5	227	80.779	20.195
Order	Aleatória Simples	80:20	10	0	80.450	20.113
Order	Aleatória Simples	80:20	10	14	80.450	20.113
Order	Aleatória Simples	80:20	10	56	80.450	20.113
Order	Aleatória Simples	80:20	10	84	80.450	20.113
Order	Aleatória Simples	80:20	10	92	80.450	20.113
Order	Aleatória Simples	80:20	10	101	80.450	20.113
Order	Aleatória Simples	80:20	10	105	80.450	20.113
Order	Aleatória Simples	80:20	10	227	80.450	20.113
Order	Aleatória Simples	90:10	5	0	90.877	10.097
Order	Aleatória Simples	90:10	5	14	90.877	10.097
Order	Aleatória Simples	90:10	5	56	90.877	10.097
Order	Aleatória Simples	90:10	5	84	90.877	10.097
Order	Aleatória Simples	90:10	5	92	90.877	10.097
Order	Aleatória Simples	90:10	5	101	90.877	10.097
Order	Aleatória Simples	90:10	5	105	90.877	10.097
Order	Aleatória Simples	90:10	5	227	90.877	10.097
Order	Aleatória Simples	90:10	10	0	90.507	10.056
Order	Aleatória Simples	90:10	10	14	90.507	10.056
Order	Aleatória Simples	90:10	10	56	90.507	10.056
Order	Aleatória Simples	90:10	10	84	90.507	10.056
Order	Aleatória Simples	90:10	10	92	90.507	10.056
Order	Aleatória Simples	90:10	10	101	90.507	10.056
Order	Aleatória Simples	90:10	10	105	90.507	10.056
Order	Aleatória Simples	90:10	10	227	90.507	10.056
Order	Aleatória Simples	95:05	10	0	95.535	5.028
Order	Aleatória Simples	95:05	10	14	95.535	5.028
Order	Aleatória Simples	95:05	10	56	95.535	5.028
Order	Aleatória Simples	95:05	10	84	95.535	5.028
Order	Aleatória Simples	95:05	10	92	95.535	5.028
Order	Aleatória Simples	95:05	10	101	95.535	5.028
Order	Aleatória Simples	95:05	10	105	95.535	5.028
Order	Aleatória Simples	95:05	10	227	95.535	5.028

Tabela A.4: Divisão do conjunto em treino e teste para o nível de Order com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Sequências de Treino	Sequências de Teste
Order	Estratificada	80:20	5	0	80.779	20.195
Order	Estratificada	80:20	5	14	80.779	20.195
Order	Estratificada	80:20	5	56	80.779	20.195
Order	Estratificada	80:20	5	84	80.779	20.195
Order	Estratificada	80:20	5	92	80.779	20.195
Order	Estratificada	80:20	5	101	80.779	20.195
Order	Estratificada	80:20	5	105	80.779	20.195
Order	Estratificada	80:20	5	227	80.779	20.195
Order	Estratificada	80:20	10	0	80.450	20.113
Order	Estratificada	80:20	10	14	80.450	20.113
Order	Estratificada	80:20	10	56	80.450	20.113
Order	Estratificada	80:20	10	84	80.450	20.113
Order	Estratificada	80:20	10	92	80.450	20.113
Order	Estratificada	80:20	10	101	80.450	20.113
Order	Estratificada	80:20	10	105	80.450	20.113
Order	Estratificada	80:20	10	227	80.450	20.113
Order	Estratificada	90:10	5	0	90.876	10.098
Order	Estratificada	90:10	5	14	90.876	10.098
Order	Estratificada	90:10	5	56	90.876	10.098
Order	Estratificada	90:10	5	84	90.876	10.098
Order	Estratificada	90:10	5	92	90.876	10.098
Order	Estratificada	90:10	5	101	90.876	10.098
Order	Estratificada	90:10	5	105	90.876	10.098
Order	Estratificada	90:10	5	227	90.876	10.098
Order	Estratificada	90:10	10	0	90.506	10.057
Order	Estratificada	90:10	10	14	90.506	10.057
Order	Estratificada	90:10	10	56	90.506	10.057
Order	Estratificada	90:10	10	84	90.506	10.057
Order	Estratificada	90:10	10	92	90.506	10.057
Order	Estratificada	90:10	10	101	90.506	10.057
Order	Estratificada	90:10	10	105	90.506	10.057
Order	Estratificada	90:10	10	227	90.506	10.057
Order	Estratificada	95:05	10	0	95.534	5.029
Order	Estratificada	95:05	10	14	95.534	5.029
Order	Estratificada	95:05	10	56	95.534	5.029
Order	Estratificada	95:05	10	84	95.534	5.029
Order	Estratificada	95:05	10	92	95.534	5.029
Order	Estratificada	95:05	10	101	95.534	5.029
Order	Estratificada	95:05	10	105	95.534	5.029
Order	Estratificada	95:05	10	227	95.534	5.029

Tabela A.5: Divisão do conjunto em treino e teste para o nível de Family com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Sequências de Treino	Sequências de Teste
Family	Aleatória Simples	80:20	5	0	66.372	16.593
Family	Aleatória Simples	80:20	5	14	66.372	16.593
Family	Aleatória Simples	80:20	5	56	66.372	16.593
Family	Aleatória Simples	80:20	5	84	66.372	16.593
Family	Aleatória Simples	80:20	5	92	66.372	16.593
Family	Aleatória Simples	80:20	5	101	66.372	16.593
Family	Aleatória Simples	80:20	5	105	66.372	16.593
Family	Aleatória Simples	80:20	5	227	66.372	16.593
Family	Aleatória Simples	80:20	10	0	65.551	16.388
Family	Aleatória Simples	80:20	10	14	65.551	16.388
Family	Aleatória Simples	80:20	10	56	65.551	16.388
Family	Aleatória Simples	80:20	10	84	65.551	16.388
Family	Aleatória Simples	80:20	10	92	65.551	16.388
Family	Aleatória Simples	80:20	10	101	65.551	16.388
Family	Aleatória Simples	80:20	10	105	65.551	16.388
Family	Aleatória Simples	80:20	10	227	65.551	16.388
Family	Aleatória Simples	90:10	5	0	74.669	8.296
Family	Aleatória Simples	90:10	5	14	74.669	8.296
Family	Aleatória Simples	90:10	5	56	74.669	8.296
Family	Aleatória Simples	90:10	5	84	74.669	8.296
Family	Aleatória Simples	90:10	5	92	74.669	8.296
Family	Aleatória Simples	90:10	5	101	74.669	8.296
Family	Aleatória Simples	90:10	5	105	74.669	8.296
Family	Aleatória Simples	90:10	5	227	74.669	8.296
Family	Aleatória Simples	90:10	10	0	73.745	8.194
Family	Aleatória Simples	90:10	10	14	73.745	8.194
Family	Aleatória Simples	90:10	10	56	73.745	8.194
Family	Aleatória Simples	90:10	10	84	73.745	8.194
Family	Aleatória Simples	90:10	10	92	73.745	8.194
Family	Aleatória Simples	90:10	10	101	73.745	8.194
Family	Aleatória Simples	90:10	10	105	73.745	8.194
Family	Aleatória Simples	90:10	10	227	73.745	8.194
Family	Aleatória Simples	95:05	10	0	77.842	4.097
Family	Aleatória Simples	95:05	10	14	77.842	4.097
Family	Aleatória Simples	95:05	10	56	77.842	4.097
Family	Aleatória Simples	95:05	10	84	77.842	4.097
Family	Aleatória Simples	95:05	10	92	77.842	4.097
Family	Aleatória Simples	95:05	10	101	77.842	4.097
Family	Aleatória Simples	95:05	10	105	77.842	4.097
Family	Aleatória Simples	95:05	10	227	77.842	4.097

Tabela A.6: Divisão do conjunto em treino e teste para o nível de Family com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Sequências de Treino	Sequências de Teste
Family	Estratificada	80:20	5	0	66.372	16.593
Family	Estratificada	80:20	5	14	66.372	16.593
Family	Estratificada	80:20	5	56	66.372	16.593
Family	Estratificada	80:20	5	84	66.372	16.593
Family	Estratificada	80:20	5	92	66.372	16.593
Family	Estratificada	80:20	5	101	66.372	16.593
Family	Estratificada	80:20	5	105	66.372	16.593
Family	Estratificada	80:20	5	227	66.372	16.593
Family	Estratificada	80:20	10	0	65.551	16.388
Family	Estratificada	80:20	10	14	65.551	16.388
Family	Estratificada	80:20	10	56	65.551	16.388
Family	Estratificada	80:20	10	84	65.551	16.388
Family	Estratificada	80:20	10	92	65.551	16.388
Family	Estratificada	80:20	10	101	65.551	16.388
Family	Estratificada	80:20	10	105	65.551	16.388
Family	Estratificada	80:20	10	227	65.551	16.388
Family	Estratificada	90:10	5	0	74.668	8.297
Family	Estratificada	90:10	5	14	74.668	8.297
Family	Estratificada	90:10	5	56	74.668	8.297
Family	Estratificada	90:10	5	84	74.668	8.297
Family	Estratificada	90:10	5	92	74.668	8.297
Family	Estratificada	90:10	5	101	74.668	8.297
Family	Estratificada	90:10	5	105	74.668	8.297
Family	Estratificada	90:10	5	227	74.668	8.297
Family	Estratificada	90:10	10	0	73.745	8.194
Family	Estratificada	90:10	10	14	73.745	8.194
Family	Estratificada	90:10	10	56	73.745	8.194
Family	Estratificada	90:10	10	84	73.745	8.194
Family	Estratificada	90:10	10	92	73.745	8.194
Family	Estratificada	90:10	10	101	73.745	8.194
Family	Estratificada	90:10	10	105	73.745	8.194
Family	Estratificada	90:10	10	227	73.745	8.194
Family	Estratificada	95:05	10	0	77.842	4.097
Family	Estratificada	95:05	10	14	77.842	4.097
Family	Estratificada	95:05	10	56	77.842	4.097
Family	Estratificada	95:05	10	84	77.842	4.097
Family	Estratificada	95:05	10	92	77.842	4.097
Family	Estratificada	95:05	10	101	77.842	4.097
Family	Estratificada	95:05	10	105	77.842	4.097
Family	Estratificada	95:05	10	227	77.842	4.097

Tabela A.7: Divisão do conjunto em treino e teste para o nível de Genus com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Sequências de Treino	Sequências de Teste
Genus	Aleatória Simples	80:20	5	0	39.070	9.768
Genus	Aleatória Simples	80:20	5	14	39.070	9.768
Genus	Aleatória Simples	80:20	5	56	39.070	9.768
Genus	Aleatória Simples	80:20	5	84	39.070	9.768
Genus	Aleatória Simples	80:20	5	92	39.070	9.768
Genus	Aleatória Simples	80:20	5	101	39.070	9.768
Genus	Aleatória Simples	80:20	5	105	39.070	9.768
Genus	Aleatória Simples	80:20	5	227	39.070	9.768
Genus	Aleatória Simples	80:20	10	0	33.346	8.337
Genus	Aleatória Simples	80:20	10	14	33.346	8.337
Genus	Aleatória Simples	80:20	10	56	33.346	8.337
Genus	Aleatória Simples	80:20	10	84	33.346	8.337
Genus	Aleatória Simples	80:20	10	92	33.346	8.337
Genus	Aleatória Simples	80:20	10	101	33.346	8.337
Genus	Aleatória Simples	80:20	10	105	33.346	8.337
Genus	Aleatória Simples	80:20	10	227	33.346	8.337
Genus	Aleatória Simples	90:10	5	0	43.954	4.884
Genus	Aleatória Simples	90:10	5	14	43.954	4.884
Genus	Aleatória Simples	90:10	5	56	43.954	4.884
Genus	Aleatória Simples	90:10	5	84	43.954	4.884
Genus	Aleatória Simples	90:10	5	92	43.954	4.884
Genus	Aleatória Simples	90:10	5	101	43.954	4.884
Genus	Aleatória Simples	90:10	5	105	43.954	4.884
Genus	Aleatória Simples	90:10	5	227	43.954	4.884
Genus	Aleatória Simples	90:10	10	0	37.515	4.168
Genus	Aleatória Simples	90:10	10	14	37.515	4.168
Genus	Aleatória Simples	90:10	10	56	37.515	4.168
Genus	Aleatória Simples	90:10	10	84	37.515	4.168
Genus	Aleatória Simples	90:10	10	92	37.515	4.168
Genus	Aleatória Simples	90:10	10	101	37.515	4.168
Genus	Aleatória Simples	90:10	10	105	37.515	4.168
Genus	Aleatória Simples	90:10	10	227	37.515	4.168
Genus	Aleatória Simples	95:05	10	0	39.599	2.084
Genus	Aleatória Simples	95:05	10	14	39.599	2.084
Genus	Aleatória Simples	95:05	10	56	39.599	2.084
Genus	Aleatória Simples	95:05	10	84	39.599	2.084
Genus	Aleatória Simples	95:05	10	92	39.599	2.084
Genus	Aleatória Simples	95:05	10	101	39.599	2.084
Genus	Aleatória Simples	95:05	10	105	39.599	2.084
Genus	Aleatória Simples	95:05	10	227	39.599	2.084

Tabela A.8: Divisão do conjunto em treino e teste para o nível de Genus com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Sequências de Treino	Sequências de Teste
Genus	Estratificada	80:20	5	0	39.070	9.768
Genus	Estratificada	80:20	5	14	39.070	9.768
Genus	Estratificada	80:20	5	56	39.070	9.768
Genus	Estratificada	80:20	5	84	39.070	9.768
Genus	Estratificada	80:20	5	92	39.070	9.768
Genus	Estratificada	80:20	5	101	39.070	9.768
Genus	Estratificada	80:20	5	105	39.070	9.768
Genus	Estratificada	80:20	5	227	39.070	9.768
Genus	Estratificada	80:20	10	0	33.346	8.337
Genus	Estratificada	80:20	10	14	33.346	8.337
Genus	Estratificada	80:20	10	56	33.346	8.337
Genus	Estratificada	80:20	10	84	33.346	8.337
Genus	Estratificada	80:20	10	92	33.346	8.337
Genus	Estratificada	80:20	10	101	33.346	8.337
Genus	Estratificada	80:20	10	105	33.346	8.337
Genus	Estratificada	80:20	10	227	33.346	8.337
Genus	Estratificada	90:10	5	0	43.954	4.884
Genus	Estratificada	90:10	5	14	43.954	4.884
Genus	Estratificada	90:10	5	56	43.954	4.884
Genus	Estratificada	90:10	5	84	43.954	4.884
Genus	Estratificada	90:10	5	92	43.954	4.884
Genus	Estratificada	90:10	5	101	43.954	4.884
Genus	Estratificada	90:10	5	105	43.954	4.884
Genus	Estratificada	90:10	5	227	43.954	4.884
Genus	Estratificada	90:10	10	0	37.514	4.169
Genus	Estratificada	90:10	10	14	37.514	4.169
Genus	Estratificada	90:10	10	56	37.514	4.169
Genus	Estratificada	90:10	10	84	37.514	4.169
Genus	Estratificada	90:10	10	92	37.514	4.169
Genus	Estratificada	90:10	10	101	37.514	4.169
Genus	Estratificada	90:10	10	105	37.514	4.169
Genus	Estratificada	90:10	10	227	37.514	4.169
Genus	Estratificada	95:05	10	0	39.598	2.085
Genus	Estratificada	95:05	10	14	39.598	2.085
Genus	Estratificada	95:05	10	56	39.598	2.085
Genus	Estratificada	95:05	10	84	39.598	2.085
Genus	Estratificada	95:05	10	92	39.598	2.085
Genus	Estratificada	95:05	10	101	39.598	2.085
Genus	Estratificada	95:05	10	105	39.598	2.085
Genus	Estratificada	95:05	10	227	39.598	2.085

Tabela A.9: Divisão do conjunto em treino e teste para o nível de Species com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Sequências de Treino	Sequências de Teste
Species	Aleatória Simples	80:20	5	0	13.521	3.380
Species	Aleatória Simples	80:20	5	14	13.521	3.380
Species	Aleatória Simples	80:20	5	56	13.521	3.380
Species	Aleatória Simples	80:20	5	84	13.521	3.380
Species	Aleatória Simples	80:20	5	92	13.521	3.380
Species	Aleatória Simples	80:20	5	101	13.521	3.380
Species	Aleatória Simples	80:20	5	105	13.521	3.380
Species	Aleatória Simples	80:20	5	227	13.521	3.380
Species	Aleatória Simples	80:20	10	0	9.334	2.333
Species	Aleatória Simples	80:20	10	14	9.334	2.333
Species	Aleatória Simples	80:20	10	56	9.334	2.333
Species	Aleatória Simples	80:20	10	84	9.334	2.333
Species	Aleatória Simples	80:20	10	92	9.334	2.333
Species	Aleatória Simples	80:20	10	101	9.334	2.333
Species	Aleatória Simples	80:20	10	105	9.334	2.333
Species	Aleatória Simples	80:20	10	227	9.334	2.333
Species	Aleatória Simples	90:10	5	0	15.211	1.690
Species	Aleatória Simples	90:10	5	14	15.211	1.690
Species	Aleatória Simples	90:10	5	56	15.211	1.690
Species	Aleatória Simples	90:10	5	84	15.211	1.690
Species	Aleatória Simples	90:10	5	92	15.211	1.690
Species	Aleatória Simples	90:10	5	101	15.211	1.690
Species	Aleatória Simples	90:10	5	105	15.211	1.690
Species	Aleatória Simples	90:10	5	227	15.211	1.690
Species	Aleatória Simples	90:10	10	0	10.500	1.167
Species	Aleatória Simples	90:10	10	14	10.500	1.167
Species	Aleatória Simples	90:10	10	56	10.500	1.167
Species	Aleatória Simples	90:10	10	84	10.500	1.167
Species	Aleatória Simples	90:10	10	92	10.500	1.167
Species	Aleatória Simples	90:10	10	101	10.500	1.167
Species	Aleatória Simples	90:10	10	105	10.500	1.167
Species	Aleatória Simples	90:10	10	227	10.500	1.167
Species	Aleatória Simples	95:05	10	0	11.084	583
Species	Aleatória Simples	95:05	10	14	11.084	583
Species	Aleatória Simples	95:05	10	56	11.084	583
Species	Aleatória Simples	95:05	10	84	11.084	583
Species	Aleatória Simples	95:05	10	92	11.084	583
Species	Aleatória Simples	95:05	10	101	11.084	583
Species	Aleatória Simples	95:05	10	105	11.084	583
Species	Aleatória Simples	95:05	10	227	11.084	583

Tabela A.10: Divisão do conjunto em treino e teste para o nível de Species com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Sequências de Treino	Sequências de Teste
Species	Estratificada	80:20	5	0	13.520	3.381
Species	Estratificada	80:20	5	14	13.520	3.381
Species	Estratificada	80:20	5	56	13.520	3.381
Species	Estratificada	80:20	5	84	13.520	3.381
Species	Estratificada	80:20	5	92	13.520	3.381
Species	Estratificada	80:20	5	101	13.520	3.381
Species	Estratificada	80:20	5	105	13.520	3.381
Species	Estratificada	80:20	5	227	13.520	3.381
Species	Estratificada	80:20	10	0	9.333	2.334
Species	Estratificada	80:20	10	14	9.333	2.334
Species	Estratificada	80:20	10	56	9.333	2.334
Species	Estratificada	80:20	10	84	9.333	2.334
Species	Estratificada	80:20	10	92	9.333	2.334
Species	Estratificada	80:20	10	101	9.333	2.334
Species	Estratificada	80:20	10	105	9.333	2.334
Species	Estratificada	80:20	10	227	9.333	2.334
Species	Estratificada	90:10	5	0	15.210	1.691
Species	Estratificada	90:10	5	14	15.210	1.691
Species	Estratificada	90:10	5	56	15.210	1.691
Species	Estratificada	90:10	5	84	15.210	1.691
Species	Estratificada	90:10	5	92	15.210	1.691
Species	Estratificada	90:10	5	101	15.210	1.691
Species	Estratificada	90:10	5	105	15.210	1.691
Species	Estratificada	90:10	5	227	15.210	1.691
Species	Estratificada	90:10	10	0	10.500	1.167
Species	Estratificada	90:10	10	14	10.500	1.167
Species	Estratificada	90:10	10	56	10.500	1.167
Species	Estratificada	90:10	10	84	10.500	1.167
Species	Estratificada	90:10	10	92	10.500	1.167
Species	Estratificada	90:10	10	101	10.500	1.167
Species	Estratificada	90:10	10	105	10.500	1.167
Species	Estratificada	90:10	10	227	10.500	1.167
Species	Estratificada	95:05	10	0	11.083	584
Species	Estratificada	95:05	10	14	11.083	584
Species	Estratificada	95:05	10	56	11.083	584
Species	Estratificada	95:05	10	84	11.083	584
Species	Estratificada	95:05	10	92	11.083	584
Species	Estratificada	95:05	10	101	11.083	584
Species	Estratificada	95:05	10	105	11.083	584
Species	Estratificada	95:05	10	227	11.083	584

APÊNDICE B – TABELA DE RESULTADOS DO FEATURE CLASSIFIER

Tabela B.1: Resultados dos experimentos com q2-feature-classifier em nível de Class com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Total de Sequências	Acurácia
Class	Aleatória Simples	80:20	5	0	24.721	0,9461
Class	Aleatória Simples	80:20	5	14	24.721	0,9473
Class	Aleatória Simples	80:20	5	56	24.721	0,9483
Class	Aleatória Simples	80:20	5	84	24.721	0,9463
Class	Aleatória Simples	80:20	5	92	24.721	0,9470
Class	Aleatória Simples	80:20	5	101	24.721	0,9449
Class	Aleatória Simples	80:20	5	105	24.721	0,9452
Class	Aleatória Simples	80:20	5	227	24.721	0,9484
Class	Aleatória Simples	80:20	10	0	24.697	0,9455
Class	Aleatória Simples	80:20	10	14	24.697	0,9442
Class	Aleatória Simples	80:20	10	56	24.697	0,9486
Class	Aleatória Simples	80:20	10	84	24.697	0,9476
Class	Aleatória Simples	80:20	10	92	24.697	0,9469
Class	Aleatória Simples	80:20	10	101	24.697	0,9476
Class	Aleatória Simples	80:20	10	105	24.697	0,9467
Class	Aleatória Simples	80:20	10	227	24.697	0,9463
Class	Aleatória Simples	90:10	5	0	12.361	0,9435
Class	Aleatória Simples	90:10	5	14	12.361	0,9484
Class	Aleatória Simples	90:10	5	56	12.361	0,9470
Class	Aleatória Simples	90:10	5	84	12.361	0,9456
Class	Aleatória Simples	90:10	5	92	12.361	0,9489
Class	Aleatória Simples	90:10	5	101	12.361	0,9441
Class	Aleatória Simples	90:10	5	105	12.361	0,9418
Class	Aleatória Simples	90:10	5	227	12.361	0,9490
Class	Aleatória Simples	90:10	10	0	12.349	0,9460
Class	Aleatória Simples	90:10	10	14	12.349	0,9427
Class	Aleatória Simples	90:10	10	56	12.349	0,9469
Class	Aleatória Simples	90:10	10	84	12.349	0,9484
Class	Aleatória Simples	90:10	10	92	12.349	0,9464
Class	Aleatória Simples	90:10	10	101	12.349	0,9482
Class	Aleatória Simples	90:10	10	105	12.349	0,9500
Class	Aleatória Simples	90:10	10	227	12.349	0,9478
Class	Aleatória Simples	95:05	10	0	6.174	0,9423
Class	Aleatória Simples	95:05	10	14	6.174	0,9409
Class	Aleatória Simples	95:05	10	56	6.174	0,9495
Class	Aleatória Simples	95:05	10	84	6.174	0,9498
Class	Aleatória Simples	95:05	10	92	6.174	0,9454
Class	Aleatória Simples	95:05	10	101	6.174	0,9498
Class	Aleatória Simples	95:05	10	105	6.174	0,9482
Class	Aleatória Simples	95:05	10	227	6.174	0,9496

Tabela B.2: Resultados dos experimentos com q2-feature-classifier em nível de Class com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Total de Sequências	Acurácia
Class	Estratificada	80:20	5	0	24.722	0,9489
Class	Estratificada	80:20	5	14	24.722	0,9466
Class	Estratificada	80:20	5	56	24.722	0,9486
Class	Estratificada	80:20	5	84	24.722	0,9511
Class	Estratificada	80:20	5	92	24.722	0,9490
Class	Estratificada	80:20	5	101	24.722	0,9476
Class	Estratificada	80:20	5	105	24.722	0,9456
Class	Estratificada	80:20	5	227	24.722	0,9475
Class	Estratificada	80:20	10	0	24.698	0,9481
Class	Estratificada	80:20	10	14	24.698	0,9459
Class	Estratificada	80:20	10	56	24.698	0,9495
Class	Estratificada	80:20	10	84	24.698	0,9504
Class	Estratificada	80:20	10	92	24.698	0,9487
Class	Estratificada	80:20	10	101	24.698	0,9469
Class	Estratificada	80:20	10	105	24.698	0,9465
Class	Estratificada	80:20	10	227	24.698	0,9479
Class	Estratificada	90:10	5	0	12.361	0,9473
Class	Estratificada	90:10	5	14	12.361	0,9467
Class	Estratificada	90:10	5	56	12.361	0,9483
Class	Estratificada	90:10	5	84	12.361	0,9498
Class	Estratificada	90:10	5	92	12.361	0,9468
Class	Estratificada	90:10	5	101	12.361	0,9466
Class	Estratificada	90:10	5	105	12.361	0,9460
Class	Estratificada	90:10	5	227	12.361	0,9482
Class	Estratificada	90:10	10	0	12.349	0,9466
Class	Estratificada	90:10	10	14	12.349	0,9461
Class	Estratificada	90:10	10	56	12.349	0,9496
Class	Estratificada	90:10	10	84	12.349	0,9491
Class	Estratificada	90:10	10	92	12.349	0,9464
Class	Estratificada	90:10	10	101	12.349	0,9463
Class	Estratificada	90:10	10	105	12.349	0,9452
Class	Estratificada	90:10	10	227	12.349	0,9483
Class	Estratificada	95:05	10	0	6.175	0,9469
Class	Estratificada	95:05	10	14	6.175	0,9453
Class	Estratificada	95:05	10	56	6.175	0,9514
Class	Estratificada	95:05	10	84	6.175	0,9503
Class	Estratificada	95:05	10	92	6.175	0,9480
Class	Estratificada	95:05	10	101	6.175	0,9456
Class	Estratificada	95:05	10	105	6.175	0,9469
Class	Estratificada	95:05	10	227	6.175	0,9475

Tabela B.3: Resultados dos experimentos com q2-feature-classifier em nível de Order com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Total de Sequências	Acurácia
Order	Aleatória Simples	80:20	5	0	20.195	0,9478
Order	Aleatória Simples	80:20	5	14	20.195	0,9476
Order	Aleatória Simples	80:20	5	56	20.195	0,9469
Order	Aleatória Simples	80:20	5	84	20.195	0,9495
Order	Aleatória Simples	80:20	5	92	20.195	0,9475
Order	Aleatória Simples	80:20	5	101	20.195	0,9459
Order	Aleatória Simples	80:20	5	105	20.195	0,9477
Order	Aleatória Simples	80:20	5	227	20.195	0,9476
Order	Aleatória Simples	80:20	10	0	20.113	0,9508
Order	Aleatória Simples	80:20	10	14	20.113	0,9538
Order	Aleatória Simples	80:20	10	56	20.113	0,9477
Order	Aleatória Simples	80:20	10	84	20.113	0,9496
Order	Aleatória Simples	80:20	10	92	20.113	0,9501
Order	Aleatória Simples	80:20	10	101	20.113	0,9481
Order	Aleatória Simples	80:20	10	105	20.113	0,9472
Order	Aleatória Simples	80:20	10	227	20.113	0,9487
Order	Aleatória Simples	90:10	5	0	10.097	0,9475
Order	Aleatória Simples	90:10	5	14	10.097	0,9515
Order	Aleatória Simples	90:10	5	56	10.097	0,9475
Order	Aleatória Simples	90:10	5	84	10.097	0,9524
Order	Aleatória Simples	90:10	5	92	10.097	0,9502
Order	Aleatória Simples	90:10	5	101	10.097	0,9446
Order	Aleatória Simples	90:10	5	105	10.097	0,9502
Order	Aleatória Simples	90:10	5	227	10.097	0,9484
Order	Aleatória Simples	90:10	10	0	10.056	0,9532
Order	Aleatória Simples	90:10	10	14	10.056	0,9544
Order	Aleatória Simples	90:10	10	56	10.056	0,9507
Order	Aleatória Simples	90:10	10	84	10.056	0,9485
Order	Aleatória Simples	90:10	10	92	10.056	0,9471
Order	Aleatória Simples	90:10	10	101	10.056	0,9488
Order	Aleatória Simples	90:10	10	105	10.056	0,9510
Order	Aleatória Simples	90:10	10	227	10.056	0,9519
Order	Aleatória Simples	95:05	10	0	5.028	0,9519
Order	Aleatória Simples	95:05	10	14	5.028	0,9541
Order	Aleatória Simples	95:05	10	56	5.028	0,9491
Order	Aleatória Simples	95:05	10	84	5.028	0,9467
Order	Aleatória Simples	95:05	10	92	5.028	0,9487
Order	Aleatória Simples	95:05	10	101	5.028	0,9461
Order	Aleatória Simples	95:05	10	105	5.028	0,9469
Order	Aleatória Simples	95:05	10	227	5.028	0,9487

Tabela B.4: Resultados dos experimentos com q2-feature-classifier em nível de Order com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Total de Sequências	Acurácia
Order	Estratificada	80:20	5	0	20.195	0,9460
Order	Estratificada	80:20	5	14	20.195	0,9496
Order	Estratificada	80:20	5	56	20.195	0,9477
Order	Estratificada	80:20	5	84	20.195	0,9475
Order	Estratificada	80:20	5	92	20.195	0,9505
Order	Estratificada	80:20	5	101	20.195	0,9452
Order	Estratificada	80:20	5	105	20.195	0,9470
Order	Estratificada	80:20	5	227	20.195	0,9496
Order	Estratificada	80:20	10	0	20.113	0,9479
Order	Estratificada	80:20	10	14	20.113	0,9487
Order	Estratificada	80:20	10	56	20.113	0,9480
Order	Estratificada	80:20	10	84	20.113	0,9487
Order	Estratificada	80:20	10	92	20.113	0,9496
Order	Estratificada	80:20	10	101	20.113	0,9470
Order	Estratificada	80:20	10	105	20.113	0,9492
Order	Estratificada	80:20	10	227	20.113	0,9492
Order	Estratificada	90:10	5	0	10.098	0,9439
Order	Estratificada	90:10	5	14	10.098	0,9485
Order	Estratificada	90:10	5	56	10.098	0,9467
Order	Estratificada	90:10	5	84	10.098	0,9502
Order	Estratificada	90:10	5	92	10.098	0,9453
Order	Estratificada	90:10	5	101	10.098	0,9456
Order	Estratificada	90:10	5	105	10.098	0,9474
Order	Estratificada	90:10	5	227	10.098	0,9487
Order	Estratificada	90:10	10	0	10.057	0,9461
Order	Estratificada	90:10	10	14	10.057	0,9503
Order	Estratificada	90:10	10	56	10.057	0,9467
Order	Estratificada	90:10	10	84	10.057	0,9507
Order	Estratificada	90:10	10	92	10.057	0,9471
Order	Estratificada	90:10	10	101	10.057	0,9470
Order	Estratificada	90:10	10	105	10.057	0,9464
Order	Estratificada	90:10	10	227	10.057	0,9522
Order	Estratificada	95:05	10	0	5.029	0,9459
Order	Estratificada	95:05	10	14	5.029	0,9499
Order	Estratificada	95:05	10	56	5.029	0,9475
Order	Estratificada	95:05	10	84	5.029	0,9483
Order	Estratificada	95:05	10	92	5.029	0,9469
Order	Estratificada	95:05	10	101	5.029	0,9483
Order	Estratificada	95:05	10	105	5.029	0,9493
Order	Estratificada	95:05	10	227	5.029	0,9547

Tabela B.5: Resultados dos experimentos com q2-feature-classifier em nível de Family com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Total de Sequências	Acurácia
Family	Aleatória Simples	80:20	5	0	16.593	0,9396
Family	Aleatória Simples	80:20	5	14	16.593	0,9411
Family	Aleatória Simples	80:20	5	56	16.593	0,9394
Family	Aleatória Simples	80:20	5	84	16.593	0,9418
Family	Aleatória Simples	80:20	5	92	16.593	0,9420
Family	Aleatória Simples	80:20	5	101	16.593	0,9390
Family	Aleatória Simples	80:20	5	105	16.593	0,9417
Family	Aleatória Simples	80:20	5	227	16.593	0,9398
Family	Aleatória Simples	80:20	10	0	16.388	0,9457
Family	Aleatória Simples	80:20	10	14	16.388	0,9428
Family	Aleatória Simples	80:20	10	56	16.388	0,9400
Family	Aleatória Simples	80:20	10	84	16.388	0,9440
Family	Aleatória Simples	80:20	10	92	16.388	0,9447
Family	Aleatória Simples	80:20	10	101	16.388	0,9410
Family	Aleatória Simples	80:20	10	105	16.388	0,9427
Family	Aleatória Simples	80:20	10	227	16.388	0,9425
Family	Aleatória Simples	90:10	5	0	8.296	0,9413
Family	Aleatória Simples	90:10	5	14	8.296	0,9402
Family	Aleatória Simples	90:10	5	56	8.296	0,9384
Family	Aleatória Simples	90:10	5	84	8.296	0,9408
Family	Aleatória Simples	90:10	5	92	8.296	0,9378
Family	Aleatória Simples	90:10	5	101	8.296	0,9368
Family	Aleatória Simples	90:10	5	105	8.296	0,9412
Family	Aleatória Simples	90:10	5	227	8.296	0,9388
Family	Aleatória Simples	90:10	10	0	8.194	0,9462
Family	Aleatória Simples	90:10	10	14	8.194	0,9433
Family	Aleatória Simples	90:10	10	56	8.194	0,9387
Family	Aleatória Simples	90:10	10	84	8.194	0,9431
Family	Aleatória Simples	90:10	10	92	8.194	0,9436
Family	Aleatória Simples	90:10	10	101	8.194	0,9412
Family	Aleatória Simples	90:10	10	105	8.194	0,9408
Family	Aleatória Simples	90:10	10	227	8.194	0,9470
Family	Aleatória Simples	95:05	10	0	4.097	0,9446
Family	Aleatória Simples	95:05	10	14	4.097	0,9404
Family	Aleatória Simples	95:05	10	56	4.097	0,9436
Family	Aleatória Simples	95:05	10	84	4.097	0,9441
Family	Aleatória Simples	95:05	10	92	4.097	0,9439
Family	Aleatória Simples	95:05	10	101	4.097	0,9424
Family	Aleatória Simples	95:05	10	105	4.097	0,9426
Family	Aleatória Simples	95:05	10	227	4.097	0,9487

Tabela B.6: Resultados dos experimentos com q2-feature-classifier em nível de Family com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Total de Sequências	Acurácia
Family	Estratificada	80:20	5	0	16.593	0,9415
Family	Estratificada	80:20	5	14	16.593	0,9391
Family	Estratificada	80:20	5	56	16.593	0,9409
Family	Estratificada	80:20	5	84	16.593	0,9426
Family	Estratificada	80:20	5	92	16.593	0,9402
Family	Estratificada	80:20	5	101	16.593	0,9372
Family	Estratificada	80:20	5	105	16.593	0,9385
Family	Estratificada	80:20	5	227	16.593	0,9420
Family	Estratificada	80:20	10	0	16.388	0,9460
Family	Estratificada	80:20	10	14	16.388	0,9436
Family	Estratificada	80:20	10	56	16.388	0,9414
Family	Estratificada	80:20	10	84	16.388	0,9439
Family	Estratificada	80:20	10	92	16.388	0,9428
Family	Estratificada	80:20	10	101	16.388	0,9450
Family	Estratificada	80:20	10	105	16.388	0,9423
Family	Estratificada	80:20	10	227	16.388	0,9429
Family	Estratificada	90:10	5	0	8.297	0,9407
Family	Estratificada	90:10	5	14	8.297	0,9383
Family	Estratificada	90:10	5	56	8.297	0,9408
Family	Estratificada	90:10	5	84	8.297	0,9419
Family	Estratificada	90:10	5	92	8.297	0,9418
Family	Estratificada	90:10	5	101	8.297	0,9400
Family	Estratificada	90:10	5	105	8.297	0,9424
Family	Estratificada	90:10	5	227	8.297	0,9406
Family	Estratificada	90:10	10	0	8.194	0,9470
Family	Estratificada	90:10	10	14	8.194	0,9440
Family	Estratificada	90:10	10	56	8.194	0,9424
Family	Estratificada	90:10	10	84	8.194	0,9439
Family	Estratificada	90:10	10	92	8.194	0,9464
Family	Estratificada	90:10	10	101	8.194	0,9454
Family	Estratificada	90:10	10	105	8.194	0,9412
Family	Estratificada	90:10	10	227	8.194	0,9430
Family	Estratificada	95:05	10	0	4.097	0,9475
Family	Estratificada	95:05	10	14	4.097	0,9424
Family	Estratificada	95:05	10	56	4.097	0,9426
Family	Estratificada	95:05	10	84	4.097	0,9400
Family	Estratificada	95:05	10	92	4.097	0,9395
Family	Estratificada	95:05	10	101	4.097	0,9426
Family	Estratificada	95:05	10	105	4.097	0,9426
Family	Estratificada	95:05	10	227	4.097	0,9439

Tabela B.7: Resultados dos experimentos com q2-feature-classifier em nível de Genus com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Total de Sequências	Acurácia
Genus	Aleatória Simples	80:20	5	0	9.768	0,7952
Genus	Aleatória Simples	80:20	5	14	9.768	0,7851
Genus	Aleatória Simples	80:20	5	56	9.768	0,7994
Genus	Aleatória Simples	80:20	5	84	9.768	0,7951
Genus	Aleatória Simples	80:20	5	92	9.768	0,7955
Genus	Aleatória Simples	80:20	5	101	9.768	0,7894
Genus	Aleatória Simples	80:20	5	105	9.768	0,7900
Genus	Aleatória Simples	80:20	5	227	9.768	0,7938
Genus	Aleatória Simples	80:20	10	0	8.337	0,8454
Genus	Aleatória Simples	80:20	10	14	8.337	0,8363
Genus	Aleatória Simples	80:20	10	56	8.337	0,8406
Genus	Aleatória Simples	80:20	10	84	8.337	0,8427
Genus	Aleatória Simples	80:20	10	92	8.337	0,8411
Genus	Aleatória Simples	80:20	10	101	8.337	0,8323
Genus	Aleatória Simples	80:20	10	105	8.337	0,8420
Genus	Aleatória Simples	80:20	10	227	8.337	0,8402
Genus	Aleatória Simples	90:10	5	0	4.884	0,7961
Genus	Aleatória Simples	90:10	5	14	4.884	0,7895
Genus	Aleatória Simples	90:10	5	56	4.884	0,8075
Genus	Aleatória Simples	90:10	5	84	4.884	0,7948
Genus	Aleatória Simples	90:10	5	92	4.884	0,8045
Genus	Aleatória Simples	90:10	5	101	4.884	0,7961
Genus	Aleatória Simples	90:10	5	105	4.884	0,7957
Genus	Aleatória Simples	90:10	5	227	4.884	0,7938
Genus	Aleatória Simples	90:10	10	0	4.168	0,8512
Genus	Aleatória Simples	90:10	10	14	4.168	0,8376
Genus	Aleatória Simples	90:10	10	56	4.168	0,8388
Genus	Aleatória Simples	90:10	10	84	4.168	0,8359
Genus	Aleatória Simples	90:10	10	92	4.168	0,8491
Genus	Aleatória Simples	90:10	10	101	4.168	0,8417
Genus	Aleatória Simples	90:10	10	105	4.168	0,8491
Genus	Aleatória Simples	90:10	10	227	4.168	0,8385
Genus	Aleatória Simples	95:05	10	0	2.084	0,8484
Genus	Aleatória Simples	95:05	10	14	2.084	0,8373
Genus	Aleatória Simples	95:05	10	56	2.084	0,8508
Genus	Aleatória Simples	95:05	10	84	2.084	0,8397
Genus	Aleatória Simples	95:05	10	92	2.084	0,8488
Genus	Aleatória Simples	95:05	10	101	2.084	0,8426
Genus	Aleatória Simples	95:05	10	105	2.084	0,8484
Genus	Aleatória Simples	95:05	10	227	2.084	0,8402

Tabela B.8: Resultados dos experimentos com q2-feature-classifier em nível de Genus com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Total de Sequências	Acurácia
Genus	Estratificada	80:20	5	0	9.768	0,7930
Genus	Estratificada	80:20	5	14	9.768	0,7989
Genus	Estratificada	80:20	5	56	9.768	0,8026
Genus	Estratificada	80:20	5	84	9.768	0,8003
Genus	Estratificada	80:20	5	92	9.768	0,7967
Genus	Estratificada	80:20	5	101	9.768	0,7962
Genus	Estratificada	80:20	5	105	9.768	0,8063
Genus	Estratificada	80:20	5	227	9.768	0,7971
Genus	Estratificada	80:20	10	0	8.337	0,8389
Genus	Estratificada	80:20	10	14	8.337	0,8369
Genus	Estratificada	80:20	10	56	8.337	0,8456
Genus	Estratificada	80:20	10	84	8.337	0,8389
Genus	Estratificada	80:20	10	92	8.337	0,8344
Genus	Estratificada	80:20	10	101	8.337	0,8383
Genus	Estratificada	80:20	10	105	8.337	0,8431
Genus	Estratificada	80:20	10	227	8.337	0,8395
Genus	Estratificada	90:10	5	0	4.884	0,8036
Genus	Estratificada	90:10	5	14	4.884	0,8018
Genus	Estratificada	90:10	5	56	4.884	0,8036
Genus	Estratificada	90:10	5	84	4.884	0,8059
Genus	Estratificada	90:10	5	92	4.884	0,7981
Genus	Estratificada	90:10	5	101	4.884	0,8124
Genus	Estratificada	90:10	5	105	4.884	0,8032
Genus	Estratificada	90:10	5	227	4.884	0,7955
Genus	Estratificada	90:10	10	0	4.169	0,8434
Genus	Estratificada	90:10	10	14	4.169	0,8388
Genus	Estratificada	90:10	10	56	4.169	0,8405
Genus	Estratificada	90:10	10	84	4.169	0,8472
Genus	Estratificada	90:10	10	92	4.169	0,8338
Genus	Estratificada	90:10	10	101	4.169	0,8412
Genus	Estratificada	90:10	10	105	4.169	0,8496
Genus	Estratificada	90:10	10	227	4.169	0,8417
Genus	Estratificada	95:05	10	0	2.085	0,8412
Genus	Estratificada	95:05	10	14	2.085	0,8436
Genus	Estratificada	95:05	10	56	2.085	0,8460
Genus	Estratificada	95:05	10	84	2.085	0,8341
Genus	Estratificada	95:05	10	92	2.085	0,8398
Genus	Estratificada	95:05	10	101	2.085	0,8494
Genus	Estratificada	95:05	10	105	2.085	0,8580
Genus	Estratificada	95:05	10	227	2.085	0,8398

Tabela B.9: Resultados dos experimentos com q2-feature-classifier em nível de Species com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Total de Sequências	Acurácia
Species	Aleatória Simples	80:20	5	0	3.380	0,7973
Species	Aleatória Simples	80:20	5	14	3.380	0,7976
Species	Aleatória Simples	80:20	5	56	3.380	0,8065
Species	Aleatória Simples	80:20	5	84	3.380	0,7991
Species	Aleatória Simples	80:20	5	92	3.380	0,8018
Species	Aleatória Simples	80:20	5	101	3.380	0,8083
Species	Aleatória Simples	80:20	5	105	3.380	0,8077
Species	Aleatória Simples	80:20	5	227	3.380	0,8044
Species	Aleatória Simples	80:20	10	0	2.333	0,0021
Species	Aleatória Simples	80:20	10	14	2.333	0,8804
Species	Aleatória Simples	80:20	10	56	2.333	0,0009
Species	Aleatória Simples	80:20	10	84	2.333	0,8714
Species	Aleatória Simples	80:20	10	92	2.333	0,8864
Species	Aleatória Simples	80:20	10	101	2.333	0,8821
Species	Aleatória Simples	80:20	10	105	2.333	0,0004
Species	Aleatória Simples	80:20	10	227	2.333	0,0017
Species	Aleatória Simples	90:10	5	0	1.690	0,8012
Species	Aleatória Simples	90:10	5	14	1.690	0,0000
Species	Aleatória Simples	90:10	5	56	1.690	0,8160
Species	Aleatória Simples	90:10	5	84	1.690	0,0006
Species	Aleatória Simples	90:10	5	92	1.690	0,0030
Species	Aleatória Simples	90:10	5	101	1.690	0,0006
Species	Aleatória Simples	90:10	5	105	1.690	0,0012
Species	Aleatória Simples	90:10	5	227	1.690	0,0030
Species	Aleatória Simples	90:10	10	0	1.167	0,0000
Species	Aleatória Simples	90:10	10	14	1.167	0,8757
Species	Aleatória Simples	90:10	10	56	1.167	0,0017
Species	Aleatória Simples	90:10	10	84	1.167	0,0000
Species	Aleatória Simples	90:10	10	92	1.167	0,0009
Species	Aleatória Simples	90:10	10	101	1.167	0,8929
Species	Aleatória Simples	90:10	10	105	1.167	0,0009
Species	Aleatória Simples	90:10	10	227	1.167	0,0026
Species	Aleatória Simples	95:05	10	0	583	0,0000
Species	Aleatória Simples	95:05	10	14	583	0,0000
Species	Aleatória Simples	95:05	10	56	583	0,0017
Species	Aleatória Simples	95:05	10	84	583	0,0000
Species	Aleatória Simples	95:05	10	92	583	0,0017
Species	Aleatória Simples	95:05	10	101	583	0,8971
Species	Aleatória Simples	95:05	10	105	583	0,0000
Species	Aleatória Simples	95:05	10	227	583	0,0034

Tabela B.10: Resultados dos experimentos com q2-feature-classifier em nível de Species com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Total de Sequências	Acurácia
Species	Estratificada	80:20	5	0	3.381	0,0012
Species	Estratificada	80:20	5	14	3.381	0,8089
Species	Estratificada	80:20	5	56	3.381	0,0015
Species	Estratificada	80:20	5	84	3.381	0,8110
Species	Estratificada	80:20	5	92	3.381	0,0018
Species	Estratificada	80:20	5	101	3.381	0,0015
Species	Estratificada	80:20	5	105	3.381	0,0015
Species	Estratificada	80:20	5	227	3.381	0,0012
Species	Estratificada	80:20	10	0	2.334	0,0009
Species	Estratificada	80:20	10	14	2.334	0,8805
Species	Estratificada	80:20	10	56	2.334	0,8830
Species	Estratificada	80:20	10	84	2.334	0,8800
Species	Estratificada	80:20	10	92	2.334	0,0009
Species	Estratificada	80:20	10	101	2.334	0,8757
Species	Estratificada	80:20	10	105	2.334	0,8839
Species	Estratificada	80:20	10	227	2.334	0,8779
Species	Estratificada	90:10	5	0	1.691	0,0012
Species	Estratificada	90:10	5	14	1.691	0,8084
Species	Estratificada	90:10	5	56	1.691	0,0012
Species	Estratificada	90:10	5	84	1.691	0,0012
Species	Estratificada	90:10	5	92	1.691	0,0012
Species	Estratificada	90:10	5	101	1.691	0,0012
Species	Estratificada	90:10	5	105	1.691	0,0012
Species	Estratificada	90:10	5	227	1.691	0,0012
Species	Estratificada	90:10	10	0	1.167	0,0009
Species	Estratificada	90:10	10	14	1.167	0,8775
Species	Estratificada	90:10	10	56	1.167	0,8946
Species	Estratificada	90:10	10	84	1.167	0,0009
Species	Estratificada	90:10	10	92	1.167	0,0009
Species	Estratificada	90:10	10	101	1.167	0,0009
Species	Estratificada	90:10	10	105	1.167	0,0009
Species	Estratificada	90:10	10	227	1.167	0,8835
Species	Estratificada	95:05	10	0	584	0,0017
Species	Estratificada	95:05	10	14	584	0,0017
Species	Estratificada	95:05	10	56	584	0,0017
Species	Estratificada	95:05	10	84	584	0,0017
Species	Estratificada	95:05	10	92	584	0,0017
Species	Estratificada	95:05	10	101	584	0,8630
Species	Estratificada	95:05	10	105	584	0,8887
Species	Estratificada	95:05	10	227	584	0,0017

APÊNDICE C – TABELA DE RESULTADOS DA SOLUÇÃO DESENVOLVIDA

Tabela C.1: Resultados dos experimentos com a solução desenvolvida em nível de Class com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Épocas	Melhor Época	Acurácia
Class	Aleatória Simples	80:20	5	0	700	306	0,9898
Class	Aleatória Simples	80:20	5	14	700	303	0,9892
Class	Aleatória Simples	80:20	5	56	700	298	0,9894
Class	Aleatória Simples	80:20	5	84	700	297	0,9909
Class	Aleatória Simples	80:20	5	92	700	305	0,9905
Class	Aleatória Simples	80:20	5	101	700	303	0,9898
Class	Aleatória Simples	80:20	5	105	700	307	0,9895
Class	Aleatória Simples	80:20	5	227	700	305	0,9897
Class	Aleatória Simples	80:20	10	0	700	304	0,9894
Class	Aleatória Simples	80:20	10	14	700	304	0,9894
Class	Aleatória Simples	80:20	10	56	700	303	0,9892
Class	Aleatória Simples	80:20	10	84	700	302	0,9876
Class	Aleatória Simples	80:20	10	92	700	310	0,9888
Class	Aleatória Simples	80:20	10	101	700	310	0,9902
Class	Aleatória Simples	80:20	10	105	700	307	0,9894
Class	Aleatória Simples	80:20	10	227	700	297	0,9898
Class	Aleatória Simples	90:10	5	0	700	301	0,9892
Class	Aleatória Simples	90:10	5	14	700	309	0,9897
Class	Aleatória Simples	90:10	5	56	700	305	0,9905
Class	Aleatória Simples	90:10	5	84	700	309	0,9921
Class	Aleatória Simples	90:10	5	92	700	299	0,9915
Class	Aleatória Simples	90:10	5	101	700	308	0,9898
Class	Aleatória Simples	90:10	5	105	700	298	0,9900
Class	Aleatória Simples	90:10	5	227	700	298	0,9887
Class	Aleatória Simples	90:10	10	0	700	307	0,9907
Class	Aleatória Simples	90:10	10	14	700	147	0,9896
Class	Aleatória Simples	90:10	10	56	700	305	0,9908
Class	Aleatória Simples	90:10	10	84	700	306	0,9885
Class	Aleatória Simples	90:10	10	92	700	308	0,9908
Class	Aleatória Simples	90:10	10	101	700	310	0,9906
Class	Aleatória Simples	90:10	10	105	700	296	0,9908
Class	Aleatória Simples	90:10	10	227	700	302	0,9895
Class	Aleatória Simples	95:05	10	0	700	307	0,9914
Class	Aleatória Simples	95:05	10	14	700	298	0,9903
Class	Aleatória Simples	95:05	10	56	700	302	0,9903
Class	Aleatória Simples	95:05	10	84	700	150	0,9903
Class	Aleatória Simples	95:05	10	92	700	300	0,9891
Class	Aleatória Simples	95:05	10	101	700	295	0,9913
Class	Aleatória Simples	95:05	10	105	700	144	0,9901
Class	Aleatória Simples	95:05	10	227	700	299	0,9906

Tabela C.2: Resultados dos experimentos com a solução desenvolvida em nível de Class com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Épocas	Melhor Época	Acurácia
Class	Estratificada	80:20	5	0	700	300	0,9901
Class	Estratificada	80:20	5	14	700	306	0,9910
Class	Estratificada	80:20	5	56	700	310	0,9897
Class	Estratificada	80:20	5	84	700	306	0,9906
Class	Estratificada	80:20	5	92	700	304	0,9904
Class	Estratificada	80:20	5	101	700	302	0,9892
Class	Estratificada	80:20	5	105	700	306	0,9890
Class	Estratificada	80:20	5	227	700	303	0,9904
Class	Estratificada	80:20	10	0	700	306	0,9909
Class	Estratificada	80:20	10	14	700	302	0,9897
Class	Estratificada	80:20	10	56	700	300	0,9899
Class	Estratificada	80:20	10	84	700	299	0,9908
Class	Estratificada	80:20	10	92	700	303	0,9906
Class	Estratificada	80:20	10	101	700	308	0,9909
Class	Estratificada	80:20	10	105	700	303	0,9890
Class	Estratificada	80:20	10	227	700	303	0,9898
Class	Estratificada	90:10	5	0	700	303	0,9908
Class	Estratificada	90:10	5	14	700	307	0,9920
Class	Estratificada	90:10	5	56	700	306	0,9912
Class	Estratificada	90:10	5	84	700	305	0,9919
Class	Estratificada	90:10	5	92	700	304	0,9909
Class	Estratificada	90:10	5	101	700	303	0,9903
Class	Estratificada	90:10	5	105	700	309	0,9884
Class	Estratificada	90:10	5	227	700	305	0,9898
Class	Estratificada	90:10	10	0	700	304	0,9908
Class	Estratificada	90:10	10	14	700	306	0,9909
Class	Estratificada	90:10	10	56	700	299	0,9914
Class	Estratificada	90:10	10	84	700	299	0,9901
Class	Estratificada	90:10	10	92	700	304	0,9903
Class	Estratificada	90:10	10	101	700	306	0,9921
Class	Estratificada	90:10	10	105	700	299	0,9892
Class	Estratificada	90:10	10	227	700	148	0,9899
Class	Estratificada	95:05	10	0	700	303	0,9922
Class	Estratificada	95:05	10	14	700	297	0,9917
Class	Estratificada	95:05	10	56	700	300	0,9911
Class	Estratificada	95:05	10	84	700	145	0,9921
Class	Estratificada	95:05	10	92	700	304	0,9906
Class	Estratificada	95:05	10	101	700	299	0,9893
Class	Estratificada	95:05	10	105	700	298	0,9891
Class	Estratificada	95:05	10	227	700	304	0,9911

Tabela C.3: Resultados dos experimentos com a solução desenvolvida em nível de Order com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Épocas	Melhor Época	Acurácia
Order	Aleatória Simples	80:20	5	0	700	310	0,9826
Order	Aleatória Simples	80:20	5	14	700	309	0,9828
Order	Aleatória Simples	80:20	5	56	700	306	0,9855
Order	Aleatória Simples	80:20	5	84	700	303	0,9844
Order	Aleatória Simples	80:20	5	92	700	308	0,9844
Order	Aleatória Simples	80:20	5	101	700	301	0,9824
Order	Aleatória Simples	80:20	5	105	700	310	0,9834
Order	Aleatória Simples	80:20	5	227	700	307	0,9824
Order	Aleatória Simples	80:20	10	0	700	304	0,9870
Order	Aleatória Simples	80:20	10	14	700	305	0,9854
Order	Aleatória Simples	80:20	10	56	700	306	0,9853
Order	Aleatória Simples	80:20	10	84	700	305	0,9837
Order	Aleatória Simples	80:20	10	92	700	307	0,9848
Order	Aleatória Simples	80:20	10	101	700	307	0,9863
Order	Aleatória Simples	80:20	10	105	700	300	0,9832
Order	Aleatória Simples	80:20	10	227	700	304	0,9844
Order	Aleatória Simples	90:10	5	0	700	309	0,9839
Order	Aleatória Simples	90:10	5	14	700	307	0,9863
Order	Aleatória Simples	90:10	5	56	700	306	0,9851
Order	Aleatória Simples	90:10	5	84	700	302	0,9850
Order	Aleatória Simples	90:10	5	92	700	298	0,9856
Order	Aleatória Simples	90:10	5	101	700	309	0,9844
Order	Aleatória Simples	90:10	5	105	700	308	0,9818
Order	Aleatória Simples	90:10	5	227	700	302	0,9848
Order	Aleatória Simples	90:10	10	0	700	306	0,9867
Order	Aleatória Simples	90:10	10	14	700	299	0,9861
Order	Aleatória Simples	90:10	10	56	700	304	0,9872
Order	Aleatória Simples	90:10	10	84	700	302	0,9862
Order	Aleatória Simples	90:10	10	92	700	295	0,9860
Order	Aleatória Simples	90:10	10	101	700	308	0,9875
Order	Aleatória Simples	90:10	10	105	700	305	0,9853
Order	Aleatória Simples	90:10	10	227	700	303	0,9880
Order	Aleatória Simples	95:05	10	0	700	296	0,9859
Order	Aleatória Simples	95:05	10	14	700	304	0,9873
Order	Aleatória Simples	95:05	10	56	700	300	0,9839
Order	Aleatória Simples	95:05	10	84	700	302	0,9841
Order	Aleatória Simples	95:05	10	92	700	301	0,9873
Order	Aleatória Simples	95:05	10	101	700	308	0,9879
Order	Aleatória Simples	95:05	10	105	700	309	0,9867
Order	Aleatória Simples	95:05	10	227	700	304	0,9857

Tabela C.4: Resultados dos experimentos com a solução desenvolvida em nível de Order com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Épocas	Melhor Época	Acurácia
Order	Estratificada	80:20	5	0	700	303	0,9846
Order	Estratificada	80:20	5	14	700	303	0,9824
Order	Estratificada	80:20	5	56	700	308	0,9863
Order	Estratificada	80:20	5	84	700	310	0,9847
Order	Estratificada	80:20	5	92	700	308	0,9851
Order	Estratificada	80:20	5	101	700	309	0,9816
Order	Estratificada	80:20	5	105	700	308	0,9852
Order	Estratificada	80:20	5	227	700	306	0,9827
Order	Estratificada	80:20	10	0	700	301	0,9849
Order	Estratificada	80:20	10	14	700	306	0,9776
Order	Estratificada	80:20	10	56	700	310	0,9832
Order	Estratificada	80:20	10	84	700	303	0,9841
Order	Estratificada	80:20	10	92	700	307	0,9837
Order	Estratificada	80:20	10	101	700	292	0,9859
Order	Estratificada	80:20	10	105	700	310	0,9854
Order	Estratificada	80:20	10	227	700	308	0,9833
Order	Estratificada	90:10	5	0	700	309	0,9862
Order	Estratificada	90:10	5	14	700	309	0,9851
Order	Estratificada	90:10	5	56	700	303	0,9864
Order	Estratificada	90:10	5	84	700	301	0,9870
Order	Estratificada	90:10	5	92	700	310	0,9863
Order	Estratificada	90:10	5	101	700	309	0,9864
Order	Estratificada	90:10	5	105	700	305	0,9834
Order	Estratificada	90:10	5	227	700	304	0,9851
Order	Estratificada	90:10	10	0	700	305	0,9863
Order	Estratificada	90:10	10	14	700	304	0,9882
Order	Estratificada	90:10	10	56	700	310	0,9874
Order	Estratificada	90:10	10	84	700	623	0,9833
Order	Estratificada	90:10	10	92	700	302	0,9861
Order	Estratificada	90:10	10	101	700	299	0,9852
Order	Estratificada	90:10	10	105	700	302	0,9862
Order	Estratificada	90:10	10	227	700	300	0,9877
Order	Estratificada	95:05	10	0	700	310	0,9853
Order	Estratificada	95:05	10	14	700	299	0,9881
Order	Estratificada	95:05	10	56	700	305	0,9867
Order	Estratificada	95:05	10	84	700	302	0,9867
Order	Estratificada	95:05	10	92	700	306	0,9865
Order	Estratificada	95:05	10	101	700	307	0,9857
Order	Estratificada	95:05	10	105	700	297	0,9881
Order	Estratificada	95:05	10	227	700	304	0,9907

Tabela C.5: Resultados dos experimentos com a solução desenvolvida em nível de Family com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Épocas	Melhor Época	Acurácia
Family	Aleatória Simples	80:20	5	0	700	305	0,9722
Family	Aleatória Simples	80:20	5	14	700	307	0,9721
Family	Aleatória Simples	80:20	5	56	700	303	0,9714
Family	Aleatória Simples	80:20	5	84	700	302	0,9722
Family	Aleatória Simples	80:20	5	92	700	302	0,9749
Family	Aleatória Simples	80:20	5	101	700	306	0,9710
Family	Aleatória Simples	80:20	5	105	700	307	0,9735
Family	Aleatória Simples	80:20	5	227	700	310	0,9708
Family	Aleatória Simples	80:20	10	0	700	309	0,9766
Family	Aleatória Simples	80:20	10	14	700	299	0,9764
Family	Aleatória Simples	80:20	10	56	700	310	0,9750
Family	Aleatória Simples	80:20	10	84	700	300	0,9779
Family	Aleatória Simples	80:20	10	92	700	303	0,9778
Family	Aleatória Simples	80:20	10	101	700	305	0,9738
Family	Aleatória Simples	80:20	10	105	700	305	0,9742
Family	Aleatória Simples	80:20	10	227	700	303	0,9771
Family	Aleatória Simples	90:10	5	0	700	305	0,9767
Family	Aleatória Simples	90:10	5	14	700	305	0,9740
Family	Aleatória Simples	90:10	5	56	700	304	0,9726
Family	Aleatória Simples	90:10	5	84	700	309	0,9729
Family	Aleatória Simples	90:10	5	92	700	309	0,9749
Family	Aleatória Simples	90:10	5	101	700	307	0,9729
Family	Aleatória Simples	90:10	5	105	700	305	0,9722
Family	Aleatória Simples	90:10	5	227	700	304	0,9728
Family	Aleatória Simples	90:10	10	0	700	303	0,9779
Family	Aleatória Simples	90:10	10	14	700	629	0,9769
Family	Aleatória Simples	90:10	10	56	700	302	0,9758
Family	Aleatória Simples	90:10	10	84	700	306	0,9808
Family	Aleatória Simples	90:10	10	92	700	310	0,9773
Family	Aleatória Simples	90:10	10	101	700	299	0,9760
Family	Aleatória Simples	90:10	10	105	700	308	0,9766
Family	Aleatória Simples	90:10	10	227	700	302	0,9800
Family	Aleatória Simples	95:05	10	0	700	299	0,9753
Family	Aleatória Simples	95:05	10	14	700	309	0,9802
Family	Aleatória Simples	95:05	10	56	700	303	0,9785
Family	Aleatória Simples	95:05	10	84	700	298	0,9807
Family	Aleatória Simples	95:05	10	92	700	305	0,9763
Family	Aleatória Simples	95:05	10	101	700	300	0,9780
Family	Aleatória Simples	95:05	10	105	700	308	0,9771
Family	Aleatória Simples	95:05	10	227	700	298	0,9810

Tabela C.6: Resultados dos experimentos com a solução desenvolvida em nível de Family com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Épocas	Melhor Época	Acurácia
Family	Estratificada	80:20	5	0	700	304	0,9773
Family	Estratificada	80:20	5	14	700	302	0,9727
Family	Estratificada	80:20	5	56	700	310	0,9744
Family	Estratificada	80:20	5	84	700	303	0,9729
Family	Estratificada	80:20	5	92	700	308	0,9738
Family	Estratificada	80:20	5	101	700	310	0,9733
Family	Estratificada	80:20	5	105	700	309	0,9734
Family	Estratificada	80:20	5	227	700	310	0,9710
Family	Estratificada	80:20	10	0	700	308	0,9760
Family	Estratificada	80:20	10	14	700	308	0,9746
Family	Estratificada	80:20	10	56	700	302	0,9732
Family	Estratificada	80:20	10	84	700	303	0,9752
Family	Estratificada	80:20	10	92	700	299	0,9741
Family	Estratificada	80:20	10	101	700	309	0,9749
Family	Estratificada	80:20	10	105	700	304	0,9763
Family	Estratificada	80:20	10	227	700	306	0,9752
Family	Estratificada	90:10	5	0	700	304	0,9751
Family	Estratificada	90:10	5	14	700	303	0,9741
Family	Estratificada	90:10	5	56	700	306	0,9755
Family	Estratificada	90:10	5	84	700	308	0,9759
Family	Estratificada	90:10	5	92	700	308	0,9753
Family	Estratificada	90:10	5	101	700	307	0,9748
Family	Estratificada	90:10	5	105	700	307	0,9781
Family	Estratificada	90:10	5	227	700	307	0,9748
Family	Estratificada	90:10	10	0	700	310	0,9796
Family	Estratificada	90:10	10	14	700	308	0,9755
Family	Estratificada	90:10	10	56	700	308	0,9769
Family	Estratificada	90:10	10	84	700	299	0,9771
Family	Estratificada	90:10	10	92	700	304	0,9729
Family	Estratificada	90:10	10	101	700	302	0,9802
Family	Estratificada	90:10	10	105	700	307	0,9774
Family	Estratificada	90:10	10	227	700	303	0,9782
Family	Estratificada	95:05	10	0	700	304	0,9800
Family	Estratificada	95:05	10	14	700	297	0,9766
Family	Estratificada	95:05	10	56	700	302	0,9761
Family	Estratificada	95:05	10	84	700	294	0,9802
Family	Estratificada	95:05	10	92	700	625	0,9766
Family	Estratificada	95:05	10	101	700	630	0,9780
Family	Estratificada	95:05	10	105	700	628	0,9788
Family	Estratificada	95:05	10	227	700	295	0,9773

Tabela C.7: Resultados dos experimentos com a solução desenvolvida em nível de Genus com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Épocas	Melhor Época	Acurácia
Genus	Aleatória Simples	80:20	5	0	700	630	0,8552
Genus	Aleatória Simples	80:20	5	14	700	623	0,8612
Genus	Aleatória Simples	80:20	5	56	700	627	0,8660
Genus	Aleatória Simples	80:20	5	84	700	626	0,8422
Genus	Aleatória Simples	80:20	5	92	700	620	0,8527
Genus	Aleatória Simples	80:20	5	101	700	628	0,8566
Genus	Aleatória Simples	80:20	5	105	700	629	0,8549
Genus	Aleatória Simples	80:20	5	227	700	622	0,8582
Genus	Aleatória Simples	80:20	10	0	700	623	0,8947
Genus	Aleatória Simples	80:20	10	14	700	627	0,9050
Genus	Aleatória Simples	80:20	10	56	700	630	0,9052
Genus	Aleatória Simples	80:20	10	84	700	629	0,9026
Genus	Aleatória Simples	80:20	10	92	700	630	0,9078
Genus	Aleatória Simples	80:20	10	101	700	623	0,9039
Genus	Aleatória Simples	80:20	10	105	700	626	0,9055
Genus	Aleatória Simples	80:20	10	227	700	629	0,9076
Genus	Aleatória Simples	90:10	5	0	700	628	0,8464
Genus	Aleatória Simples	90:10	5	14	700	629	0,8618
Genus	Aleatória Simples	90:10	5	56	700	627	0,8700
Genus	Aleatória Simples	90:10	5	84	700	626	0,8616
Genus	Aleatória Simples	90:10	5	92	700	623	0,8595
Genus	Aleatória Simples	90:10	5	101	700	618	0,8755
Genus	Aleatória Simples	90:10	5	105	700	629	0,8753
Genus	Aleatória Simples	90:10	5	227	700	624	0,8657
Genus	Aleatória Simples	90:10	10	0	700	625	0,9048
Genus	Aleatória Simples	90:10	10	14	700	305	0,8959
Genus	Aleatória Simples	90:10	10	56	700	629	0,9119
Genus	Aleatória Simples	90:10	10	84	700	622	0,9100
Genus	Aleatória Simples	90:10	10	92	700	624	0,9081
Genus	Aleatória Simples	90:10	10	101	700	619	0,9088
Genus	Aleatória Simples	90:10	10	105	700	618	0,9083
Genus	Aleatória Simples	90:10	10	227	700	626	0,9095
Genus	Aleatória Simples	95:05	10	0	700	619	0,8983
Genus	Aleatória Simples	95:05	10	14	700	624	0,9007
Genus	Aleatória Simples	95:05	10	56	700	303	0,9103
Genus	Aleatória Simples	95:05	10	84	700	629	0,9107
Genus	Aleatória Simples	95:05	10	92	700	627	0,9093
Genus	Aleatória Simples	95:05	10	101	700	612	0,9107
Genus	Aleatória Simples	95:05	10	105	700	629	0,9203
Genus	Aleatória Simples	95:05	10	227	700	310	0,9079

Tabela C.8: Resultados dos experimentos com a solução desenvolvida em nível de Genus com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Épocas	Melhor Época	Acurácia
Genus	Estratificada	80:20	5	0	700	630	0,8740
Genus	Estratificada	80:20	5	14	700	629	0,8706
Genus	Estratificada	80:20	5	56	700	630	0,8695
Genus	Estratificada	80:20	5	84	700	625	0,8768
Genus	Estratificada	80:20	5	92	700	630	0,8762
Genus	Estratificada	80:20	5	101	700	622	0,8742
Genus	Estratificada	80:20	5	105	700	628	0,8631
Genus	Estratificada	80:20	5	227	700	624	0,8585
Genus	Estratificada	80:20	10	0	700	630	0,9104
Genus	Estratificada	80:20	10	14	700	621	0,9037
Genus	Estratificada	80:20	10	56	700	616	0,9099
Genus	Estratificada	80:20	10	84	700	626	0,9078
Genus	Estratificada	80:20	10	92	700	625	0,9103
Genus	Estratificada	80:20	10	101	700	621	0,9057
Genus	Estratificada	80:20	10	105	700	628	0,9040
Genus	Estratificada	80:20	10	227	700	628	0,9050
Genus	Estratificada	90:10	5	0	700	630	0,8657
Genus	Estratificada	90:10	5	14	700	622	0,8759
Genus	Estratificada	90:10	5	56	700	305	0,8690
Genus	Estratificada	90:10	5	84	700	628	0,8868
Genus	Estratificada	90:10	5	92	700	624	0,8792
Genus	Estratificada	90:10	5	101	700	628	0,8681
Genus	Estratificada	90:10	5	105	700	625	0,8870
Genus	Estratificada	90:10	5	227	700	627	0,8837
Genus	Estratificada	90:10	10	0	700	300	0,9120
Genus	Estratificada	90:10	10	14	700	630	0,9086
Genus	Estratificada	90:10	10	56	700	627	0,9081
Genus	Estratificada	90:10	10	84	700	624	0,9089
Genus	Estratificada	90:10	10	92	700	623	0,9069
Genus	Estratificada	90:10	10	101	700	627	0,9122
Genus	Estratificada	90:10	10	105	700	630	0,9098
Genus	Estratificada	90:10	10	227	700	628	0,9156
Genus	Estratificada	95:05	10	0	700	307	0,9137
Genus	Estratificada	95:05	10	14	700	624	0,9074
Genus	Estratificada	95:05	10	56	700	299	0,9113
Genus	Estratificada	95:05	10	84	700	625	0,9079
Genus	Estratificada	95:05	10	92	700	620	0,9127
Genus	Estratificada	95:05	10	101	700	307	0,8993
Genus	Estratificada	95:05	10	105	700	627	0,9151
Genus	Estratificada	95:05	10	227	700	629	0,9223

Tabela C.9: Resultados dos experimentos com a solução desenvolvida em nível de Species com amostragem aleatória simples.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Épocas	Melhor Época	Acurácia
Species	Aleatória Simples	80:20	5	0	700	628	0,8515
Species	Aleatória Simples	80:20	5	14	700	625	0,8710
Species	Aleatória Simples	80:20	5	56	700	625	0,8648
Species	Aleatória Simples	80:20	5	84	700	609	0,8604
Species	Aleatória Simples	80:20	5	92	700	628	0,8571
Species	Aleatória Simples	80:20	5	101	700	617	0,8734
Species	Aleatória Simples	80:20	5	105	700	611	0,8660
Species	Aleatória Simples	80:20	5	227	700	622	0,8521
Species	Aleatória Simples	80:20	10	0	700	596	0,9250
Species	Aleatória Simples	80:20	10	14	700	572	0,9374
Species	Aleatória Simples	80:20	10	56	700	616	0,9177
Species	Aleatória Simples	80:20	10	84	700	609	0,9216
Species	Aleatória Simples	80:20	10	92	700	615	0,9396
Species	Aleatória Simples	80:20	10	101	700	616	0,9254
Species	Aleatória Simples	80:20	10	105	700	615	0,9336
Species	Aleatória Simples	80:20	10	227	700	621	0,9340
Species	Aleatória Simples	90:10	5	0	700	615	0,8751
Species	Aleatória Simples	90:10	5	14	700	609	0,8722
Species	Aleatória Simples	90:10	5	56	700	623	0,8905
Species	Aleatória Simples	90:10	5	84	700	611	0,8817
Species	Aleatória Simples	90:10	5	92	700	625	0,8787
Species	Aleatória Simples	90:10	5	101	700	625	0,8893
Species	Aleatória Simples	90:10	5	105	700	611	0,8817
Species	Aleatória Simples	90:10	5	227	700	618	0,8846
Species	Aleatória Simples	90:10	10	0	700	628	0,9263
Species	Aleatória Simples	90:10	10	14	700	611	0,9366
Species	Aleatória Simples	90:10	10	56	700	617	0,9400
Species	Aleatória Simples	90:10	10	84	700	627	0,9263
Species	Aleatória Simples	90:10	10	92	700	623	0,9400
Species	Aleatória Simples	90:10	10	101	700	609	0,9357
Species	Aleatória Simples	90:10	10	105	700	622	0,9383
Species	Aleatória Simples	90:10	10	227	700	629	0,9409
Species	Aleatória Simples	95:05	10	0	700	589	0,9228
Species	Aleatória Simples	95:05	10	14	700	616	0,9365
Species	Aleatória Simples	95:05	10	56	700	603	0,9537
Species	Aleatória Simples	95:05	10	84	700	610	0,9297
Species	Aleatória Simples	95:05	10	92	700	616	0,9331
Species	Aleatória Simples	95:05	10	101	700	599	0,9365
Species	Aleatória Simples	95:05	10	105	700	627	0,9434
Species	Aleatória Simples	95:05	10	227	700	628	0,9262

Tabela C.10: Resultados dos experimentos com a solução desenvolvida em nível de Species com amostragem estratificada.

Nível	Amostragem	Proporção	Limite Mínimo	Seed	Épocas	Melhor Época	Acurácia
Species	Estratificada	80:20	5	0	700	615	0,8660
Species	Estratificada	80:20	5	14	700	628	0,8844
Species	Estratificada	80:20	5	56	700	626	0,8953
Species	Estratificada	80:20	5	84	700	626	0,8746
Species	Estratificada	80:20	5	92	700	626	0,8909
Species	Estratificada	80:20	5	101	700	612	0,8657
Species	Estratificada	80:20	5	105	700	625	0,8648
Species	Estratificada	80:20	5	227	700	619	0,8802
Species	Estratificada	80:20	10	0	700	607	0,9374
Species	Estratificada	80:20	10	14	700	614	0,9447
Species	Estratificada	80:20	10	56	700	608	0,9370
Species	Estratificada	80:20	10	84	700	585	0,9319
Species	Estratificada	80:20	10	92	700	615	0,9336
Species	Estratificada	80:20	10	101	700	622	0,9336
Species	Estratificada	80:20	10	105	700	612	0,9357
Species	Estratificada	80:20	10	227	700	617	0,9284
Species	Estratificada	90:10	5	0	700	608	0,8847
Species	Estratificada	90:10	5	14	700	611	0,8900
Species	Estratificada	90:10	5	56	700	617	0,8936
Species	Estratificada	90:10	5	84	700	614	0,8953
Species	Estratificada	90:10	5	92	700	620	0,8912
Species	Estratificada	90:10	5	101	700	625	0,8882
Species	Estratificada	90:10	5	105	700	611	0,8894
Species	Estratificada	90:10	5	227	700	619	0,8971
Species	Estratificada	90:10	10	0	700	607	0,9383
Species	Estratificada	90:10	10	14	700	609	0,9469
Species	Estratificada	90:10	10	56	700	607	0,9443
Species	Estratificada	90:10	10	84	700	619	0,9349
Species	Estratificada	90:10	10	92	700	617	0,9323
Species	Estratificada	90:10	10	101	700	606	0,9383
Species	Estratificada	90:10	10	105	700	617	0,9443
Species	Estratificada	90:10	10	227	700	627	0,9366
Species	Estratificada	95:05	10	0	700	609	0,9401
Species	Estratificada	95:05	10	14	700	602	0,9486
Species	Estratificada	95:05	10	56	700	604	0,9332
Species	Estratificada	95:05	10	84	700	603	0,9486
Species	Estratificada	95:05	10	92	700	627	0,9332
Species	Estratificada	95:05	10	101	700	613	0,9435
Species	Estratificada	95:05	10	105	700	608	0,9503
Species	Estratificada	95:05	10	227	700	588	0,9384

APÊNDICE D – TABELA DE TEMPOS DO FEATURE CLASSIFIER

Tabela D.1: Duração das etapas de treino e teste dos experimentos usando q2-feature-classifier.

Nível	Proporção	Tempo de Treino			Tempo de Teste		
		Mínimo	Médio	Máximo	Mínimo	Médio	Máximo
Class	80:20	0:01:41	0:01:43	0:01:49	0:00:38	0:00:38	0:00:39
Class	90:10	0:01:50	0:01:52	0:01:56	0:00:28	0:00:29	0:00:30
Class	95:05	0:01:54	0:01:56	0:01:58	0:00:24	0:00:24	0:00:25
Order	80:20	0:01:35	0:01:37	0:01:41	0:00:36	0:00:37	0:00:39
Order	90:10	0:01:43	0:01:45	0:01:49	0:00:28	0:00:28	0:00:30
Order	95:05	0:01:47	0:01:47	0:01:49	0:00:24	0:00:24	0:00:25
Family	80:20	0:01:33	0:01:37	0:01:44	0:00:36	0:00:37	0:00:39
Family	90:10	0:01:41	0:01:46	0:01:55	0:00:28	0:00:28	0:00:30
Family	95:05	0:01:45	0:01:46	0:01:50	0:00:24	0:00:24	0:00:25
Genus	80:20	0:01:17	0:01:41	0:02:12	0:00:32	0:00:37	0:00:44
Genus	90:10	0:01:22	0:01:49	0:02:25	0:00:26	0:00:29	0:00:33
Genus	95:05	0:01:25	0:01:26	0:01:29	0:00:23	0:00:23	0:00:24
Species	80:20	0:00:38	0:00:46	0:00:55	0:00:22	0:00:24	0:00:27
Species	90:10	0:00:40	0:00:48	0:00:58	0:00:21	0:00:22	0:00:24
Species	95:05	0:00:40	0:00:40	0:00:41	0:00:20	0:00:20	0:00:21

APÊNDICE E – TABELA DE TEMPOS DA SOLUÇÃO DESENVOLVIDA

Tabela E.1: Duração das etapas de treino e teste dos experimentos usando a solução desenvolvida.

Nível	**Proporção**	**Tempo de Treino**			**Tempo de Teste**		
		Mínimo	**Médio**	**Máximo**	**Mínimo**	**Médio**	**Máximo**
Class	80:20	1:32:02	1:33:02	1:35:19	0:00:37	0:00:38	0:00:43
Class	90:10	1:33:20	1:37:24	1:40:18	0:00:17	0:00:19	0:00:23
Class	95:05	1:41:30	1:42:12	1:43:30	0:00:12	0:00:13	0:00:17
Order	80:20	1:15:54	1:16:39	1:18:25	0:00:34	0:00:34	0:00:35
Order	90:10	1:20:53	1:22:28	1:25:55	0:00:17	0:00:17	0:00:22
Order	95:05	1:22:54	1:23:40	1:26:08	0:00:12	0:00:13	0:00:16
Family	80:20	1:02:43	1:03:31	1:05:03	0:00:30	0:00:31	0:00:34
Family	90:10	1:06:08	1:07:16	1:09:21	0:00:15	0:00:16	0:00:20
Family	95:05	1:08:58	1:11:24	1:13:02	0:00:11	0:00:12	0:00:15
Genus	80:20	0:34:33	0:38:04	0:42:09	0:00:16	0:00:17	0:00:22
Genus	90:10	0:35:24	0:39:26	0:44:05	0:00:12	0:00:13	0:00:17
Genus	95:05	0:36:03	0:36:44	0:38:42	0:00:10	0:00:11	0:00:14
Species	80:20	0:08:51	0:11:58	0:15:21	0:00:10	0:00:11	0:00:15
Species	90:10	0:09:02	0:12:25	0:15:47	0:00:09	0:00:10	0:00:14
Species	95:05	0:09:14	0:09:35	0:10:05	0:00:08	0:00:09	0:00:11

APÊNDICE F – REVISÃO SISTEMÁTICA DA LITERATURA

No presente apêndice é apresentada a revisão sistemática da literatura realizada com a finalidade de fundamentar esse trabalho. Dado que parte do conteúdo gerado durante o processo de revisão da literatura foi diretamente utilizado na definição do escopo principal desse estudo, algumas seções a seguir, em especial a Introdução(F.1), são compartilhadas com a estrutura principal desse documento.

F.1 Introdução

No campo da biologia, taxonomia é a ciência de nomear, descrever e classificar organismos de acordo com características em comum. Existem diversas classificações taxonômicas, com diferentes critérios e estruturas, das quais a mais conhecida atualmente teve origem no trabalho Systema Naturae [17] publicado em 1735, do botânico sueco Carl Linnæus, e foi sendo aprimorada resultando no que é conhecido como Taxonomia Moderna.

Atualmente, as principais técnicas de classificação são baseadas na análise de sequencias genéticas por meio de algoritmos, permitindo a aplicação em contextos mais complexos e com maior volume de dados. Além disso, novos métodos de sequenciamento genético que foram desenvolvidos ao longo dos últimos anos tornaram as pesquisas na área mais acessíveis e escaláveis, como por exemplo biomonitoramento baseado em eDNA [8]

Environmental DNA, ou eDNA, refere-se ao material genético encontrado em um ambiente [30]. O conjunto de amostras de um local representam as espécies presentes no mesmo, servindo para identificar o perfil do ecossistema. O estudo de eDNA tem se mostrado com grande aplicabilidade em diversos cenários, tais como na conservação de ecossistemas, identificação e mapeamento da biodiversidade. [3][4][23]

Por vezes, estudos que analisam amostras de conjuntos de organismos necessitam de meios de sequenciamento genético mais eficientes e que possam ser aplicados diretamente na amostra, visto que a separação dos organismos em geral não é viável. Nesses casos, é utilizado o *Metabarcoding*, normalmente nos segmentos genéticos 16S rRNA e 18S rRNA, para identificar as espécies presentes nas amostras.

A classificação taxonômica e identificação de espécies por algoritmos são feitas a partir da comparação da sequência genética da amostra com as sequencias de referências, sendo que os critérios de comparação variam de acordo com o algoritmo escolhido. Atualmente, o algoritmo mais utilizado é o Naive Bayes Classifiers[33], um algoritmo de

machine learning cujo método busca avaliar a semelhança entre sequencias pela observação de fragmentos em comum.

Tendo em vista modelos mais recentes de *machine learning*, especialmente *deep learning*, e a capacidade demonstrada pelos mesmos em identificar padrões ao mesmo tempo que conseguem generalizar o aprendizado, levanta-se o questionamento sobre a possível aplicação desses modelos na classificação taxonômica. Atendo-se a esse escopo, este trabalho objetiva identificar, por meio de uma revisão sistemática da literatura, qual o estado da arte da classificação taxonômica usando *deep learning*, quais os resultados e como são comparados com os métodos atuais.

F.1.1 Taxonomia Moderna

Linnæus, o qual ficou conhecido como pai da taxonomia moderna, descreveu em seu trabalho o sistema hierárquico que havia desenvolvido, cuja composição possuía três grupos primários denominados reinos: *Regnum Animale* (reino animal), *Regnum Vegetabile* (reino vegetal) e *Regnum Lapideum* (reino mineral). Durante seus anos de pesquisa, Linnæus foi aprimorando o sistema proposto de diversas formas, tendo publicado ao todo 12 edições do Systema Naturae sob seu nome.

O sistema hierárquico de Linnæus, apesar de ter sido muito importante e já abordar níveis como Classes, Ordens, Gêneros e Espécies; ainda continha diversas inconsistências e lacunas. Um dos exemplos de problema recorrente tinha-se quando diferentes indivíduos eram agrupados em um mesmo *taxon* (como são chamados os grupos individuais; plural *taxa*) por não ter uma divisão adequada para os indivíduos em questão, criando uma falsa associação entre os mesmos. Com o passar dos anos, outros pesquisadores usaram o sistema para criar variações, com melhor entendimento dos organismos, novos conhecimentos e critérios mais precisos e consistentes. Dentre os diversos estudos que contribuíram para o desenvolvimento de versões aprimoradas do sistema hierárquico, um dos que teve maior impacto na forma que a hierarquia é estruturada e nos critérios para classificação foi o "On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life"[7], de Charles Darwin.

O trabalho de Darwin teve grande impacto na classificação taxonômica pois foi um dos primeiros estudos a abordar a relação entre espécies de acordo com um ancestral em comum. Tal mudança permitiu agrupar organismos semelhantes de forma mais consistente e precisa, uma vez que organismos que possuem mesma origem tendem a compartilhar características. Além disso, ainda que não tenha sido abordado o código genético em si na época, o fato de indivíduos de um mesmo grupo possuírem genomas semelhantes possibilitou que novas técnicas fossem desenvolvidas para a utilização de

sequencias genéticas na identificação e classificação de organismos baseados na sua semelhança com sequencias de referência para cada *taxon*.

Atualmente, os principais níveis taxonômicos utilizados são, em ordem hierárquica: Domínio, Reino, Filo, Classe, Ordem, Família, Gênero e Espécie. Dentre esses, o Domínio o mais recente a ser adotado, tendo sua inclusão na estrutura atribuída a Royall T. Moore[19]. Além dos principais níveis, subníveis também são frequentemente utilizados, variam de acordo com cada ramificação onde se encontram. No entanto, devido às novas descobertas sobre os organismos e de novas variáveis que impactam nos critérios de classificação, nem sempre as estruturas hierárquicas tradicionais mostram-se adequadas para a classificação de um grupo de organismos. De forma semelhante ao problema de organismos distintos em um mesmo grupo que ocorria no sistema proposto por Linnæus, a gama de características associadas a cada organismo varia tanto que torna-se difícil definir critérios de classificação que funcionem em todos os casos.

F.1.2 Taxonomia Baseada em OTUs

Uma OTU, do inglês *Operational Taxonomic Unit*, é a unidade básica presente em um sistema de taxonomia numérica. A abordagem da taxonomia numérica é baseada no agrupamento de organismos semelhantes entre si, sendo cada grupo uma OTU diferente.

As técnicas de taxonomia por meio de OTUs são amplamente utilizadas, tendo recebido destaque quando aplicadas no agrupamento de organismos com base na similaridade dos códigos genéticos sequenciados, cenários em que recebe o nome de Molecular Operational Taxonomic Unit (MOTU)[9]. Além disso, um dos benefícios dessa técnica é a possibilidade de agrupar sem definir características em comum para cada grupo, apenas usando a similaridade entre os indivíduos, tornando-a uma alternativa interessante em casos onde as estruturas hierárquicas convencionais não conseguem representar adequadamente grupos distintos.

F.1.3 Metabarcoding

O processo de *metabarcoding* é semelhante ao *barcoding*, tendo como principal diferença o primeiro ser aplicado em uma amostra composta por múltiplos organismos, enquanto o segundo é aplicado em um único indivíduo. Ambos os processos seguem quatro etapas fundamentais: extração do DNA, amplificação por PCR, sequenciamento e análise. [23] [18]

Apesar dos grandes avanços, existem desafios para garantir a qualidade dos resultados das classificações. Contaminação da amostra, viés causado pela escolha do

primer, produção de *chimeras* (réplicas falhas) durante a replicação do material e a classificação taxonômica baseada em referências incompatíveis com a amostra são alguns dos problemas encontrados durante o processo. [2] [18] [26]

F.2 Objetivos

Este trabalho visa identificar o estado da arte no uso de *deep learning* para a classificação taxonômica de organismos vivos eucariontes e/ou procariontes por meio da sequência genética, em especial os segmentos 18S rRNA e 16S rRNA.

F.2.1 Objetivos Específicos

Objetivando a definição clara de um escopo, foram elaboradas três perguntas para guiar esta pesquisa:

- Quais tipos de modelos de *deep learning* estão sendo utilizados para classificação taxonômica?
- Qual a acurácia dos métodos de classificação com *deep learning* atuais?
- Teriam as abordagens utilizando *deep learning* melhor performance que os métodos de classificação com Naïve Bayes?

F.3 Metodologia

Neste trabalho foi utilizada a revisão sistemática da literatura como metodologia de pesquisa para identificação, seleção e análise dos trabalhos relacionados, bem como para o levantamento dos algoritmos e técnicas aplicadas na classificação taxonômica de organismos.

F.3.1 Bases de Pesquisa

Foram selecionadas quatro bases para pesquisar trabalhos para estudo. Os critérios para a escolha dessas bases envolvem reconhecimento internacional, volume de publicações, qualidade e relevância dos trabalhos. Além disso, foi considerada a área de conhecimento a qual a base foca, uma vez que a pesquisa desenvolvida envolve não

apenas o campo da inteligência artificial, como também da biologia, medicina e outros relacionados ao estudo genético.

As bases selecionadas foram:

- IEEE
- Springer
- ScienceDirect
- PubMed

F.3.2 Estratégia de Busca

Realizou-se a busca por meio de uma *string* elaborada contendo os termos relevantes à pesquisa, os quais auxiliam na identificação e seleção de trabalhos relacionados. Para a elaboração da *string* de busca, foi utilizada a estratégia PICO, gerando a Tabela F.1.

Tabela F.1: Definição dos termos de busca aplicando PICO.

	Termos
População	18S rRNA, 16S rRNA
Intervenção	Deep Learning, Machine Learning, CNN
Controle	Naive Bayes, Accuracy
Desfechos (**O**utcome)	Taxonomy Classification

Com isso foi possível gerar a seguinte *string* de busca:

("18S rRNA" OR "16S rRNA") AND ("Deep Learning" OR "Machine Learning" OR "CNN") AND ("Accuracy" OR "Naive Bayes") AND ("Taxonomy" AND "Classification")

F.3.3 Critérios de Inclusão e Exclusão

Com o avanço das técnicas de IA e de sequenciamento genético, muitos estudos têm sido publicados e nem todos, ainda que selecionados pela busca, são relevantes para este estudo. Assim sendo, critérios de inclusão e exclusão foram determinados para auxiliar na seleção dos artigos relevantes.

Os critérios escolhidos buscam restringir o conjunto de artigos de acordo com o tipo de pesquisa desenvolvida, os recursos e resultados compartilhados, a validade da

publicação e com o quão atual é a publicação, visto que os algoritmos de aprendizado profundo abordados neste trabalho datam predominantemente dos últimos cinco anos. Assim sendo, são filtradas para análise as publicações que satisfizerem os seguintes critérios de inclusão:

I1. abordar como tema principal a classificação taxonômica,

I2. deve abordar técnicas de *deep learning*,

I3. ter realizado a implementação e teste das técnicas abordadas,

I4. ser replicável,

I5. ter sido publicado em inglês,

I6. ter sido publicado após 2018; e

I7. possuir um identificador DOI válido.

Além disso, as publicações não podem se enquadrar em um dos seguintes critérios de exclusão:

E1. ser duplicata,

E2. não prover informações sobre os parâmetros do algoritmo; ou

E3. não apresentar métricas de avaliação.

F.3.4 Planejamento de Busca e Seleção

A busca foi realizada nas bases de dados separadamente e, quando possível, já com filtros relacionados aos critérios pré-definidos. As etapas do processo de busca e seleção de trabalhos relacionados foram executadas de acordo com o diagrama apresentado na figura F.1.

F.4 Resultado da Busca

Com base na *string* de busca previamente definida, realizou-se buscas automatizadas nas bases de dados de interesse. Durante esse processo, alguns filtros foram aplicados de acordo com os critérios de inclusão e exclusão, quando disponíveis na ferramenta de pesquisa. Na tabela F.2 são apresentados os números de publicações resultantes das buscas realizadas em cada base, bem como quais os campos que tiveram filtros aplicados.

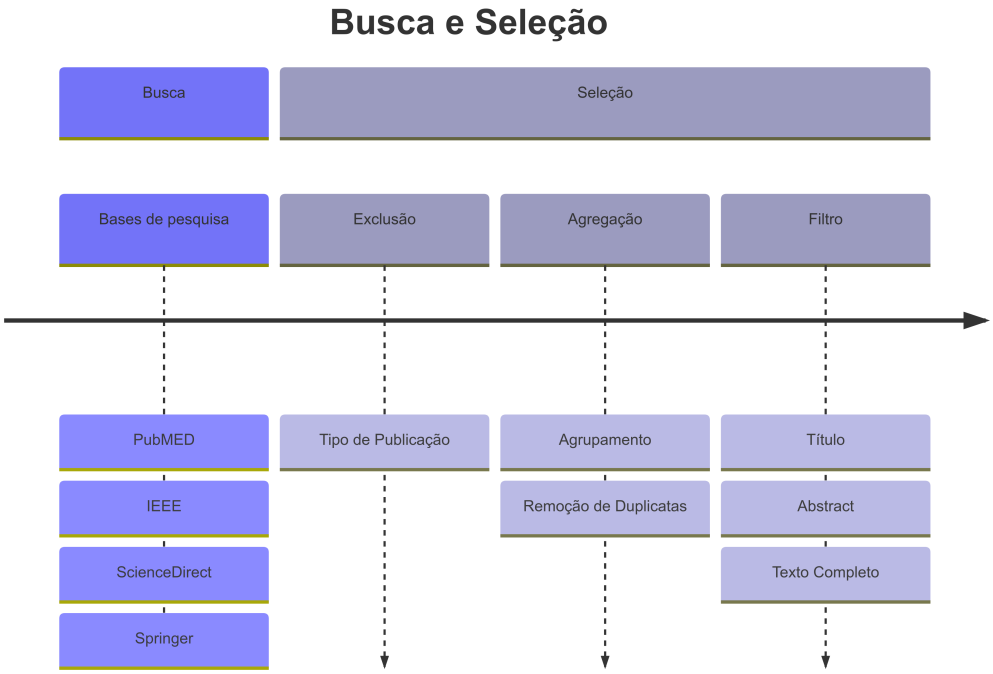

Figura F.1: Diagrama do processo de busca e seleção de publicações.

Tabela F.2: Resultados das buscas.

Base de Pesquisa	Filtros	Número de Publicações
PubMED	Data e Idioma	7
IEEE	Data	3
ScienceDirect	Data e Tipo de Publicação	429
Springer	Data, Idioma e Tipo de Publicação	363

F.5 Seleção

O processo de seleção foi aplicado no conjunto de trabalhos encontrados durante as buscas nas bases de dados selecionadas, seguindo os critérios de inclusão e exclusão previamente definidos. As etapas que compõem o processo de seleção foram executadas de forma sequencial, seguindo a ordem apresentada na tabela F.3.

Na primeira etapa, filtrou-se apenas publicações de acordo com os respectivos tipos. Nessa etapa foram removidos, principalmente, *posters* e *abstracts* provenientes da base de pesquisa Springer, reduzindo o número de resultados de 363 para 317.

Na segunda etapa, realizou-se o agrupamento dos resultados em um único conjunto e removeu-se as duplicatas. Com isso, foi obtido um conjunto com 752 registros.

A terceira etapa teve como finalidade a aplicação dos filtros nos títulos e palavras-chave, reduzindo o tamanho do conjunto de 752 para 125. Nesse momento, foram iden-

Tabela F.3: Resultados da aplicação dos filtros no conjunto de publicações.

Etapa	Removidos	Restantes
1- Filtro por tipo de publicação	46	756
2- Agrupamento e remoção de duplicatas	4	752
3- Aplicação dos filtros nos títulos	627	125
4- Aplicação dos filtros nos *abstracts*	94	31
5- Aplicação dos filtros nos textos completos	23	8

tificados diversos trabalhos cujos contextos de estudo não correspondiam aos critérios estabelecidos. Alguns dos tópicos mais recorrentes nos estudos removidos eram referentes a:

- Saúde, nutrição e sistema gastrointestinal
- Diagnóstico, análise e/ou classificação de câncer
- Análise de impacto, correlação e a associação entre substâncias com doenças ou características do ambiente

Na quarta etapa, aplicou-se os filtros considerando os *abstracts* das publicações. Além de tópicos fora do escopo desejado como os encontrados anteriormente, foi identificada a recorrência de estudos sobre a escolha de biomarcadores para utilização em sistemas de classificação taxonômica e análise biológica. Apesar dos biomarcadores serem utilizados junto a classificadores, os trabalhos avaliavam o impacto dos marcadores e não dos classificadores, não tendo relevância para esta revisão. Após essa parte, restaram 31 dos 125 trabalhos.

Por último, aplicou-se os filtros nos textos completos das publicações. Nesse caso, os trabalhos removidos eram predominantemente referentes ao agrupamento de organismos semelhantes ou a análise de biomas, mas sem abordar a classificação taxonômica. Além disso, foi aberta uma exceção ao não remover o trabalho “Convolutional neural networks improve fungal classification”[31], o qual não aborda os segmentos 18S rRNA e 16S rRNA, porém compartilha diversas características com o contexto de interesse. Com isso, foram selecionados 8 dos 31 trabalhos, listados na tabela F.4.

F.6 Análise

A partir da análise do material selecionado, é possível ter uma visão geral de como estão as pesquisas para aplicação de *machine learning* na classificação taxonômica. Dentre os pontos abordados nos estudos, alguns dos principais são referentes aos dados utilizados e como os modelos são arquitetados, além da aplicabilidade das soluções.

Tabela F.4: Publicações selecionadas.

Título	Data
AmpliconNet: Sequence Based Multi-layer Perceptron for Amplicon Read Classification Using Real-time Data Augmentation [15]	2018
Beating Naive Bayes at Taxonomic Classification of 16S rRNA Gene Sequences. [33]	2021
BERT contextual embeddings for taxonomic classification of bacterial DNA sequences [12]	2022
Convolutional neural networks improve fungal classification [31]	2020
Investigation of machine learning algorithms for taxonomic classification of marine metagenomes [22]	2023
Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. [5]	2018
Taxonomic classification of metagenomic sequences from Relative Abundance Index profiles using deep learning [14]	2021
Taxonomy metagenomic analysis for microbial sequences in three domains system via machine learning approaches [1]	2018

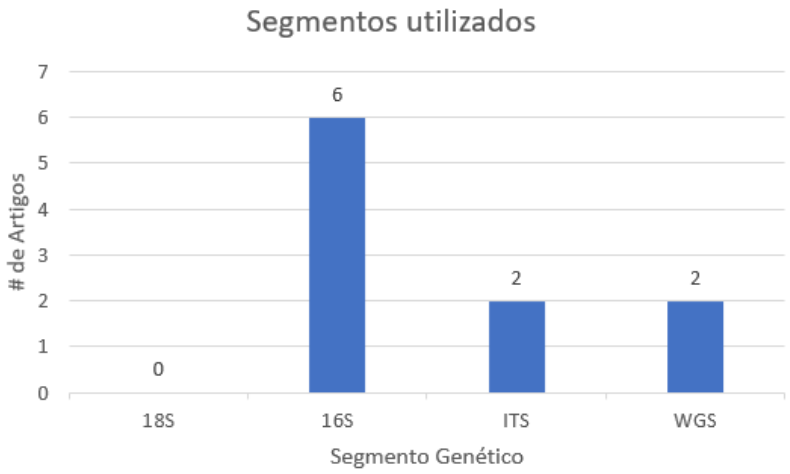

Figura F.2: Gráfico da contagem de artigos que utilizam cada um dos segmentos genéticos.

Em relação aos dados, como mostra a figura F.2, 16S rRNA é o segmento mais utilizado. Essa informação reflete o que foi observado durante o processo de busca e seleção dos artigos, visto que percebeu-se um grande volume de pesquisas de classificação de bactérias e o segmento 16S rRNA é um dos mais utilizados nos estudos genéticos das mesmas. Em contrapartida, nenhuma das publicações selecionadas utilizou o segmento 18S rRNA em seus experimentos.

Ademais, nota-se que não há um padrão na escolha do tamanho das sequências. Observa-se que para o segmento 16S rRNA, o tamanho das sequências utilizadas varia desde 50 bases em alguns experimentos até aproximadamente 1500 bases em outros. A escolha do tamanho tem grande impacto no desenvolvimento, execução e aplicabilidade dos modelos, pois sequências menores são mais baratas de produzir e rápidas para classificar, ao passo que sequências longas geram melhores resultados. Esse fato é reconhecido em "Taxonomy metagenomic analysis for microbial sequences in three domains system via machine learning approaches"[1].

Também é possível perceber que os conjuntos de dados usados são desbalanceados, dificultando tanto o treinamento quanto a avaliação dos resultados, como em casos com possível *overfiting*. Em "Investigation of machine learning algorithms for taxonomic classification of marine metagenomes"[22] é evidenciado esse problema quando identifica-se que as classificações erradas costumam ser atribuídas as classes mais frequentes no conjunto de treino.

O tamanho dos *datasets* também varia bastante entre os estudos, sendo o menor contendo apenas 1.267 sequências enquanto outros chegam a trabalhar com mais de 60.000 sequências. Contudo, sabe-se que a maioria dos modelos de *deep learning* precisam de um volume de dados consideravelmente grande e representativo para que os padrões sejam identificados corretamente e as devidas generalizações sejam aplicadas durante o treinamento.

Já na etapa de testes, alguns trabalhos utilizaram amostras estratificadas para manter a representatividade e tentar garantir que o modelo consiga classificar todas as classes, independentemente da quantidade de registros das mesmas. Ainda assim, a pequena quantidade de registros disponíveis em alguns casos faz com que seja difícil assegurar a relevância estatística dos resultados, visto que por vezes algumas classes não estão presentes no conjunto de teste.

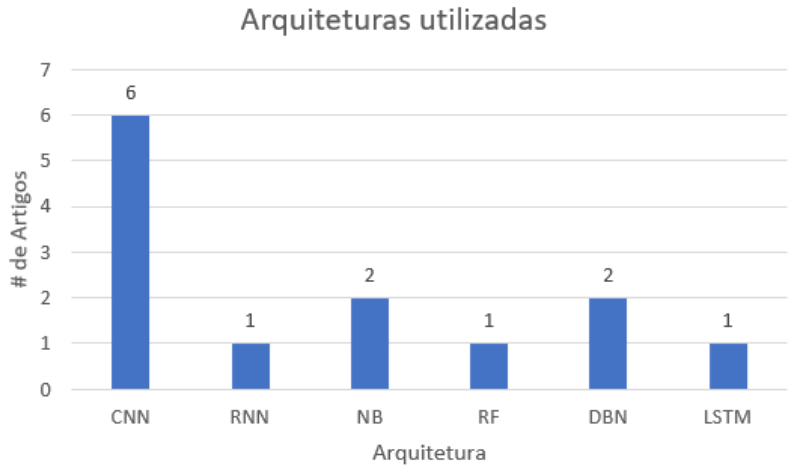

Figura F.3: Gráfico com a quantidade de artigos que implementam cada arquitetura.

Em relação aos tipos modelos de inteligência artificial implementados, destaca-se a recorrente utilização da arquitetura CNN, presente em 6 dos 8 trabalhos. Os resultados mostraram potencial aplicabilidade, com acurácia superando 90% em diversos testes.

Uma forma de melhorar a acurácia que alguns trabalhos exploraram foi adicionar informações sobre a frequência ou distribuição dos registros entre as classes do mesmo nível durante o treinamento. Essas informações acabam sendo utilizadas como pesos das classes, compensando em parte o desbalanceamento dos dados.

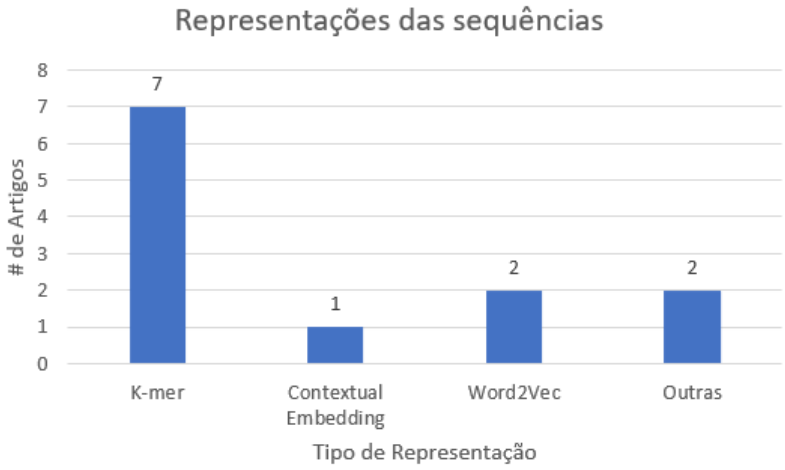

Figura F.4: Gráfico do número de artigos que aplicou cada técnica de representação de sequências.

A escolha da forma de representar a informação das sequências também pode impactar no desempenho dos modelos, sendo a representação K-mer a mais utilizada, como mostrado na figura F.4. K-mers maiores tendem a alcançar resultados um pouco melhores, mas demandam muito mais recursos computacionais. Sendo assim, a maioria dos estudos focou em usar K-mers com tamanho entre 6 e 8.

Em "Taxonomic classification of metagenomic sequences from Relative Abundance Index profiles using deep learning"[14] foi utilizado Relative Abundance Index (RAI) junto com K-mer para a representação dos dados, tornando-a mais informativa. Além disso, outra forma de representação que obteve bons resultados foi a *contextual embedding*, gerada por meio de um modelo BERT em "BERT contextual embeddings for taxonomic classification of bacterial DNA sequences"[12].

Quanto aos problemas recorrentemente encontrados no desenvolvimento, foram citadas limitações de *hardware*, as quais impediram que alguns experimentos aprofundassem o desenvolvimento das arquiteturas dos modelos, em especial do tipo CNN. No entanto, nenhum dos estudos, que reportaram esse problema, fundamentou a definição das estruturas das redes neurais, como tamanho dos *kernels* ou das camadas totalmente conectadas. Essas características podem ter enorme impacto na demanda de recursos computacionais e nem sempre modelos maiores alcançam melhores resultados. Consi-

derando esse cenário, questiona-se a qualidade da definição e otimização dos modelos propostos.

A otimização dos parâmetros de treinamento também é questionável e conflitante entre os trabalhos. Enquanto em "Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin."[5] é apresentada melhoria significativa de acordo com a escolha dos parâmetros, em "Investigation of machine learning algorithms for taxonomic classification of marine metagenomes"[22] afirma-se que o refinamento dos valores não implica em resultados significativamente melhores. Além disso, nesse último trabalho foram executadas apenas 3 épocas de treino com o pretexto de evitar *overfiting*, mas sem a garantia de que o número de épocas seja o fator responsável pelo problema e nem possibilitando o aprendizado por parte do modelo, o que pode justificar a acurácia máxima de 30.5% do modelo CNN proposto.

O tempo de treinamento também é um fator criticado em alguns trabalhos, podendo variar de poucas horas até dias, dependendo de características como a definição da arquitetura, volume de dados e parâmetros de treinamento. Ainda assim, deve-se considerar que a etapa de treinamento de um modelo complexo é a parte mais demorada e ocorre apenas uma vez. Já na etapa de predição não há problema quanto ao tempo de execução, normalmente demorando poucos segundos.

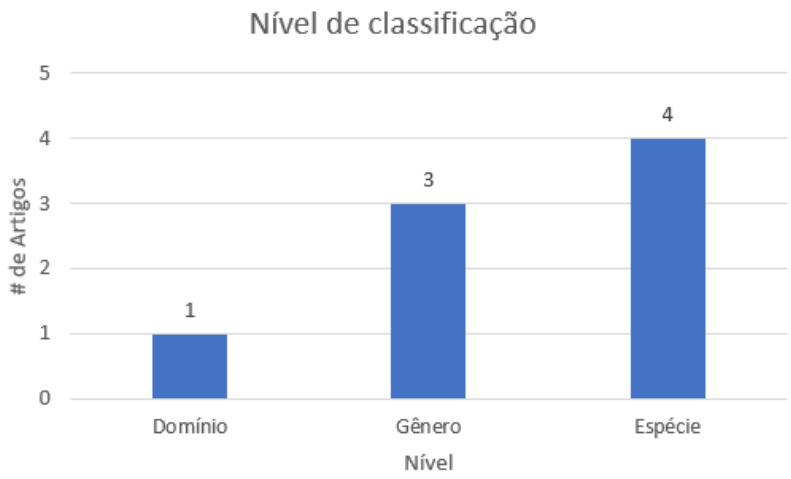

Figura F.5: Quantidade de artigos por nível de classificação mais específico.

Da aplicabilidade e resultados das novas técnicas, é sabido que níveis taxonômicos mais específicos são mais difíceis de classificar, tendo menor acurácia. Dos 8 trabalhos, apenas 4 abordaram a classificação em nível de espécie e outros 3 classificaram no máximo em gênero. O trabalho "Taxonomy metagenomic analysis for microbial sequences in three domains system via machine learning approaches"[1] classificou apenas em nível de domínio (com as classes *Archaea*, *Bacteria* e *Eukarya*) e com isso reportou acurácia superior a 99%. No entanto, não apresenta indícios de replicabilidade em outros níveis. Inclusive, uma das arquiteturas propostas foi a *Deep Belief Network*(DBN), cujo desempe-

nho em níveis mais específicos se mostrou inferior à outras arquiteturas em comparações feitas em outros estudos.

Outro problema identificado é o de *overclassification*. Esse problema fica ainda mais evidente quando tenta-se classificar sequências cuja classe de nível mais alto estava presente no conjunto de treino, mas a classe de nível mais específico não estava. Nesse caso, o modelo é incapaz de classificar corretamente o nível específico, atribuindo a classe mais provável dentre as presentes no treino. Esse cenário impacta diretamente na aplicabilidade das técnicas de classificação propostas, pois encontrar espécies ainda não listadas é algo esperado em tarefas como mapeamento da biodiversidade ou análise de biomas.

F.7 Conclusões

Tendo em mente as perguntas definidas na sessão F.2.1 e que guiaram o desenvolvimento do presente estudo, pôde-se chegar a algumas respostas. Primeiramente, observa-se que os principais trabalhos foram desenvolvidos usando a arquitetura CNN, junto a um crescente interesse em explorar formas de pré-processamento dos dados. Também é possível verificar o potencial da aplicação de *deep learning* para a classificação taxonômica, visto que em diversos casos a acurácia alcançada foi superior ao algoritmo de referência, o Naïve Bayes.

No entanto, é necessário ressaltar que cada experimento foi realizado considerando o escopo definido pelo próprio estudo, não havendo um padrão de dados, requisitos ou limitações compartilhadas que possibilitem a comparação direta entre estudos. Dessa forma, mostra-se interessante a definição de um protocolo avaliativo para comparação das soluções propostas em diferentes cenários de aplicação.

Extrapolando a pesquisa para além das perguntas definidas, nota-se a escassez de informações relacionadas as escolhas das estruturas das redes neurais e dos parâmetros utilizados. Os processos de otimização dos modelos são questionáveis e pouco fundamentados pelos textos apresentados nos trabalhos avaliados, indicando um potencial ramo a ser explorado para alcançar melhores resultados.

Ademais, observa-se que o segmento 16S rRNA está sendo amplamente abordado na área. Todavia, outros segmentos carecem de estudos mais aprofundados. Ainda quanto aos dados, há falta de conjuntos de dados curados com maior abrangência das classes e maior volume de sequências para referência.

Por fim, a aplicabilidade das técnicas também varia de acordo com a situação em que são utilizadas. Por exemplo, em alguns casos pode ser interessante priorizar modelos com menor ocorrência de falsos positivos a modelos com maior abrangência da classificação. Esse tipo de escolha impacta diretamente na seleção de métricas e

formas de avaliação. Melhores definições dos cenários de aplicação, origem e qualidade dos dados, demandas e níveis de confiança esperados são características primordiais no desenvolvimento de novas técnicas.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Pesquisa e Pós-Graduação
Av. Ipiranga, 6681 – Prédio 1 – Térreo
Porto Alegre – RS – Brasil
Fone: (51) 3320-3513
E-mail: propesq@pucrs.br
Site: www.pucrs.br