UNIVERSIDADE ESTADUAL DE CAMPINAS
Instituto de Biologia

FÁBIO MALTA DE SÁ PATRONI

APRENDIZADO DE MÁQUINA APLICADO AO ESTUDO DE SUBTIPOS DAS MALIGNIDADES MIELOIDES

MACHINE LEARNING MODELS APPLIED TO THE SUBTYPES CLASSIFICATION OF THE MYELOID MALIGNANCIES

CAMPINAS
2024

FÁBIO MALTA DE SÁ PATRONI

APRENDIZADO DE MÁQUINA APLICADO AO ESTUDO DE SUBTIPOS DAS MALIGNIDADES MIELOIDES

MACHINE LEARNING MODELS APPLIED TO THE SUBTYPES CLASSIFICATION OF THE MYELOID MALIGNANCIES

Tese apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Genética e Biologia Molecular, na área de Bioinformática.

Dissertation presented to the Institute of Biology of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Genetics and Molecular Biology, in the area of Bioinformatics.

Supervisor/Orientador:

Prof. Dr. Marcelo Mendes Brandão

ESTE TRABALHO CORRESPONDE À VERSÃO
FINAL DA TESE DEFENDIDA PELO ALUNO FÁBIO
MALTA DE SÁ PATRONI, E ORIENTADA PELO
PROF. DR. MARCELO MENDES BRANDÃO.

CAMPINAS

2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

Patroni, Fábio Malta de Sá, 1993-
P276a Aprendizado de máquina aplicado ao estudo de subtipos das malignidades mieloides / Fábio Malta de Sá Patroni. – Campinas, SP : [s.n.], 2024.

Orientador: Marcelo Mendes Brandão.
Tese (doutorado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Biologia.

1. Ciência de dados. 2. Aprendizado de máquina. 3. Leucemia mieloide aguda. 4. Síndromes mielodisplásicas. 5. Leucemia Mieloide - Classificação. I. Brandão, Marcelo Mendes, 1974-. II. Universidade Estadual de Campinas (UNICAMP). Instituto de Biologia. III. Título.

Informações Complementares

Título em outro idioma: Machine learning models applied to the subtypes classification of the myeloid malignancies
Palavras-chave em inglês:
Data science
Machine learning
Acute myeloid leukemia
Myelodysplastic syndromes
Myeloid leucemia - Classification
Área de concentração: Bioinformática
Titulação: Doutor em Genética e Biologia Molecular
Banca examinadora:
Marcelo Mendes Brandão [Orientador]
Paula de Melo Campos
Mariana Camargo Maschietto
Gabriel Ravanhani Schleder
Paulo Sergio Lopes de Oliveira
Data de defesa: 08-08-2024
Programa de Pós-Graduação: Genética e Biologia Molecular

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-0221-8380
- Currículo Lattes do autor: http://lattes.cnpq.br/6242216792745630

Campinas, 08 de Agosto de 2024

COMISSÃO EXAMINADORA

Prof. Dr. Marcelo Mendes Brandão

Profa. Dra. Paula de Melo Campos

Profa. Dra. Mariana Camargo Maschietto

Prof. Dr. Gabriel Ravanhani Schleder

Prof. Dr. Paulo Sergio Lopes de Oliveira

Os membros da Comissão Examinadora acima assinaram a Ata de Defesa, que se encontra no processo de vida acadêmica do aluno.

A Ata de Defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

AGRADECIMENTOS

A Deus.

À Carolina minha esposa, amiga e companheira, por todo amor, carinho e apoio nas mais diversas tribulações, facilitando a minha jornada sendo o sorriso nos momentos difíceis.

À minha família, pelo amor e companheirismo que sempre me acompanha.

Ao Dr. Marcelo Mendes Brandão pela confiança em me receber em seu laboratório, e pela disposição, e orientação. Obrigado pelo apoio.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de estudos fornecida sob o processo n° 140610/2020-0, a qual possibilitou o desenvolvimento deste trabalho.

Ao grupo LaBIS e colegas do CBMEG, pelo suporte durante esses dois anos, além da amizade.

“The results shown here are in whole or part based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga) and by the Therapeutically Applicable Research to Generate Effective Treatments (https://www.cancer.gov/ccg/research/genome-sequencing/target) initiative, phs000218. The data used for this analysis are available at the Genomic Data Commons (https://portal.gdc.cancer.gov)." *O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.*

RESUMO

Na ciência de dados diversos métodos são utilizados para analisar grandes volumes de dados (Big Data) e extrair informações valiosas. Tal área é integrada pela estatística, ciência da computação, matemática aplicada e visualização para transformar vastas quantidades de dados em novos insights e conhecimentos. Atualmente, a ciência de dados se subdivide em várias áreas, como a Inteligência Artificial (IA), que, apesar de muitas vezes ser confundida com aprendizado de máquina (ML), é uma área mais abrangente que inclui ML como uma de suas subáreas. Tom Mitchell definiu aprendizado de máquina em 1997 como a capacidade de um programa de computador melhorar seu desempenho em uma tarefa através da experiência. Em outras palavras, um programa é considerado como "aprendendo" se sua performance em uma tarefa melhora com a experiência acumulada. O aprendizado de máquina (ML) envolve o uso de algumas áreas como probabilidade, estatística e álgebra linear, as quais possibilitam o aprendizado de forma automática. Os modelos de ML identificam regras e padrões, tornando-se mais complexos com mais atributos e mais bem ajustados com mais dados observacionais, algo que seria inviável para o cérebro humano processar. Eles podem ser empregados em diversas áreas do conhecimento, como na medicina de precisão, integrando diversos tipos de dados, identificando características importantes e respondendo a questões biológicas aplicáveis na prática médica para compreender doenças humanas. As abordagens de ML visam preencher a lacuna entre a capacidade de gerar grandes volumes de dados e a compreensão dos sistemas biomédicos. As aplicações emergentes prometem um futuro promissor para a integração de dados biomédicos, permitindo novas abordagens baseadas em diferentes modelos de dados e resultados biomédicos. O presente trabalho está dividido em dois capítulos. O primeiro aborda o uso de ML com dados clínicos de pacientes portadores de Malignidades Mieloides, especialmente a Leucemia Mielóide Aguda (LMA), e a Síndrome Mielodisplásica (SMD), evidenciando o potencial desses modelos para entender a doença e seus subtipos. Foi explorado o melhor modelo treinado para cada doença, através do uso de métodos como SHAP e Lime com o objetivo de proporcionar maior interpretabilidade e explicabilidade dos resultados. Esses modelos demonstram potencial para auxiliar na compreensão

dessas doenças e fornecer novos insights sobre seus subtipos. O segundo capítulo trata da aplicação de aprendizado de máquina em dados de pacientes com câncer na cavidade oral. Este trabalho resultou em uma publicação que demonstra o potencial de seis glicanos específicos para estratificar pacientes com e sem metástase linfonodal. Este estudo evidencia como modelos de ML podem ser aplicados em diferentes contextos clínicos para melhorar a precisão diagnóstica e as estratégias de tratamento. Em conclusão, a ciência de dados, e em particular o aprendizado de máquina, oferece ferramentas poderosas para transformar dados complexos em conhecimentos acionáveis, com aplicações significativas na área da saúde. O progresso contínuo nesta área promete avanços ainda maiores na personalização e eficácia dos tratamentos médicos, beneficiando pacientes e profissionais de saúde.

Palavras-chave: Ciência de dados; Aprendizado de Máquina; Malignidades Mieloides; Leucemia Mieloide Aguda; Subtipos; Classificação

ABSTRACT

In data science, various methods are used to analyze large volumes of data (Big Data) and extract valuable information. This area integrates statistics, computer science, applied mathematics and visualization to transform vast amounts of data into new insights and knowledge. Today, data science is subdivided into several areas, such as Artificial Intelligence (AI), which, although it is often confused with machine learning (ML), is a broader area that includes ML as one of its subareas. Tom Mitchell defined machine learning in 1997 as the ability of a computer program to improve its performance on a task through experience. In other words, a program is considered to be “learning” if its performance on a task improves with accumulated experience. Machine learning (ML) involves the use of areas such as probability, statistics and linear algebra, which make it possible to learn automatically. ML models identify rules and patterns, becoming more complex with more attributes and better adjusted with more observational data, something that would be unfeasible for the human brain to process. They can be used in various areas of knowledge, such as precision medicine, integrating different types of data, identifying important characteristics and answering biological questions applicable in medical practice to understand human diseases. ML approaches aim to bridge the gap between the ability to generate large volumes of data and the understanding of biomedical systems. Emerging applications promise a promising future for biomedical data integration, enabling new approaches based on different data models and biomedical outcomes. This work is divided into two chapters. The first addresses the use of ML with clinical data from patients with Myeloid Malignancies, especially Acute Myeloid Leukemia (AML), and Myelodysplastic Syndrome (MDS), highlighting the potential of these models to understand the disease and its subtypes. The best trained model for each disease was explored using methods such as SHAP and Lime in order to provide greater interpretability and explainability of the results. These models show the potential to help understand these diseases and provide new insights into their subtypes. The second chapter deals with the application of machine learning to data from patients with cancer of the oral cavity. This work resulted in a publication that demonstrates the potential of six specific glycans to stratify patients with and without lymph node metastasis. This study highlights how ML models can be applied in

different clinical contexts to improve diagnostic accuracy and treatment strategies. In conclusion, data science, and in particular machine learning, offers powerful tools for transforming complex data into actionable knowledge, with significant applications in healthcare. Continued progress in this area promises even greater advances in the personalization and effectiveness of medical treatments, benefiting patients and healthcare professionals alike.

Keywords: Data science; Machine Learning; Myeloid malignancies; Acute Myeloid Leukemia; Subtypes; Classification

SUMÁRIO

1. Introdução..... 13
1.1. A Ciência de Dados..... 13
1.2. O que é o Aprendizado de Máquina?..... 14
1.2.1. Aplicações na saúde..... 14
1.3. Divisão de capítulos..... 16
2. Capítulo 1. APRENDIZADO DE MÁQUINA APLICADO AO ESTUDO DE SUBTIPOS DAS MALIGNIDADES MIELOIDES..... 17
2.1. Introdução..... 17
2.1.1. As Malignidades Mieloides..... 17
2.1.1.1. Leucemia Mieloide Aguda..... 20
2.1.1.2. Síndrome Mielodisplásica..... 23
2.1.2. Os Bancos de Dados Biológicos..... 24
2.2. Objetivos..... 27
2.2.1. Objetivo Geral..... 27
2.2.2. Objetivos Específicos..... 27
2.3. Materiais E Métodos..... 27
2.3.1. Aquisição dos dados..... 28
2.3.2. A biblioteca de ML..... 30
2.3.3. Pré-processamento dos dados..... 30
2.3.4. Seleção e validação dos modelos..... 33
2.3.5. Interpretabilidade e Explicabilidade..... 38
2.4. Resultados..... 40
2.4.1. Aquisição e pré-processamento dos dados..... 41
2.4.2. Treino e seleção de modelos..... 42
2.4.2.1. LMA..... 44
2.4.2.1.1. Dados unificados..... 45
2.4.2.1.2. Dados individualizados..... 49
2.4.2.1.2.1. TCGA..... 49
2.4.2.1.2.2. TARGET..... 52
2.4.2.1.2.3. Hemocentro da Unicamp..... 54
2.4.2.2. SMD..... 57
2.4.2.2.1. Dados unificados..... 58
2.4.2.2.1.1. Predição do sistema da OMS 2016..... 58
2.4.2.2.1.2. Predição do índice de prognóstico (IPSS-R)..... 61
2.4.2.2.2. Dados individualizados..... 64
2.4.2.2.2.1. Dados UFCeará..... 64
2.4.2.2.2.2. MDS-IWG..... 67
2.4.2.2.2.3. Hemocentro..... 70
2.4.3. Interpretabilidade e Explicabilidade dos melhores modelos..... 73

2.5. Conclusão, Próximos Passos e Doutorado Sanduíche..... 76
2.6. Conclusão e Próximos Passos..... 76
2.7. Doutorado sanduíche..... 78
2.8. Referências..... 79
3. Capítulo 2. Comprehensive Glycoprofiling of Oral Tumors Associates N-Glycosylation With Lymph Node Metastasis and Patient Survival..... 84
3.1. Introduction..... 86
3.2. Methods..... 87
3.2.1. OSCC primary tumor tissues..... 87
3.2.2. Tissue sample preparation..... 88
3.2.3. N-glycan release..... 89
3.2.4. N-glycome profiling..... 89
3.2.5. N-glycan data analysis..... 90
3.2.6. Protein digestion and peptide desalting..... 91
3.2.7. TMT labeling of peptides..... 91
3.2.8. Glycopeptide enrichment..... 92
3.2.9. High-pH reversed phase prefractionation..... 93
3.2.10. (Glyco)proteome profiling by LC-MS/MS..... 93
3.2.11. N-glycoproteomics data analysis..... 94
3.2.12. Proteomics data analysis..... 95
3.2.13. Experimental design and statistical rationale..... 95
3.2.14. Clinical association and survival analysis..... 96
3.2.15. ROC curve analysis by logistic regression and random forest models... 97
3.2.16. Enrichment analysis of gene ontology (GO) terms..... 97
3.3. Results..... 97
3.3.1. Comprehensive N-glycoprofiling of primary tumors from OSCC patients. 97
3.3.2. Comparative N-glycome profiling of OSCC tissues from N+ and N0 patients..... 98
3.3.3. Comparative N-glycoproteome profiling of OSCC primary tumor tissues from N+ and N0 patients..... 100
3.3.4. N-glycosylation-guided clustering of OSCC patients..... 103
3.3.5. Associations between N-glycoproteome components, clinical features and biological processes..... 105
3.3.6. Subsets of OSCC N-glycans and N-glycopeptides associate with key clinical outcomes..... 107
3.4. Discussion..... 109
3.5. Author Contributions..... 114
3.6. References..... 115
4. Referências Bibliografias Gerais..... 125
5. Apêndices..... 140
5.1. Apêndice A..... 140

5.2. Apêndice B..... 153
6. Anexos..... 169
6.1. Anexo A..... 169
6.2. Anexo B..... 170
6.3. Anexo C..... 171

1. Introdução

1.1. A Ciência de Dados

A ciência de dados envolve o uso de métodos automatizados para analisar grandes quantidades de dados (*Big Data*) e extrair informações. Ao combinar aspectos de estatística, ciência da computação, matemática aplicada e visualização, a ciência de dados pode transformar uma grande quantidade de dados gerados, em novas ideias e conhecimentos. Atualmente a ciência de dados está envolvida em diversas áreas, por exemplo, no aprendizado de máquina e na Inteligência Artificial (Figura 1). Isso ocorre pois para compreender certos processos é preciso integrar sistematicamente diversos tipos de informações, utilizando relacionamentos relevantes (RICHARD A. DUSCHL *et al.*, 2007).

Figura 1. ***As subdivisões e relacionamento entre os diferentes campos da Ciência de Dados até o nível do Deep Learning.*** Embora confundidos como o mesmo conceito, o aprendizado de máquina é uma sub-área dentro da área de inteligência artificial *(CHEN et al., 2022)*. Traduzido de Singapore Computer Society.

1.2. O que é o Aprendizado de Máquina?

Em 1997, Tom Mitchell utilizou a seguinte frase para responder a seguinte pergunta “O que é o Aprendizado de Máquina?”:

Diz-se que um programa de computador aprende a partir da experiência ‘E’ em relação a alguma tarefa ‘T’ e alguma medida de desempenho ‘P’, se o seu desempenho em ‘T’, como medido por ‘P’, melhora com a experiência ‘E’ -- Tom Mitchell, 1997.

Espera-se que a inclusão da experiência ‘E’ em suas tarefas acabe aprimorando o aprendizado, com objetivo final de torná-lo automático (GÉRON, 2017). O processo de aprendizagem (treinamento) e a saída que esse processo produz é chamado de modelo. Por sua vez, um modelo pode receber novos dados e pode argumentar sobre essa nova informação, com base no que foi aprendido anteriormente.

Os modelos de ML determinam um conjunto de regras, usando grandes quantidades de poder de computação, de forma que um cérebro humano seria incapaz de processar. Quanto aos dados, mais atributos (preditores) tornam o modelo mais complexo, por outro lado mais observações (pacientes, por exemplo) ou podem torná-lo mais preciso (HASTIE *et al.*, 2009).

1.2.1. Aplicações na saúde

Os fatores relevantes para a compreensão de uma doença comumente não são encontrados num único tipo de dado, por exemplo expressão gênica. Por conta de tal demanda, métodos integrativos que combinam dados de várias tecnologias têm sido usados para atender a esse desafio (ZITNIK *et al.*, 2018). Assim sendo, um modelo treinado e validado com diferentes tipos de dados pode realizar predições com maior performance. Dessa forma, o modelo de aprendizagem automática identifica características importantes e prevê resultados, respondendo a perguntas biológicas que podem ser aplicadas na prática médica (ZITNIK *et al.*, 2018).

Um estudo recente demonstrou o potencial do uso de métodos de aprendizado de máquina para a predição de pacientes com Diabetes Mellitus tipo-2 (GANIE; MALIK, 2022). Neste trabalho, foram utilizados dados de 1939 pacientes e 11 parâmetros. Os dez primeiros eram compostos de variáveis independentes relacionados à biologia e estilo de vida deles, e o último era a variável dependente binária, sendo 1 caso o paciente fosse diabético (GANIE; MALIK, 2022). Após testar diversos modelos, uma vez que a priori nenhum modelo é melhor que o outro para uma determinada tarefa, o modelo de "*bagged decision tree"* foi o selecionado, com um alto valor de acurácia (99.41%). Portanto, este modelo treinado tem o potencial de auxiliar um profissional da saúde no diagnóstico de uma possível suspeita de Diabetes Mellitus tipo-2.

Na área oncológica também existem alguns exemplos de aplicação de ML. Em uma das problemáticas, pacientes com câncer podem ser submetidos a uma combinação terapêutica de radiação e vários medicamentos todos os meses. Entretanto, esses medicamentos podem causar múltiplos efeitos colaterais (INCA, 2019). Em 2018, pesquisadores do MIT desenvolveram um modelo capaz de melhorar a qualidade de vida dos pacientes com glioblastoma (câncer que afeta cérebro e medula espinal), reduzindo as doses de quimioterapia e de radioterapia (YAUNEY; SHAH, 2018). Esse modelo pode ajudar nas decisões de tratamento tornando os regimes de dosagem menos tóxicos, todavia ainda eficazes. O modelo foi alimentado com os dados de tratamento, atualmente em uso, e ele ajusta as doses iterativamente, para otimizar a quimioterapia e os regimes de dosagem dos ensaios clínicos. O modelo aprendeu regimes de dosagem os quais reduziam os diâmetros médios tumorais (MTD) em ensaios clínicos simulados que envolviam quimioterapia com temozolomida (TMZ) e PCV. Além da redução do MTD, a abordagem também reduziu com sucesso os ciclos de tratamento e as concentrações de dosagem em comparação com os regimes de especialistas humanos, fornecendo um novo método adaptável a outras tarefas de aprendizagem não estruturadas (YAUNEY; SHAH, 2018). Eventualmente, um plano de tratamento com o mesmo potencial dos regimes tradicionais era encontrado, com menor potência (de um quarto a metade) e frequência de doses possível (YAUNEY; SHAH, 2018).

Essas abordagens visam preencher a lacuna entre a capacidade de gerar grandes quantidades de dados (o *Big Data*) e a compreensão dos sistemas

biomédicos. As aplicações emergentes do ML prometem um futuro promissor e inspirador para a integração de dados biomédicos, como nunca antes realizado. Dessa forma, serão possíveis novas abordagens, de acordo com diferentes tipos de modelos de dados e de resultados biomédicos. A mineração de conhecimento permite que a medicina de precisão seja orientada por dados. Entretanto, os conjuntos de dados biomédicos refletem o conhecimento científico descoberto até o momento. Essa limitação ocasiona uma piora no desempenho preditivo, uma vez que nem todas as variáveis necessárias para modelar um determinado problema biológico são conhecidas. Por isso, é desafiador implementar sistemas de aprendizado de máquina, para apoiar a tomada de decisões em práticas clínicas (MULLAINATHAN; OBERMEYER, 2017).

1.3. Divisão de capítulos

O presente trabalho está subdividido em 2 capítulos. O capítulo 1 apresenta ao leitor uma aplicação do aprendizado de máquina com foco em dados de pacientes afetados por alguma Malignidade Mielóide, como Leucemia Mielóide Aguda e Síndrome Mielodisplásica. O melhor modelo treinado para cada uma das duas doenças foi explorado com o propósito de interpretabilidade e explicabilidade. Por fim, os resultados mostram o potencial de aplicação destes modelos para auxiliar o entendimento da doença, assim como fornecer um novo ponto de vista dos subtipos vigentes. No capítulo 2, a aprendizagem supervisionada foi aplicada em dados de pacientes com câncer na cavidade oral, resultando em uma publicação. Neste trabalho foi demonstrado o potencial de separação de seis glicanos ("polímeros de açúcares") para estratificar pacientes sem, e com metástase linfonodal, isso é, células tumorais que espalharam para outra parte do corpo, os linfonodos.

2. Capítulo 1. APRENDIZADO DE MÁQUINA APLICADO AO ESTUDO DE SUBTIPOS DAS MALIGNIDADES MIELOIDES

2.1. Introdução

2.1.1. As Malignidades Mieloides

As malignidades mieloides são doenças clonais das células-tronco hematopoiéticas ou progenitoras (MURATI *et al.*, 2012). Elas são um grupo heterogêneo caracterizado por proliferação excessiva, auto-renovação anormal e/ou defeitos de diferenciação de células hematopoiéticas (origina todos os tipos de células sanguíneas) e células progenitoras mieloides (origina as células da linhagem mieloide sanguíneas) (Figura 2). Esses distúrbios são resultantes de alterações genéticas e epigenéticas nas células-tronco cancerígenas e nas células do nicho da medula óssea. Dessa forma, processos-chave como a auto-renovação, a proliferação e a diferenciação são prejudicados (NTZIACHRISTOS *et al.*, 2013) Modificações genéticas e epigenéticas também foram observadas em células estromais da medula óssea em Síndromes Mielodisplásicas e em Leucemia Mieloide Aguda (KORN; MÉNDEZ-FERRER, 2017).

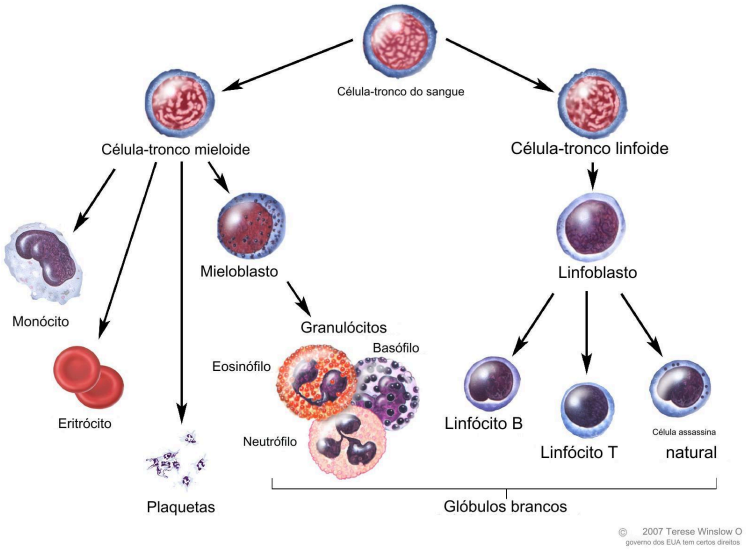

Figura 2. ***Esquema da hematopoiese fisiológica.*** A partir de uma célula tronco sanguínea são originadas as duas linhagens que compõem o sangue humano, a linhagem Mieloide (à esquerda) e a linhagem Linfoide (à direita). Traduzido de 2007 Terese Winslow.

Com base na morfologia, citoquímica, imunofenótipo, genética e características clínicas dos distúrbios, em 2022, o Consenso Internacional de Classificação (ICC) previu a separação em 12 subtipos principais (ARBER *et al.*, 2022):

1. Citopenias clonais pré-malignas e síndromes mielodisplásicas (SMD);
2. Doenças pediátricas e/ou associadas a mutações da linha germinal;
3. Leucemia aguda de linhagem ambígua;
4. Leucemia/linfoma linfoblástico B;
5. Leucemia/linfoma linfoblástico T;
6. Leucemia Mieloide Aguda (LMA);
7. Mastocitose;
8. Neoplasia blástica de células dendríticas plasmocitóides;
9. Neoplasias mielodisplásicas/mieloproliferativas (NMD/NMP);
10. Neoplasias mielóides/linfóides com eosinofilia e fusões do gene da tirosina quinase;
11. Neoplasias Mieloproliferativas (NMP);
12. Proliferações mieloides associadas à síndrome de Down.

A classificação foi possível graças a uma colaboração entre a Organização Mundial da Saúde (OMS), a Sociedade de Hematopatologia e a Associação Europeia de Hematopatologia (ARBER *et al.*, 2022). Distúrbios mieloides, como a LMA, são considerados malignos há muito tempo, enquanto outros até recentemente eram considerados distúrbios sanguíneos não malignos, ou de "pré-leucemia", com potencial maligno, como ocorreu com a SMD. No entanto, descobertas científicas recentes indicaram que esses distúrbios sanguíneos "pré-leucemia" eram na verdade formas de câncer de sangue de crescimento lento (MA, 2012).

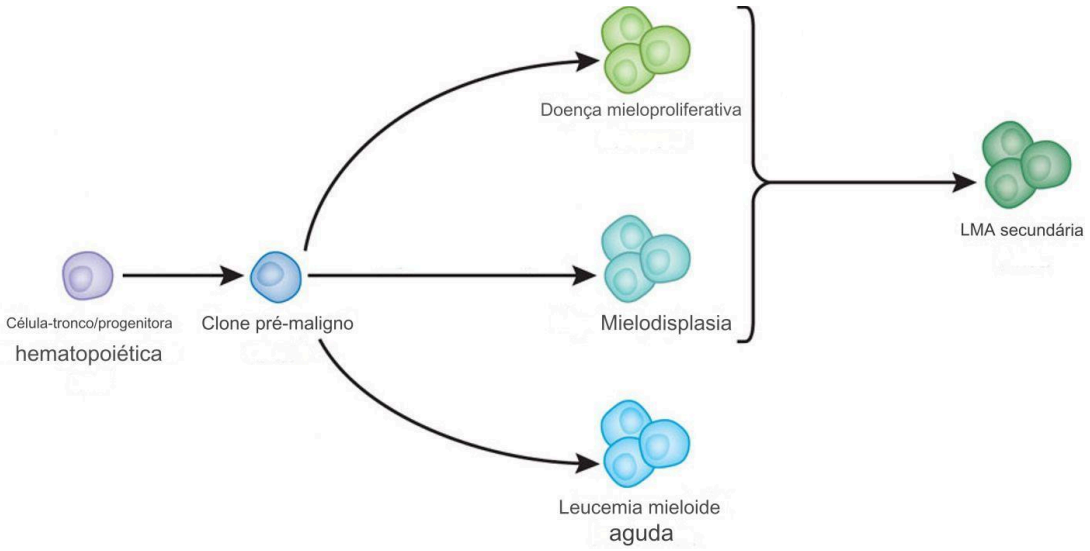

Figura 3. ***Esquema demonstrando potenciais pontos de origem das malignidades mieloides.*** Mutações, alterações cromossômicas e alterações epigenéticas estão entre as potenciais causas da formação dos clones pré-malignos bem como das células indiferenciadas. Adaptado de Mullighan, C. TET2 mutations in myelodysplasia and myeloid malignancies. Nat Genet 41, 766–767 (2009) doi:10.1038/ng0709-766

O escopo de análise do projeto em questão visa dois dos principais de malignidades mieloides: a LMA e a SMD (Figura 3). O fator limitante do escopo são os tipos de malignidades mieloides com dados disponíveis nos bancos de dados públicos e no Hemocentro da UNICAMP. A correta classificação de um paciente dentro de uma doença é importante para definir tratamentos personalizados e prognóstico dos pacientes, bem como é de interesse acadêmico.

Classificar os pacientes destas duas doenças requer um especialista, muitas vezes presentes nos grandes centros. Já os exames de referência para classificação são específicos, custosos, além de não serem oferecidos e realizados com a mesma disponibilidade em todo território nacional. Portanto, seria de grande importância um sistema de suspeita diagnóstica para encaminhamento precoce treinado com dados mais simples de se obter, e que auxilia o corpo clínico para o encaminhamento imediato do paciente para um determinado tratamento mais adequado.

2.1.1.1. Leucemia Mieloide Aguda

A leucemia é um termo para o câncer das células do sangue. Essa doença é caracterizada pelo acúmulo de glóbulos brancos malignos na medula óssea e sangue, comprometendo a hematopoese (HOFFBRAND; MOSS, 2015). Os principais sintomas são anemia, trombocitopenia (número reduzido de plaquetas) e neutropenia (número reduzido de neutrófilos). Fadiga, falta de ar, palpitação e dor de cabeça são alguns dos sintomas/sinais da anemia. A neutropenia pode predispor ao desenvolvimento de infecções severas. Já os sangramentos, que são mais comuns na pele, equimoses e petéquias, são causados pela trombocitopenia. Os quatro tipos principais de leucemia são: Leucemia Mieloide Aguda (LMA), Leucemia Linfoide Aguda (LLA), Leucemia Linfoide Crônica (LLC) e Leucemia Mieloide Crônica (LMC).

O acompanhamento da incidência nacional das leucemias é realizado a cada dois anos pelo Instituto Nacional de Câncer (INCA) que divulga dados de diversos tipos de câncer no Brasil. A Figura 4 ilustra o avanço das leucemias no país baseado nestes dados. Neste gráfico é possível identificar que, num intervalo de dez anos, houve um aumento no número de casos das leucemias. Nesse mesmo intervalo, o número de homens acometidos com leucemias foi em média ~20% maior quando comparado ao de mulheres.

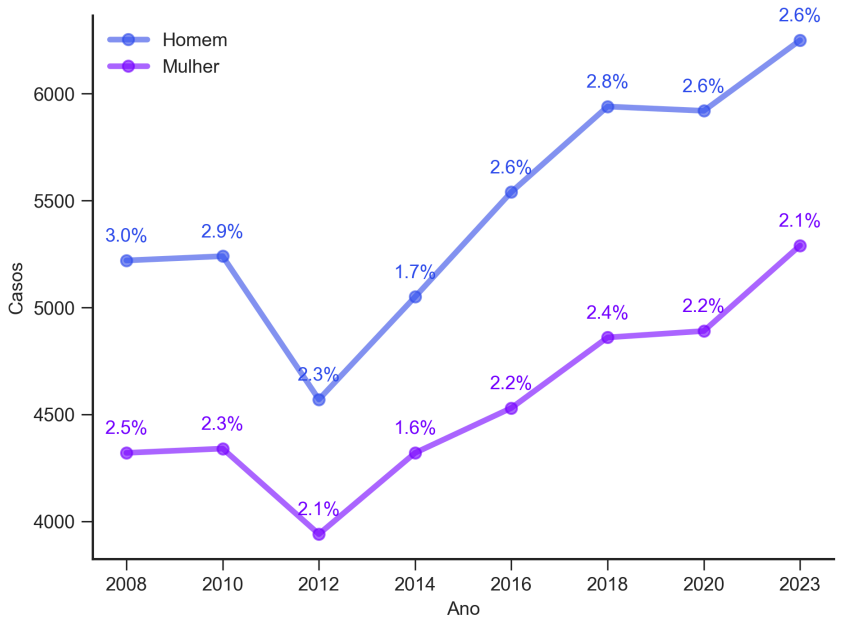

*Figura 4**. Incidência das leucemias no Brasil entre 2008 e 2020.** Na média, nos últimos dez anos houve um aumento no número de casos de Leucemia. Também aumentou a porcentagem*

corresponde à Leucemia do total dos casos de câncer (exceto pele não melanoma) por 100 mil habitantes, no Brasil. Os números acima dos pontos do gráfico indicam a porcentagem de leucêmicos, em relação ao total de cânceres em cada ano. A tendência de acometer mais homens do que mulheres se manteve (dados de incidência de câncer no Brasil de 2008 a 2020 - INCA).

A LMA é a forma mais comum de leucemia aguda em adultos e torna-se cada vez mais comum com o aumento da idade, com um início mediano de 65 anos (HOFFBRAND; MOSS, 2015). Esta doença afeta não só pessoas no Brasil como também é relevante em outros países, como na Alemanha. A Figura 5 apresenta um quadro comparativo entre a prevalência de diversos tipos de cânceres no Brasil e na Alemanha. Neste comparativo é possível identificar que a LMA foi o 12º tipo de câncer em número de vítimas no Brasil e o 7º na Alemanha (Figura 5). A LMA é geralmente uma doença agressiva e definida como um distúrbio clonal causado pela transformação maligna de células-tronco hematopoiéticas ou progenitores iniciais (HOFFBRAND; MOSS, 2015). Seu desenvolvimento foi associado a uma variedade de síndromes herdadas, adquiridas e familiares que resultam de desequilíbrios ou instabilidades cromossômicas, defeitos no reparo do DNA, alteração do receptor de citocinas ou ativação da via de transdução de sinal e síntese protéica alterada (WEST *et al.*, 2014).

Outro ponto importante foi a pesquisa para a descoberta de alterações genéticas específicas que tem permitido o desenvolvimento de terapias mais eficientes aos subtipos da doença. Na década de 1970, a LMA foi dividida em subtipos com base nos critérios morfológicos e na maturidade celular (WALTER *et al.*, 2013). Essa classificação conhecida como FAB (*French-American-British*) subdividiu a LMA em oito níveis, de M0 a M7 (Tabela 1) (WALTER *et al.*, 2013).

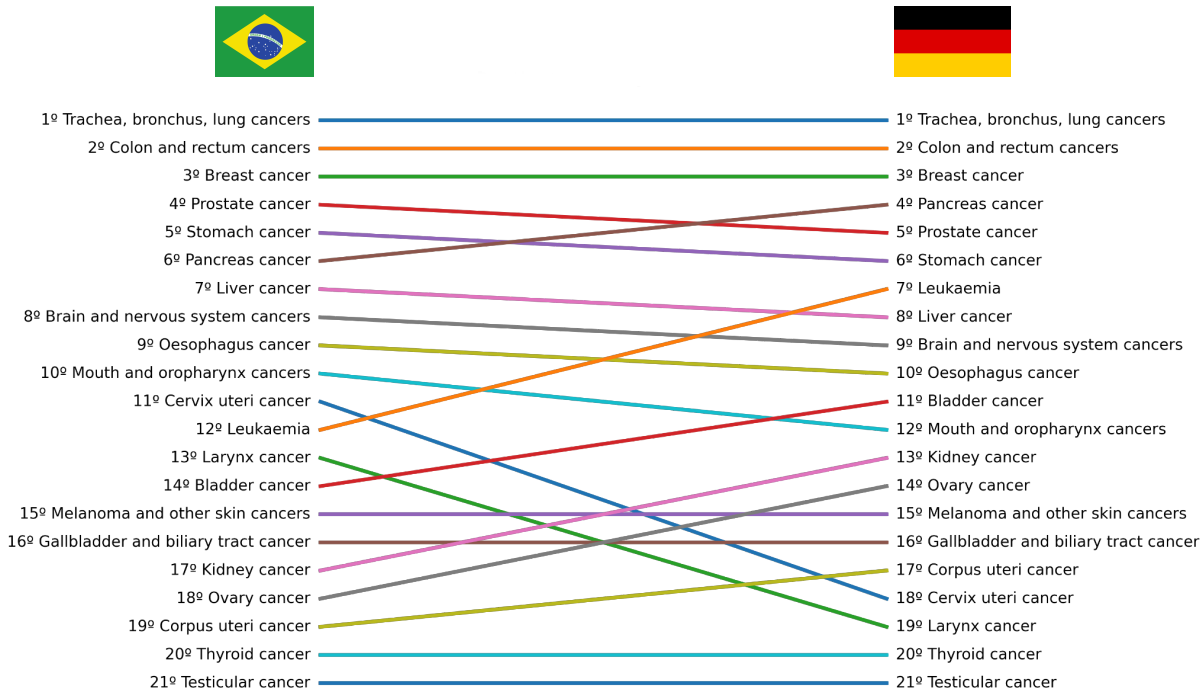

*Figura 5**. Rank dos diferentes tipos de câncer ordenados pelo número de mortes entre Brasil e Alemanha em 2019.** No ano anterior ao início da pandemia, 2019, as leucemias representaram a 12ª maior causa de mortes por câncer no Brasil. Por outro lado, em países como a Alemanha, ela estava em 7º, evidenciando que comumente alguns tipos tumorais são mais prevalentes em alguns países em relação a outros, enquanto que os cânceres de pulmão, cólon e mama, respectivamente, são os três mais comuns em ambos os países. Fonte: GHO OData API data. Global Health Estimates 2019: Deaths by Cause, by Country and by Region. Geneva, World Health Organization; 2019.*

Atualmente, a classificação das LMA se baseia em características adicionais, além das utilizadas anteriormente, de acordo com a Classificação de Tumores Hematopoiéticos e Tecidos Linfóides (ARBER *et al.*, 2022), sendo elas: clínicas, imunofenotípicas e genéticas. Embora essa seja a mais recente, a classificação a ser utilizada como atributo alvo de nossos modelos experimentais será a FAB, uma vez que tanto os dados do TCGA e TARGET (ambos citado nos agradecimentos como requerido pelos autores), quanto do Hemocentro da UNICAMP a utilizam.

Tabela 1. Classificação da Leucemia Mieloide Aguda segundo proposto pela FAB (BENNETT et al., 1976).

Subtipo	Descrição
M0	Leucemia Aguda Mieloblástica indiferenciada
M1	Leucemia Aguda Mieloblástica com maturação mínima
M2	Leucemia Aguda Mieloblástica com maturação
M3	Leucemia Aguda Promielocítica (APL)
M4	Leucemia Aguda Mielomonocítica (com e sem eosinofilia)
M5	Leucemia Aguda Monocítica
M6	Leucemia Aguda Eritróide
M7	Leucemia Aguda Megacarioblástica

2.1.1.2. Síndrome Mielodisplásica

A Síndrome Mielodisplásica (SMD) inclui um grupo de distúrbios clonais das células-tronco hematopoiéticas que levam à insuficiência da medula óssea e baixa contagem de células sanguíneas (REILLY *et al.*, 2019). Uma característica da doença é o aumento da proliferação e apoptose das células hematopoiéticas (hematopoiese ineficaz), levando ao paradoxo de uma medula óssea hipercelular, mas à pancitopenia no sangue periférico. Há uma tendência à progressão da doença para a LMA (LOPES *et al.*, 2017).

Na SMD, algumas células da medula óssea são displásicas (anormais) e apresentam problemas na produção de novas células sanguíneas. Diferentes tipos de células podem ser afetadas, embora o achado mais comum na SMD seja a anemia (HOFFBRAND; MOSS, 2015). A análise molecular mostra que a SMD comumente carrega mutações em dois ou três sítios, as quais são condutoras geralmente acometendo genes envolvidos em processos epigenéticos, metilação do DNA e modificação da cromatina (REILLY *et al.*, 2019). A SMD foi recentemente classificada pela em 2022 com base no hemograma, aparência morfológica e número de células blásticas no sangue ou medula óssea, além da análise citogenética (ARBER *et al.*, 2022). Entretanto, da mesma forma que ocorreu com os dados de LMA os dados obtidos de nossos parceiros em 2021 contemplava a classificação de 2016, na época a mais recente, sendo esta a utilizada no presente trabalho (ARBER *et al.*, 2016). Nesta classificação de 2016 os subtipos foram agrupados em oito tipos principais de SMD, os quais foram listados na Tabela 2. Outra forma de estratificar os pacientes de SMD é utilizando um índice de risco de

prognóstico, conhecido pela comunidade como IPSS-R (Revised International Prognostic Scoring System). Este índice é definido por cinco níveis de risco: Muito Baixo, Baixo, Intermediário, Alto e Muito Alto (GREENBERG *et al.*, 2012).

O IPSS-R tem sua importância na clínica médica pois ajuda na criação de estratégias de tratamento, especialmente para pacientes de baixo ou alto risco (os extremos). Os pacientes classificados como de risco intermediário apresentam uma dúvida clínica sobre se devem receber um tratamento de alto ou baixo risco. Isso constitui um problema (LIAPIS *et al.*, 2021). Portanto, determinar a probabilidade dos pacientes do grupo intermediário serem também classificados como de baixo ou alto risco, bem como os fatores que nos permitiriam definir isso é de vital importância. O IPSS-R é composto por uma equação, a qual leva em consideração para o cálculo do risco certos fatores de doença: citogenética, contagem de blastos/plaquetas, contagem absoluta de neutrófilos e hemoglobina (GREENBERG *et al.*, 2012). Vale ressaltar que a análise citogenética necessária em ambos métodos de classificação de SMD é custosa e não é sempre acessível a todos em todos os centros hospitalares nacionais.

Tabela 2. Classificação da Síndrome Mielodisplásica modificado de OMS 2016.

Subtipo	Descrição
RCUD	Citopenia refratária com displasia unilinhagem (Anemia refratária (RA), Neutropenia refratária (RN), Trombocitopenia refratária (RT))
RCUD-RS	RCUD com sideroblastos em anel
RARS	Anemia refratária com sideroblastos em anel
RCMD	Citopenia refratária com displasia multilinhagem
RCMD-RS	RCMD com sideroblastos em anel
RAEB-1	Anemia refratária com excesso de blastos tipo 1
RAEB-2	Anemia refratária com excesso de blastos tipo 2
Del5q	Síndromes mielodisplásicas associadas com isolada del(5q)

2.1.2. Os Bancos de Dados Biológicos

O termo "banco de dados" refere-se a uma coleção de informações estruturadas e armazenadas eletronicamente. Eles funcionam como repositórios digitais, permitindo a gestão, acesso e armazenamento de grandes quantidades de

informações. Para garantir a integridade e a segurança dos dados, eles permitem consultas, atualizações e outras operações.

O Instituto Nacional do Câncer dos Estados Unidos (NCI, *National Cancer Institute*) hospeda uma grande quantidade de dados sobre diversos tipos de câncer, através do *GDC Data Portal*. Segundo a política de proteção dos dados, e manutenção do anonimato dos pacientes, quaisquer dados que possam servir para rastreá-los, ou revelar sua identidade são protegidos em ambos os bancos sob os níveis 1 e 2. Por exemplo, os dados brutos de alinhamento de *reads* (no formato BAM) são dados protegidos. Independente do nível do dado, os pacientes são anonimizados para garantir a política de proteção de identidade. O *GDC Data Portal* possui um total de 39 tumores humanos, vindos de 12 projetos. No total, aproximadamente 193 mil arquivos possuem acesso livre (nível 3). Dentre eles estão os dados disponibilizados de projetos, como o do projeto da investigação terapeuticamente aplicável para gerar tratamentos eficazes (TARGET, *The Therapeutically Applicable Research to Generate Effective Treatments*) e do Atlas do Genoma Humano (TCGA, *The Cancer Genome Atlas*) estão disponíveis. As referências de ambos projetos encontram-se na parte de agradecimentos desta tese, conforme requerido pelos autores.

O câncer pediátrico é uma doença rara, mas fatal para crianças e jovens adultos. Muitos tipos da doença ainda não foram compreendidos e não há tratamento eficaz para eles. A iniciativa TARGET foi fundada em 2007 para abordar essas questões, identificando mudanças genéticas em cânceres pediátricos difíceis de tratar por meio de caracterização molecular abrangente. Inicialmente a iniciativa concentrou-se em subtipos de leucemia linfoblástica aguda e neuroblastoma, entretanto posteriormente passou a estudar também a LMA, osteossarcoma e tumores renais. A aceleração das descobertas moleculares e a transferência imediata dos resultados para a clínica dependeu da cooperação entre pesquisadores e médicos, especialmente do Children's Oncology Group. A comunidade de pesquisa pode usar os dados criados para criar novas terapias e melhorar significativamente os resultados de alguns pacientes pediátricos.

O TCGA por sua vez foi uma iniciativa do NCI e do Instituto Nacional de Pesquisa do Genoma Humano (NHGRI, *National Human Genome Research Institute*), ambos filiados ao Instituto Nacional de Saúde (NIH*, National Institutes of Health*). Utilizando técnicas como sequenciamento em larga escala de RNA, o

projeto visou compreender as bases moleculares do câncer (NIH, 2019). Durante os dez anos de sua existência (2006 - 2016) o projeto TCGA gerou 2,5 Petabytes de dados, descrevendo 33 tumores, sendo 10 raros, baseados em tecidos normais e tumorais de 11 mil pacientes. Se esses dados fossem gravados em DVDs, a quantidade necessária de discos seria equivalente a uma pilha de 530 metros de altura, ou aproximadamente equivalente a um prédio de 177 andares.

Um outro relevante portal de dados é o cBioPortal (CERAMI *et al.*, 2012), o qual foi projetado para pesquisadores e clínicos integrando dados de várias fontes, e de diversos tipos de dados, como dados genômicos, e clínicos. A plataforma conta com a adição contínua de novos dados e funcionalidades, mantendo a plataforma atualizada com os datasets das últimas publicações envolvendo ômicas. Dentre os dados presentes na plataforma estão os dados de SMD (BERNARD *et al.*, 2022) gerados pelo Grupo de Trabalho Internacional para o Prognóstico na SMD (IWG-PM). O IWG-PM é composto por um grupo de investigadores internacionais alinhados pela *MDS Foundation*. Seu objetivo é definir as características clínicas e biológicas da SMD, fornecendo as bases para compreender a natureza e o potencial de progressão deste espectro de doenças. Para atingir esse objetivo, o grupo desenvolveu iniciativas que conduziram a projetos e publicações cruciais para caracterizar e classificar a doença. Atualmente, o grupo está realizando pesquisas para determinar os efeitos das características mutacionais, que revelam mais sobre o estado da doença, e para identificar novos alvos terapêuticos para a SMD.

2.2. Objetivos

2.2.1. Objetivo Geral

Ajustar modelos preditivos para os subtipos de LMA e de SMD a fim de avaliar o potencial destes modelos na digitalização do conhecimento médico estabelecido.

2.2.2. Objetivos Específicos

1. Adquirir e preparar os dados padronizando as características comuns;
2. Treinar algoritmos de aprendizado de máquina utilizando dados clínicos de pacientes diagnosticados com LMA ou SMD;
3. Selecionar dois melhores modelos um para LMA e outro para SMD com base nos valores das métricas de avaliação, como acurácia balanceada e AUC-ROC;
4. Descrever e comparar diferentes abordagens para avaliar a importância dos preditores utilizados no ajuste dos modelos preditivos selecionados.

2.3. Materiais E Métodos

A máquina em que as análises foram geradas possui a seguinte configuração: CPU AMD Ryzen 7 3800X 8-Core Processor × 16 threads, 32 GiB de memória RAM, Sistema operacional Linux Mint 21.3 Cinnamon Linux Kernel 6.5.0-41-generic. A versão Python utilizada foi a 3.8.16, e os pacotes foram gerenciados pelo Conda 23.1.0. Uma visão geral dos métodos empregados no presente trabalho tanto para LMA quanto para SMD é exibida na Figura 6.

O presente trabalho valeu-se do uso de banco de dados públicos, os quais contém informações clínicas de pacientes de SMD e LMA de diferentes trabalhos publicados. Os bancos de dados usados foram os seguintes: TCGA, TARGET (hospedados no GDC Data Portal), e MDS-IWG (BERNARD *et al.*, 2022) (hospedado no cBioPortal). As colaborações estabelecidas durante o presente trabalho com grupos de pesquisa clínica foram fundamentais para aquisição de um

maior volume de dados. O grupo da Profª. Drª. Sara Saad (Hemocentro UNICAMP) nos possibilitou utilizar dados clínicos anonimizados para LMA e SMD. Já o grupo do Prof. Dr. Ronald Feitosa Pinheiro da Universidade Federal do Ceará (UFC) nos forneceu dados de histórico médico de pacientes de SMD anonimizados.

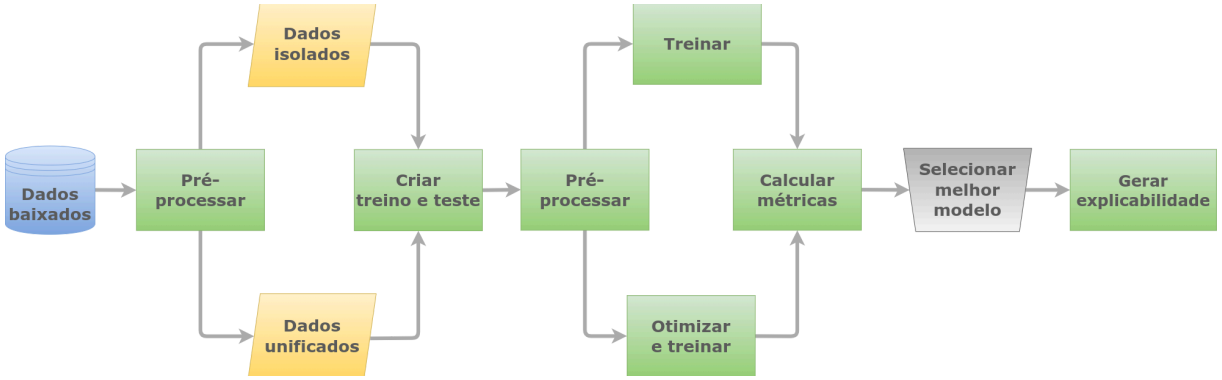

Figura 6. ***Organograma geral dos métodos utilizados no presente trabalho.*** *Em azul a entrada de dados, em verde está representado ações e processos executados pelo sistema, em amarelo a saída e entrada simultânea de dados, por fim em cinza está representando uma entrada manual feita por uma pessoa. Imagem elaborada de acordo com ISO 5807:1985 (Information processing - Documentation symbols and conventions for data, program and system flowcharts, program network charts and system resources charts).*

2.3.1. Aquisição dos dados

A aquisição dos dados públicos hospedados pelo *GDC Data Portal* foi realizada utilizando o pacote R desenvolvido durante o mestrado do presente autor, denominado DOAGDC e disponível em https://github.com/Facottons/DOAGDC. Na página do pacote é possível encontrar instruções detalhadas de como utilizar as diferentes funções oferecidas. A obtenção dos dados é realizada pela função denominada `download_gdc`. De forma sucinta, os parâmetros necessários são: tipo de dado, código do tumor, abreviação de um dos dois bancos de dados disponíveis e diretório de trabalho desejado. Alguns parâmetros são específicos dos tipos de dados. Por exemplo, para a aquisição dos valores de expressão gênicas existe um parâmetro necessário, o qual informa o tipo de normalização dos dados, caso necessário. O código para a aquisição dos dados clínicos do tumor LMA,

hospedados no *GDC Data Portal* será: `download_gdc("clinical_supplement", "LAML", "gdc", "~/Desktop")`.

Os dados de SMD foram obtidos de estudos publicamente hospedados no cBioPortal utilizando a ferramenta de download oferecida pela própria plataforma. Além das plataformas citadas anteriormente, dados do Hemocentro da UNICAMP para os dois tumores foram disponibilizados pela equipe da Profª. Drª. Sara Saad, do Hemocentro da UNICAMP. Por sua vez, dados adicionais de SMD foram fornecidos pelo grupo do Prof. Dr. Ronald Feitosa Pinheiro. Os dados de cada paciente são compostos de atributos físicos (por exemplo, idade de diagnóstico e sexo) e laboratoriais (por exemplo, a contagem de blastos e eritrócitos).

A Tabela 3 reúne as dimensões dos conjuntos de dados antes e depois de processados, enquanto a última coluna informa a composição das matrizes após realizar a operação de unificação dos bancos. Nos conjuntos de dados disponíveis existem diferentes preditores, sua descrição, bem como uma análise exploratória deles se encontram nos notebooks python hospedados no repositório deste trabalho: https://github.com/fpatroni/defesa_doutorado.

Tabela 3. Relação de pacientes e atributos obtidos brutos e após o pré-processamento.

Câncer	Origem	Bruto		Pré-processado		Total	
		Pacientes	Atributos	Pacientes	Atributos	Pacientes	Atributos comuns
LMA	TCGA	200	88	178	19	1023	4
	TARGET	833	65	789	5		
	Hemocentro	129	58	67	13		
SMD		89	26	69	5	1943	5
	UFCeará	151	15	117	6		
	MDS-IWG	3323	38	1862	6		

Os conjuntos de dados obtidos foram utilizados em oito cenários experimentais para analisar qual combinação gerava os melhores resultados. Para os dados de LMA quatro cenários, apenas dados do TCGA, apenas dados do TARGET, apenas dados do Hemocentro e com o conjunto resultante da intersecção dos atributos entre eles. Este apresenta mais pacientes e menos atributos, enquanto os outros ensaios possuem menor número de pacientes e maior número de

variáveis. Os outros quatro ensaios restantes foram com os dados de SMD, dois com os dados dos Hemocentros parceiros, um com os dados do IWG (cbioportal) e o último com a unificação destes dados (Tabela 3). Um 9º cenário experimental foi adicionado durante o desenvolvimento do presente trabalho. Neste cenário, a resposta de interesse é um índice disponível apenas para SMD, o IPSS-R. Para tal, este cenário contou com o dataset resultante da intersecção entre os dados de SMD utilizados no presente trabalho.

2.3.2. A biblioteca de ML

O desenvolvimento dos scripts utilizados neste projeto foram baseados na linguagem de programação Python. Esta linguagem tem sido amplamente utilizada na área de análise de dados e ML. Dos 17 modelos de classificação utilizados neste trabalho, 14 são do Scikit-Learn (PEDREGOSA *et al.*, 2012), o qual é um importante módulo escrito em Python para o aprendizado de máquina. Ele foi lançado em 2010 e reúne algoritmos para o aprendizado supervisionado, não-supervisionado, além de métodos para pré-processamento dos dados e para avaliação dos modelos treinados (PEDREGOSA *et al.*, 2012). Um de seus pontos fortes é a pequena curva de aprendizagem e a vasta documentação, por outro lado o não uso de GPU para agilizar o processo de treinamento dos modelos é um de seus maiores pontos fracos. O Scikit-Learn conquistou seu espaço na comunidade em detrimento de outras alternativas similares, código-livre e/ou pagas, por utilizar o ecossistema científico do Python, composto por bibliotecas famosas na área, como Numpy, Pandas e Scipy (PEDREGOSA *et al.*, 2012).

Os três modelos não implementados pelo Scikit-Learn são oriundos de três fontes distintas. Todos são métodos ensemble de árvores de decisão, sendo eles o CatBoost (PROKHORENKOVA *et al.*, 2019), o LightGBM (KE *et al.*, 2017) e o XGBoost (CHEN; GUESTRIN, 2016).

2.3.3. Pré-processamento dos dados

O pré-processamento realizado possui dois objetivos: preparar os dados para as análises de ML e adequar as colunas para a posterior união das matrizes. Resumidamente, ele foi composto de dez etapas:

1) Padronizar os separadores decimais para ponto (“.”);
2) Ajustar nome das colunas dos conjuntos de dados;
3) Remover as variáveis sem utilidade, como IDs;
4) Eliminar colunas (atributos) com mais de 10% de valores faltantes;
5) Eliminar todas as linhas (pacientes) com valores faltantes;
6) Remover os subtipos com menos de 10 pacientes (ao menos dois pacientes no conjunto de teste, isso é 20% dos dados);
7) Realizar a codificação de variáveis categóricas e dos alvos (target encode);
8) Separar os dados em conjunto de treino (80%) e de teste (20%), outro valor também utilizado é o 70-30 quando se possui uma maior quantidade de dados;
9) Excluir colunas com baixa variância;
10) Checar e extinguir variáveis com alta colinearidade, ou multicolinearidade;
11) Normalizar os dados usando *z-score* e escala entre zero e um para *Multinomial Naive Bayes* e um segundo modelo de *Multi-layer Perceptron*.

A diferença de unidades e abreviações das colunas inviabilizou um procedimento automatizado de unificação dos conjuntos de dados, provenientes de fontes distintas. Assim, para transformá-los em um conjunto de dados coesos, fez-se necessária a harmonização manual, com a seleção de variáveis relevantes e em comum entre as fontes, tanto para os dados da LMA quanto para os dados da SMD. Outra importância dessa etapa de seleção de atributos é evitar o evento denominado “vazamento de dados” (*Data Leakage*). Este problema ocorre quando dentre as variáveis preditoras existe uma variável, a qual sozinha é capaz de predizer o alvo além de ter uma relação íntima e direta com ele. Por exemplo, em um conjunto de dados usado para treinar um modelo para prever se um paciente tem diabetes, a presença de uma coluna "índice glicêmico em jejum" pode causar vazamento de dados. Isso acontece porque há uma relação conhecida entre diabetes e o índice glicêmico em jejum: pacientes com diabetes geralmente têm um valor igual ou superior a 126 mg/dL.

Outro cuidado durante o pré-processamento é lidar com possíveis valores faltantes. Os dados de LMA e SMD apresentam esses valores, os quais não se pode confirmar sua natureza aleatória e, portanto, foram removidos. Para cada caso, foi analisado em qual dimensão deveria ser realizada a remoção do paciente ou da variável clínica a fim de minimizar as perdas.

As variáveis categóricas restantes foram codificadas para valores numéricos (*feature engineering*), uma vez que os modelos utilizados não aceitam valores de texto (*strings*) como atributos (*features*). A função `OrdinalEncoder()` foi empregada quando os valores eram ordinais, e `LabelEncoder()` quando binárias, a qual substitui os valores categóricos em 0 e 1. Após essa etapa, os atributos restantes com alta colinearidade entre si (> 0.7 point-biserial, pearson, spearman ou kendall) foram removidos ou usados para gerar outra variável em seus lugares (*feature extraction* ou *feature synthesis*) afim de evitar redundância e instabilidade nos modelos. Depois, um filtro de variância foi utilizado, no qual todos os atributos com baixa variância foram removidos utilizando a função `VarianceThreshold()` com corte sendo 0,1.

Os conjuntos de dados resultantes das etapas anteriores foram individualmente separados em treino e teste, na proporção 4:1, respectivamente. Logo, do total de pacientes nos bancos públicos, 80% serão escolhidos aleatoriamente (com estratificação categórica) formando o conjunto de treino, enquanto o restante será utilizado para testar o modelo. A estratificação categórica é importante nesta etapa, pois ela certifica que a mesma proporção entre as classes será mantida no conjunto de treino e de teste, a fim de evitar uma situação, na qual o modelo tenha que predizer um alvo para o qual ele não foi treinado previamente.

Os atributos não dicotômicos (binários) foram então normalizados utilizando o `StandardScaler()`, dessa forma cada uma das variáveis apresentou média zero e variância unitária. Entretanto, essa normalização gera valores negativos não sendo adequados para estimadores como o *Multinomial Naive Bayes*. Portanto, uma outra normalização também foi realizada nos atributos contínuos com a função `MinMaxScaler()`. Essa normalização escala todos os valores de cada variável entre um valor mínimo e máximo, neste caso, zero e um. Dessa forma, valores negativos não são gerados atendendo o requerimento de alguns modelos. Essa última normalização foi utilizada exclusivamente para o modelos de *Multinomial*

Naive Bayes e um segundo modelo de *Multi-layer Perceptron*, o qual também foi treinado com a primeira normalização apresentada.

Nesta etapa é importante transformar os dados de treino e teste utilizando apenas o escalador calculado com os dados de treino, uma vez que não se deve assumir o conhecimento da média e do desvio padrão do conjunto teste. Além do ganho de velocidade para execução de alguns algoritmos, as discrepâncias de escalas entre as variáveis e a presença de *outliers* não impactam a qualidade das predições.

2.3.4. Seleção e validação dos modelos

Os modelos padrões e com alguns de seus hiperparâmetros otimizados foram testados quanto à sua performance e comparados, a fim de selecionar os melhores para cada uma das experimentações. Essa etapa se faz necessária, pois, nenhum algoritmo é a priori o melhor. A otimização de hiperparâmetros foi realizada com validação cruzada (5-*fold cross-validation*), para cada um dos modelos de classificação utilizando a função `RandomizedSearchCV()`, a qual além de ser mais rápida que sua contra parte, `GridSearchCV()`, testa combinações de valores de hiperparâmetros de forma aleatória dentro de uma distribuição. A otimização de hiperparâmetros tende a gerar um algoritmo menos complexo evitando sobreajuste (*overfitting*), além de generalizar o máximo possível para predições mais acuradas, quando um novo conjunto de dados for apresentado (na fase de teste do modelo). Ao todo 17 modelos de classificação foram empregados (Tabela 4).

Tabela 4. Modelos de classificação utilizados com suas respectivas siglas utilizadas no trabalho.

Nome do Modelo	Tipo	Descrição
AdaBoost (ADA)	Ensemble	Método de boosting adaptativo que combina múltiplos classificadores fracos para construir um único classificador mais robusto (FREUND; SCHAPIRE, 1995).
Árvore de Decisão (DTR)	Decision Trees	Modelo que divide os dados em subconjuntos baseados em features, construindo uma estrutura em forma de árvore que representa as decisões tomadas para chegar a um resultado final (BREIMAN *et al.*, 1984).

CatBoost (CAT)	Ensemble	Algoritmo de boosting de gradiente que lida automaticamente com variáveis categóricas.
Gaussian Naive Bayes (GNB)	Naive Bayes	Classificador probabilístico que assume independência entre as features.
Gaussian Processes (GP)	Gaussian Processes	Método não-paramétrico baseado em processos gaussianos para problemas de regressão e classificação.
Gradient Boosting (GB)	Ensemble	Técnica de boosting que cria um modelo preditivo forte a partir da combinação de vários modelos fracos.
k-Nearest Neighbors (KNN)	Nearest Neighbors	Algoritmo que classifica amostras com base nas k amostras de treinamento mais próximas.
LightGBM (LGB)	Ensemble	Algoritmo de boosting de gradiente baseado em histogramas, otimizado para desempenho e eficiência.
Multi-layer Perceptron (MLP)	Neural Network	Rede neural feedforward que consiste em múltiplas camadas de perceptrons.
Multi-layer Perceptron (MLP2)	Neural Network	MLP com normalização dos dados entre zero e um, melhorando a performance em certos casos.
Multinomial Naive Bayes (MBN)	Naive Bayes	Variante do Naive Bayes para dados distribuídos multinomialmente, como texto.
Quadratic Discriminant Analysis (QDA)	Discriminant Analysis	Método de análise discriminante que assume diferentes matrizes de covariância para cada classe.
Random Forest (RF)	Ensemble	Conjunto de árvores de decisão treinadas com diferentes subconjuntos de dados para melhorar a robustez.
Regressão Logística (LR)	Linear Models	Modelo de regressão utilizado para prever a probabilidade de uma classe binária.
Stochastic Gradient Descent (SGD)	Linear Models	Classificador que utiliza gradiente descendente estocástico para otimização de funções convexas.
Support Vector Machine (SVM)	Support Vector Machines	Algoritmo de classificação que encontra o hiperplano que melhor separa as classes.
XGBoost (XGB)	Ensemble	Implementação eficiente de boosting de gradiente que melhora a velocidade e a performance.

A próxima etapa se deu com a fase de teste, os algoritmos foram ranqueados utilizando as métricas de avaliação dos modelos na seguinte ordem: acurácia, valor da área sob a curva ROC e macro-média F1-score. A avaliação também foi auxiliada pela geração de imagens como matrizes de confusão e curvas

ROC. As seguintes métricas foram utilizadas para avaliar o desempenho dos modelos após treinamento:

1. Acurácia balanceada (Balanced *Accuracy*):
    a. Calcula a média da taxa de acertos (recall) para cada classe individualmente, e então tira a média dessas taxas. Ajuda a identificar modelos que são robustos e que não estão simplesmente super ajustados à classe majoritária. Os valores variam de 0 a 1, sendo 1 o melhor valor possível;
    b. $Balanced_acc = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$;
2. Área sob a curva ROC (*Area Under the Receiver Operating Characteristic Curve*):
    a. Avalia a capacidade do modelo de distinguir entre classes, plotando a taxa de verdadeiros positivos (Sensitividade/*True Positive Rate*(TPR)) contra a taxa de falsos positivos (1 - Especificidade ou *False Positive Rate* (FPR)) para diferentes limiares de decisão. É uma métrica robusta para conjuntos de dados desbalanceados, pois não depende da distribuição de classes. A área sob essa curva fornece uma métrica agregada do desempenho do modelo em todos os limiares possíveis. Ela retorna valores variando de 0 a 1, sendo 1 perfeita discriminação, e o valor de 0.5 evidencia ausência de capacidade discriminativa (equivalente a uma escolha aleatória);
    b. $TPR = \frac{TP}{TP+FN}$,

    $FPR = \frac{FP}{FP+TN}$;
3. F1-score macro-média (*Macro-averaged F1 score*):
    a. Calcula a média harmônica da precisão e da revocação, fornecendo um único número que equilibra esses dois aspectos. A Macro-Média permite que esse equilíbrio seja mantido em todas as classes. Adequado para avaliar desempenho quando as classes têm importância igual, independentemente de seu tamanho. Os valores variam de 0 a 1, sendo 1 o melhor valor possível;

   b. $F1 = 2 \times \left(\frac{Precisão \times Revocação}{Precisão + Revocação}\right)$;

4. F1-*score* ponderado (*Weighted average of the F1-score*):
   a. Calcula a média ponderada do F1-score levando em conta o tamanho de cada classe ao calcular o F1-score global. Classes com mais exemplos contribuem mais para a métrica final, refletindo a importância relativa de cada classe no conjunto de dados. Isso contrasta com a macro-média, a qual trata todas as classes igualmente, independentemente de seu tamanho. Os valores variam de 0 a 1, sendo 1 o melhor valor possível;
   b. $F1Weighted = \frac{\sum_{i=1}^{N}\left(F1_i \times n_i\right)}{\sum_{i=1}^{N} n_i}$;
5. Perda Média de Hamming (*Average Hamming loss*):
   a. Mede a proporção média de rótulos incorretamente previstos em relação ao total de rótulos. Os valores variam de 0 a 1, e diferentemente das métricas até então descritas, um valor de Perda Média de Hamming próximo de zero indica que o modelo está fazendo poucas previsões incorretas, enquanto valores mais altos indicam um maior número de erros;
   b. $Hamming\ Loss = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L}\sum_{j=1}^{L} I\left(y_{ij} \neq \widehat{y_{ij}}\right)$;
6. Coeficiente de Similaridade de Jaccard (*Jaccard similarity coefficient score)*:
   a. Mede a similaridade entre conjuntos de amostras, tendo a interseção dividida pela união dos conjuntos. Em termos de classificação, ele mede a proporção de rótulos corretamente previstos em relação ao número total de rótulos únicos presentes nas previsões e nos verdadeiros. Os valores variam de 0 a 1, onde 0 indica que não há interseção entre os conjuntos e 1 indica que os conjuntos são idênticos;
   b. $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$;
7. Perda de Classificação Zero-Um (*Zero-one classification loss*):

a. Mede a taxa de erro do modelo contando o número de previsões incorretas e dividindo pelo número total. Ela é menos informativa que outras métricas em problemas desbalanceados, como o do presente trabalho. Os valores variam de 0 a 1, e da mesma forma que ocorre com a Perda Média de Hamming, o valor resultante é inversamente proporcional ao desejado, sendo zero o melhor valor possível;

b. $ZeroOne\ Loss\ = \frac{1}{N}\sum_{i=1}^{N} I\left(y_{ij} \neq \widehat{y_{ij}}\right)$;

8. Precisão (*Precision*):
    a. Mede a proporção de verdadeiros positivos sobre todas as previsões positivas. Esta métrica é importante quando o custo de falsos positivos é alto. Os valores variam de 0 a 1, quanto mais alta a precisão, mais confiáveis são as previsões positivas;
    b. $Precision\ = \frac{TP}{TP + FP}$;
9. Revocação/Sensibilidade (*Recall*):
    a. Mede a proporção de verdadeiros positivos sobre todos os casos reais positivos. Desta vez, o foco está no custo de falsos negativos, sendo mais importante do que minimizar falsos positivos. A Revocação fornece uma medida de abrangência do modelo em relação às instâncias positivas. É uma métrica robusta mesmo em cenários onde há um desequilíbrio significativo entre as classes. Quanto mais próxima de 1, maior a capacidade do modelo de capturar todas as instâncias positivas;
    b. $Recall\ = \frac{TP}{TP + FN}$;

Os modelos treinados exibidos a seguir utilizaram os seguintes parâmetros. Para garantir reprodutibilidade dos modelos foi utilizado “seed” como `random_state = 2024`. O parâmetro `njobs = -1` foi empregado para que seja utilizado todos os processadores disponíveis na máquina. A otimização de hiperparâmetros contou com o valor de 5-fold de validação cruzada `n_cv=5`,

`scoring='f1_weighted'`, e com `n_iter=50` para a estratégia de busca aleatória `RandomizedSearchCV()`.

2.3.5. Interpretabilidade e Explicabilidade

O grande objetivo da Interpretabilidade e Explicabilidade de um modelo é garantir a confiabilidade e reduzir dúvidas de funcionamento. Diferentemente da crença popular, modelos de aprendizado de máquina não são necessariamente caixas pretas, sendo possível interpretar e explicar seu funcionamento (RIBEIRO *et al.*, 2016).

Na área médica é vital que os profissionais da saúde saibam a causa de uma decisão de um modelo. Portanto, o modelo selecionado deve apresentar grau de interpretação suficiente para justificar as decisões do modelo, assim como antecipar as consequências dessas decisões (RIBEIRO *et al.*, 2016). Dessa forma, é possível que uma pessoa consiga predizer o resultado do modelo sabendo como ele manipula e interpreta os atributos (*features*) (AREOSA; TORGO, 2019; THE MYTHOS OF MODEL INTERPRETABILITY - ACM QUEUE, [*s. d.*]).

Os métodos de interpretabilidade podem ser subdivididos em quadrantes (Figura 7) sendo sua classificação a combinação de um par dos seguintes tipos de métodos: Específicos, Agnósticos, Locais e Globais. Como regra geral, métodos globais são menos acurados que os locais, porém podem explicar melhor as predições e refletir o comportamento global ou médio. Métodos específicos são usados apenas em modelos próprios, por exemplo o método "*XGBoost Feature Importances - get_fscore()*" é global e específico para avaliação de atributos de modelos XGBoost treinados (CHEN; GUESTRIN, 2016). Um exemplo de método específico e local seria o método "Saliency Maps", ele é em visão computacional e tem como funcionamento revelar a importância de uma região da imagem (pixels) para visão humana.

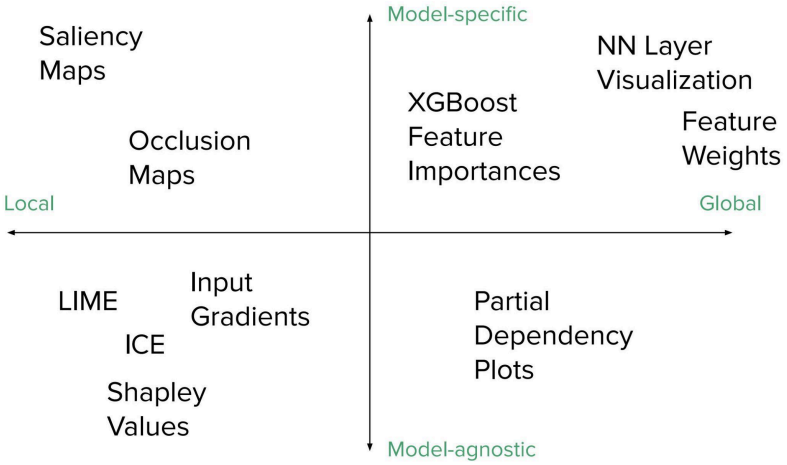

*Figura 7**. Os diferentes métodos de interpretabilidade separados em quadrantes.** Fonte: Adaptado de 2020-ComputeFest Harvard-IACS.*

Métodos agnósticos por sua vez podem ser usados com quaisquer modelos, sendo mais flexíveis. Seu método de funcionamento é baseado na manipulação de atributos de entrada e na subsequente avaliação de desempenho do modelo após essa mudança (EXPLAINABLE AND INTERPRETABLE MODELS IN COMPUTER VISION AND MACHINE LEARNING, [*s. d.*]). Caso o atributo seja importante para predição é esperado uma substancial queda ou aumento no rendimento do modelo, dependendo do método.

Os PDP (*Partial Dependence Plots*) (FRIEDMAN, 2001) são um exemplo de métodos agnósticos global e eles são capazes de expor como um atributo afeta a predição e quais variáveis mais afetam a predição. Para tal, um gráfico é produzido para demonstrar a relação entre um ou dois preditores (gráfico bi-variado) de interesse (eixo X) e a média das probabilidades das predições do modelo (eixo Y), para cada iteração (FRIEDMAN, 2001). É importante notar que para se ter confiabilidade nos resultados dos PDP é importante que haja independência entre a variável de interesse e as demais variáveis.

A mesma atenção deve ser levada em consideração ao utilizar a vertente local dos PDP, o ICE (*Individual Conditional Expectation*) (GOLDSTEIN *et al.*, 2015). O ICE funciona de forma similar aos PDP, entretanto ao invés de computar a média das predições do modelo em cada iteração, as predições são plotadas individualmente para cada observação (GOLDSTEIN *et al.*, 2015). Portanto, o ICE é um método local e pode ser usado para melhorar a informação global quando combinado com os PDP. Outra observação importante sobre os PDP e ICE é a

possibilidade de serem criadas observações improváveis, ou até mesmo impossíveis dada a natureza de funcionamento de ambos os métodos. Dois outros métodos agnósticos locais são: o LIME (*Local Interpretable Model-Agnostic Explanations*) (RIBEIRO *et al.*, 2016) e o SHAP (*SHapley Additive exPlanations*) (LUNDBERG; LEE, 2017).

O LIME utiliza modelos substitutos (lineares ou baseados em árvore de decisão) para explicar predições em relação a uma variável preditora de interesse. Esse método é capaz de gerar explicações usando apenas as variáveis locais mais importantes. A primeira etapa do LIME consiste em sintetizar dados aleatórios ao redor da variável desejada. Em seguida são calculadas as distâncias entre os dados aleatórios e as observações originais. O modelo então utiliza os dados sintetizados e realiza as predições, sendo testado o número mínimo de atributos preditores que otimiza o desempenho do modelo em relação às predições realizadas com os dados originais. Então, um modelo mais simples (geralmente linear) é treinado com os dados sintetizados dos atributos selecionados na etapa anterior, sendo comparado com as métricas do modelo treinado com os dados originais. Esta comparação final é utilizada como explicação local do modelo.

O SHAP utiliza um método da teoria dos jogos denominado Shapley values, proposto em 1953 (SHAPLEY, 1953). Para computar a importância de uma variável preditora Xi, são criados subconjuntos de atributos não contendo Xi. O próximo passo é computar o efeito nas predições após concatenar Xi a todos subconjuntos criados anteriormente. Embora esse método possa ser computacionalmente custoso, existem otimizações para certos modelos.

2.4. Resultados

Os notebooks com os códigos Python empregados em todo o processo realizado nas análises aqui descritas, os dados anonimizados utilizados e os resultados destas análises estão hospedados no repositório GitHub (https://github.com/fpatroni/defesa_doutorado). A lista de pacotes empregados com suas respectivas versões também se encontra no repositório citado anteriormente.

2.4.1. Aquisição e pré-processamento dos dados

O download e o pré-processamento dos dados foi realizado como descrito anteriormente. A etapa seguinte de pré-processamento contou com o ajuste de separadores decimais. Algumas colunas continham mais de um separador decimal, na mesma coluna, portanto o padrão de separadores decimais foi ajustado para ser o ponto final ".".

Visando a adequação das colunas para unificação, a tradução do nome das colunas foi realizada, uma vez que os dados fornecidos pelos colaboradores nacionais estavam em português e os dados dos bancos públicos em inglês. Além da tradução, os padroẽs de nomenclatura estavam diferentes, sendo necessário realizar uma padronização para se obter uma equivalência entre os conjuntos de dados clínicos. Outra padronização foi necessária por conta da diferença de unidades entre os bancos. Para exemplificar, no banco do TCGA os neutrófilos e linfócitos estavam em valores porcentuais, já no banco de dados do Hemocentro da UNICAMP estavam em contagens brutas. Portanto, a solução encontrada foi converter esses atributos para valores percentuais ao dividi-los pela contagem de leucócitos. Em outros casos, como para plaquetas, a unidade de G/L (gramas por litro) foi convertida para células/μL, sendo 1 G/L equivalente à $1x10^3$ células/μL (KRATZ *et al.*, 2004).

A penúltima etapa de pré-processamento envolve a checagem de variáveis com alta colinearidade, entre si. Para ilustrar esta etapa, utilizou-se o conjunto de dados de SMD do Hemocentro da Unicamp (Figura 8), no qual foi detectado a alta colinearidade entre "Neutrófilos" e "Leucócitos" (Figura 8A) e entre "Blastos MO (%)" e "IPSS" (Figura 8A). Posteriormente, as duas primeiras foram manipuladas para formar uma terceira variável chamada NLR (*neutrophil-to-leukocyte ratio*), técnica essa conhecida como Engenharia de atributos (*Feature engineering*) e a coluna "IPSS" removida (Figura 8B).

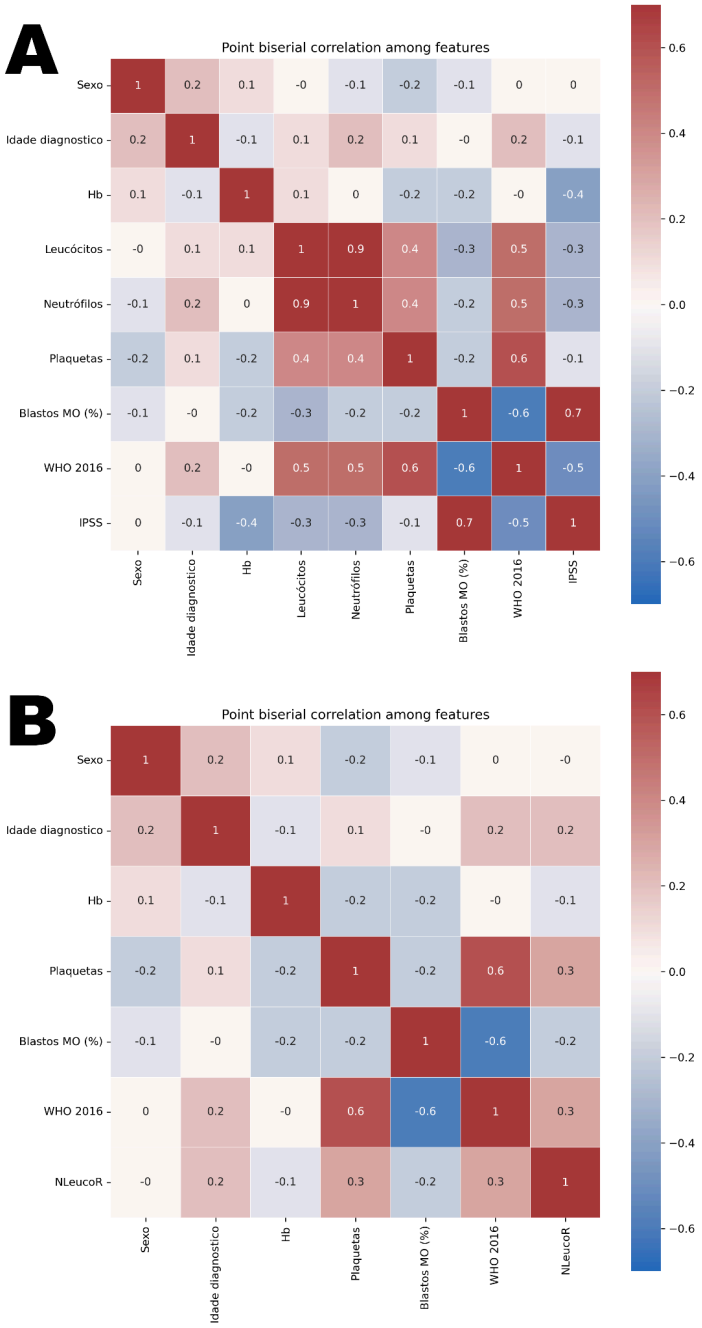

Figura 8. ***Mapa de calor da correlação de Point-biserial entre os atributos de SMD do Hemocentro.*** *A imagem ilustra a alta correlação entre as variáveis "Neutrófilos" e "Leucócitos" e também correlação alta entre "Blastos MO (%)" e "IPSS" (**A**). Após o pré-processamento de dados foi criado um novo atributo chamado NLR (neutrophil-to-leukocyte ratio) e a coluna "IPSS" foi removida (**B**).*

2.4.2. Treino e seleção de modelos

Ao todo foram testados 17 modelos de classificação, em 9 cenários experimentais diferentes. Três cenários utilizando o conjunto de dados unificados de

LMA e de SMD, sendo dois para SMD unificado (WHO 2016 e IPSS-R). Outros três cenários envolvem os dados de LMA (TCGA, Hemocentro UNICAMP e TARGET) e três os dados de SMD (Hemocentro UNICAMP, dados UFCeará e dados do MDS-IWG).

Em todos os cenários experimentais a seguir foram encontrados os mesmos padrões de desbalanceamento entre as classes alvos (resposta de interesse), como é possível de ser visualizado na Figura 9. Essa disparidade pode prejudicar o desempenho de modelos de machine learning, pois eles tendem a ser tendenciosos para a classe majoritária, negligenciando a classe minoritária. Existem técnicas para lidar com o desbalanceamento de dados como, o random sampling (BERGSTRA; BENGIO, 2012), o Tomek links (TOMEK, 1976), o SMOTE (CHAWLA *et al.*, 2002), e uma combinação de SMOTE e Tomek. Entretanto, o presente trabalho optou por contornar o problema de desbalanceamento de classes utilizando métricas que levam o peso e distribuição delas em conta durante os cálculos.

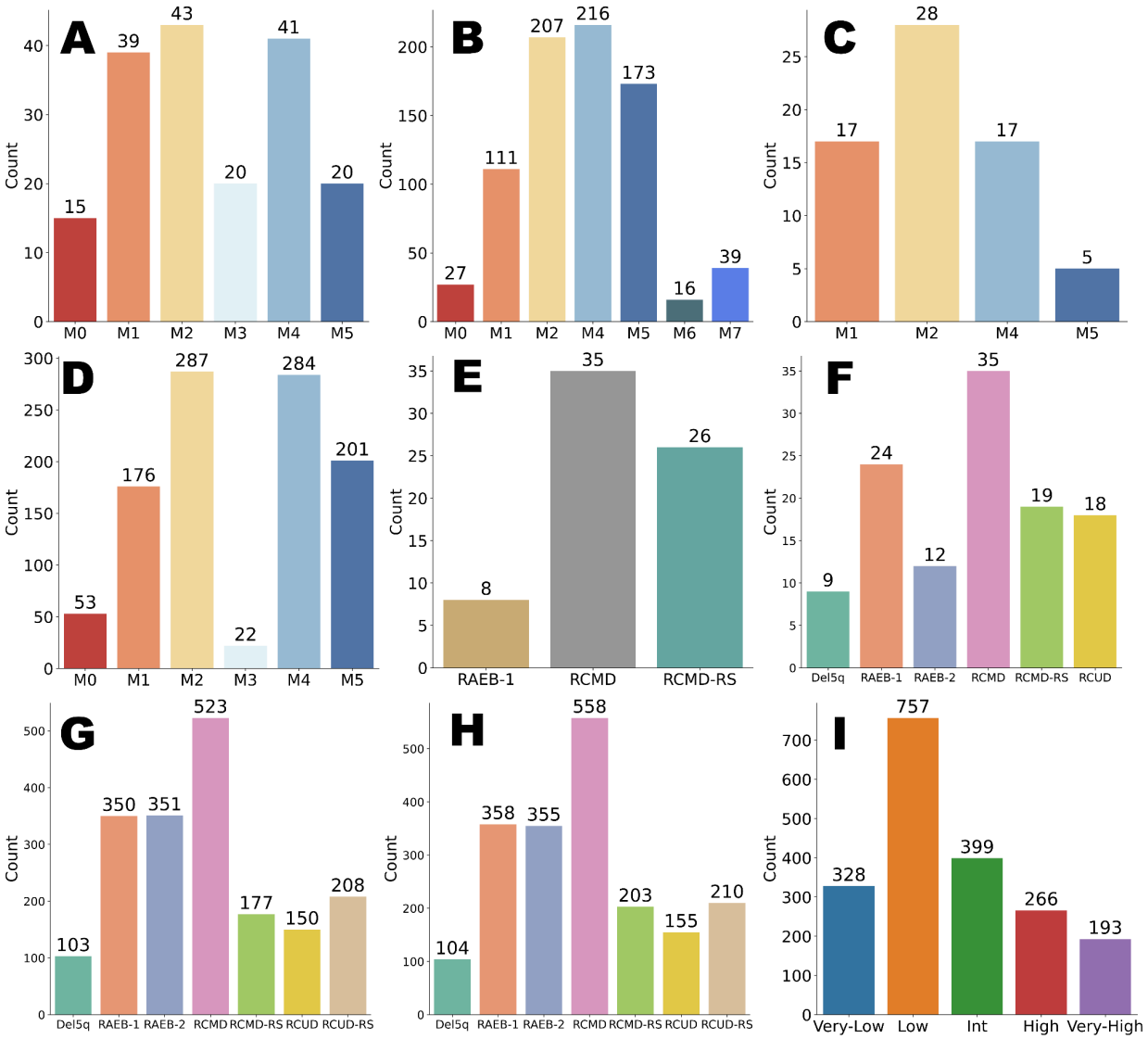

Figura 9. ***Distribuição das classes resposta.*** *Essa disparidade pode prejudicar o desempenho de modelos de machine learning, pois eles tendem a ser tendenciosos para a classe majoritária, negligenciando a classe minoritária. O eixo-y contém os valores de frequência de ocorrências, enquanto que o eixo-x representa os subtipos em cada cenário experimental. LMA TCGA* ***(A)****, LMA TARGET* ***(B)****, LMA Hemocentro* ***(C)****, LMA* ***(D)****, SMD Hemocentro* ***(E)****, SMD UFCeará* ***(F)****, SMD-IWG* ***(G)****, SMD reunido* ***(H)****, SMD reunido IPSS-R* ***(I)****.*

2.4.2.1. LMA

No Apêndice A podem ser encontradas as matrizes de confusão (Figuras S1-S8), e a tabela com todas as métricas utilizadas para os modelos treinados com os dados de LMA (Tabelas S1-S8).

2.4.2.1.1. Dados unificados

O conjunto unificado é composto por 1023 pacientes e 4 atributos (Tabela 3). As visualizações das distribuições dos atributos no conjunto de dados foram realizadas em relação a cada categoria de predição, ou seja, por subtipo de LMA segundo sistema de classificação FAB (Figura 10).

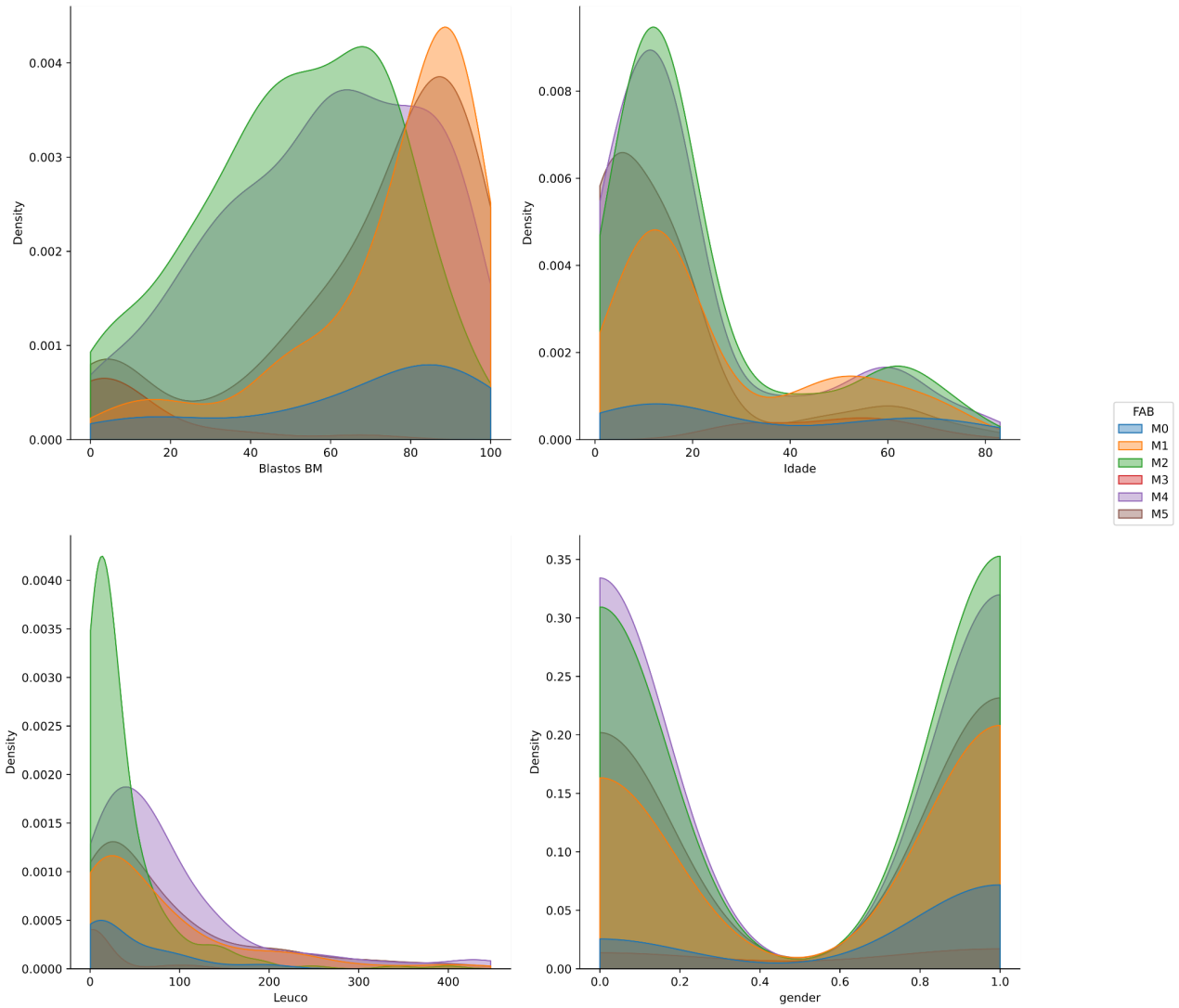

Figura 10. ***KDE plots com os valores de distribuição de cada um dos atributos utilizados no treinamento dos modelos.*** *Nesta visualização a distribuição de densidade de cada um dos atributos (antes da normalização) é separada por subtipo de LMA com o intuito de visualmente determinar se algum atributo isoladamente consegue distinguir os subtipos.*

Inicialmente, os modelos foram treinados sem otimização de hiperparâmetros e posteriormente foi realizada a otimização, a fim de testar e

quantificar o ganho da otimização de hiperparâmetros. Isso se fez importante, uma vez que a otimização de hiperparâmetros exige tempo de máquina e altos recursos computacionais (Figura 11). As flutuações no ranqueamento dos modelos antes e depois da otimização dos hiperparâmetros está presente nos notebooks python no repositório do GitHub mencionado pelo presente trabalho.

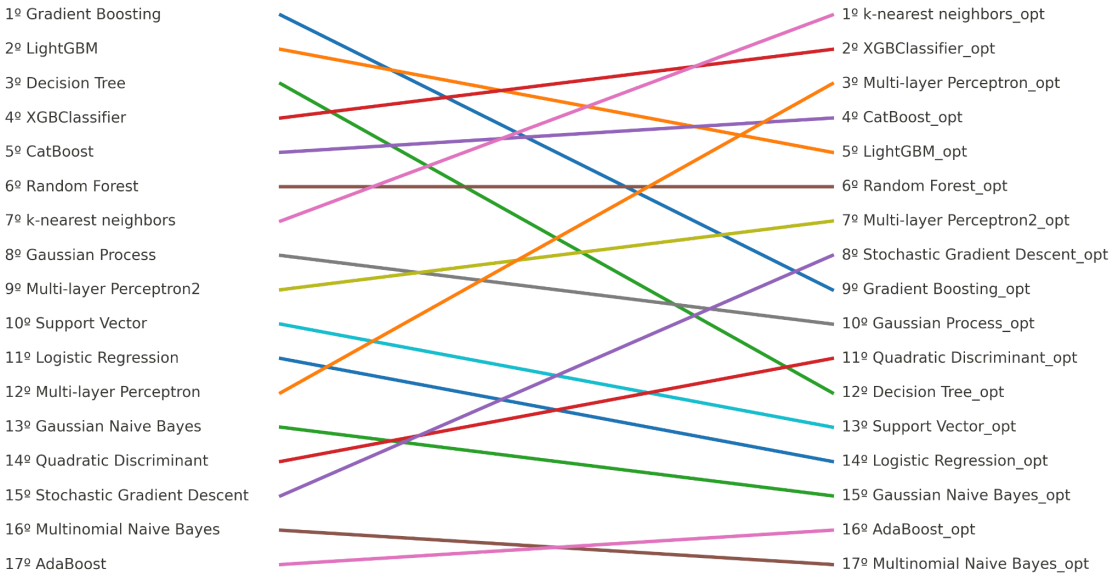

Figura 11. ***Ranqueamento dos 17 modelos treinados antes e após a otimização de hiperparâmetros.*** *Os modelos treinados foram testados com o conjunto de treino e ordenados segundo a seguinte ordem: acurácia balanceada, valor da área sob a curva ROC e macro-média F1-score. As linhas representam a mudança de posição neste ranqueamento, antes e após a otimização de hiperparâmetros.*

A Figura 12 reúne alguns valores de métricas dos top 3 modelos treinados, segundo a seguinte ordem: acurácia balanceada (*balanced_acc*), valor da área sob a curva ROC (*auc_roc*) e macro-média F1-score (f1_na_avg). Neste experimento, o modelo não otimizado melhor ranqueado foi o Gradient Boosting com 0,46 de acurácia balanceada e 0,76 para média ponderada de AUC-ROC (Figura 12).

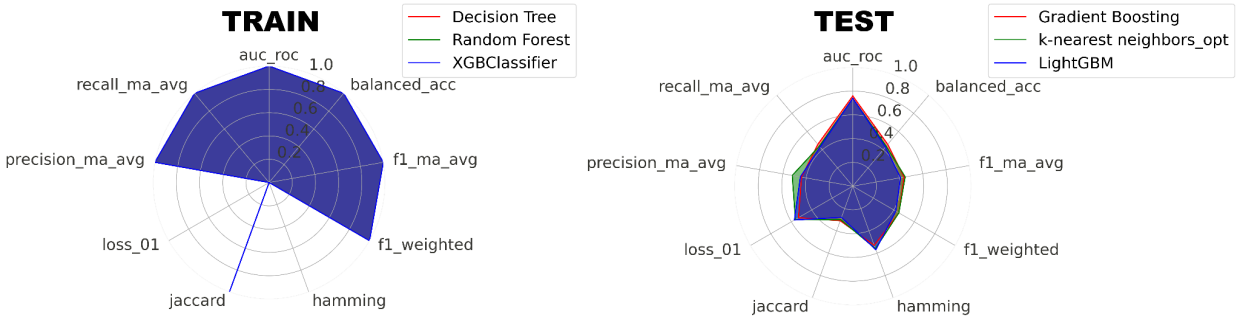

Figura 12. ***Radar plot ilustrando as métricas utilizadas para avaliar os modelos de classificação treinados com os dados de LMA unificados.*** À esquerda a visualização das métricas dos três melhores modelos utilizando os dados de treinamento na predição dos modelos, à direita utilizando os dados de teste. As métricas na imagem são as seguintes, a partir da posição central: AUC-ROC, acurácia balanceada, F1-score macro-média, F1-score ponderado, Perda Média de Hamming, Coeficiente de Similaridade de Jaccard, Perda de Classificação Zero-Um, Precisão, e Revocação.

Nota-se que não houve uma melhora significativa nas métricas dos modelos após otimização, nem mesmo quando o julgamento visual é feito com a ajuda da matrizes de confusão (Figura 13) e da AUC-ROC (Figura 14). Provavelmente a baixa quantidade de atributos (*features*) do conjunto de dados unificados pode estar colaborando com a baixa eficiência dos modelos, mesmo após otimização dos hiperparâmetros. Portanto, o próximo passo foi re-treinar os modelos, porém dessa vez utilizando os conjuntos de dados de LMA individualmente, sem unificação. Dessa forma, os modelos serão treinados com menor número de amostras, entretanto com maior número de atributos (*features*).

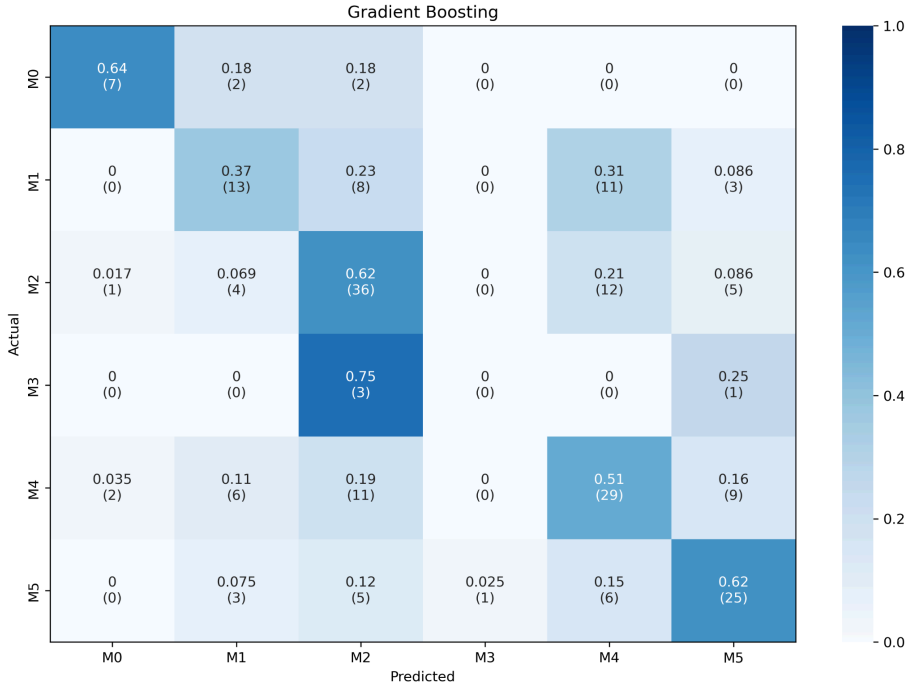

Figura 13. ***Matriz de confusão do melhor modelo de classificação do cenário com os dados de LMA unificados.*** *Na imagem o valor acima dos parênteses simboliza o valor percentual e dentro dos parênteses o valor bruto, tanto de erros quanto de acertos.*

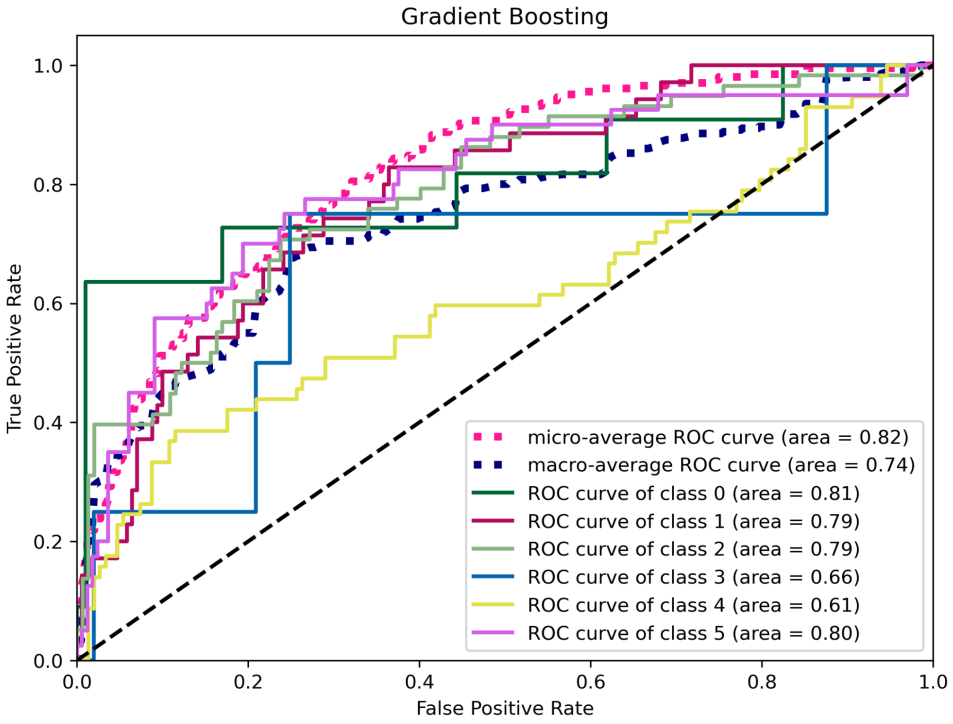

Figura 14. ***AUC-ROC do melhor modelo de classificação do cenário com os dados de LMA unificados.*** *Na imagem o melhor desempenho foi para a classe 0 (M0) apresentando um valor de área igual à 0,81. Por outro lado, o pior desempenho foi da classe 4 (M4) com um valor de área igual à 0,61. A linha preta tracejada diagonal representa uma curva com AUC de 0.5, indicando o limiar de aleatoriedade.*

2.4.2.1.2. Dados individualizados

2.4.2.1.2.1. TCGA

Os 17 modelos foram re-treinados neste novo cenário experimental apenas com os dados de LMA do TCGA. Da mesma forma que foi realizado anteriormente, os modelos foram inicialmente treinados sem otimização e posteriormente com a otimização dos respectivos hiperparâmetros. As dimensões do conjunto de dados do TCGA estão presentes na Tabela 3, sendo 178 pacientes e 19 parâmetros clínicos. As distribuições dos atributos no conjunto de dados foram realizadas em relação a cada subtipo de LMA segundo sistema de classificação FAB (Figura 15).

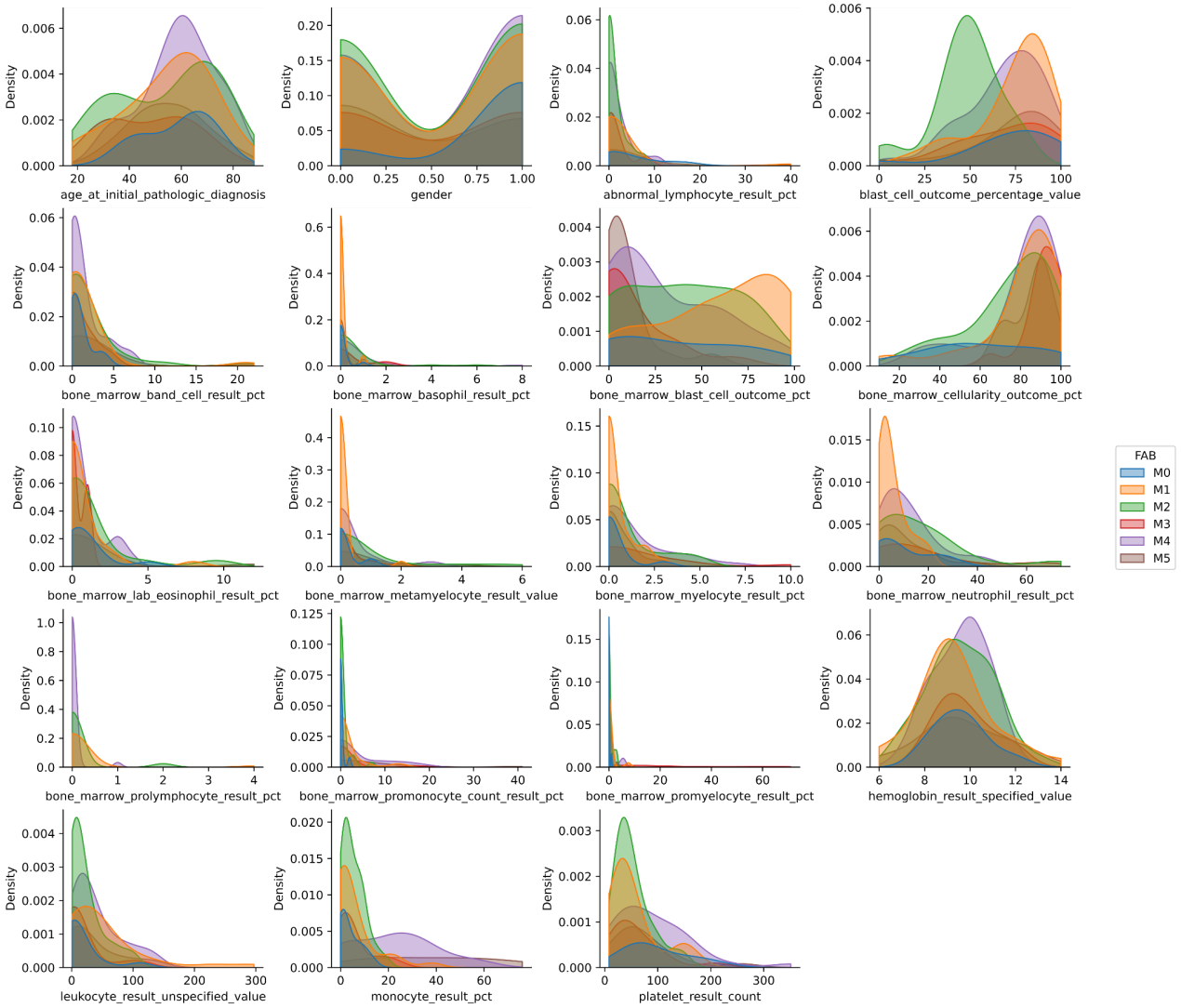

Figura 15. ***KDE plots com os valores de distribuição de cada um dos atributos utilizados no treinamento dos modelos.*** *Nesta visualização a distribuição de densidade de cada um dos atributos (antes da normalização) é separada por subtipo de LMA com o intuito de visualmente determinar se algum atributo isoladamente consegue distinguir os subtipos.*

Neste experimento, o modelo melhor ranqueado foi o SVM com hiperparâmetros otimizados, apresentando acurácia balanceada de 0,57 e 0,78 para média ponderada de AUC-ROC (Figura 16). As imagens para auxiliar no julgamento deste modelo estão ilustradas, sendo a primeira a matriz de confusão (Figura 17) e a segunda a AUC-ROC (Figura 18).

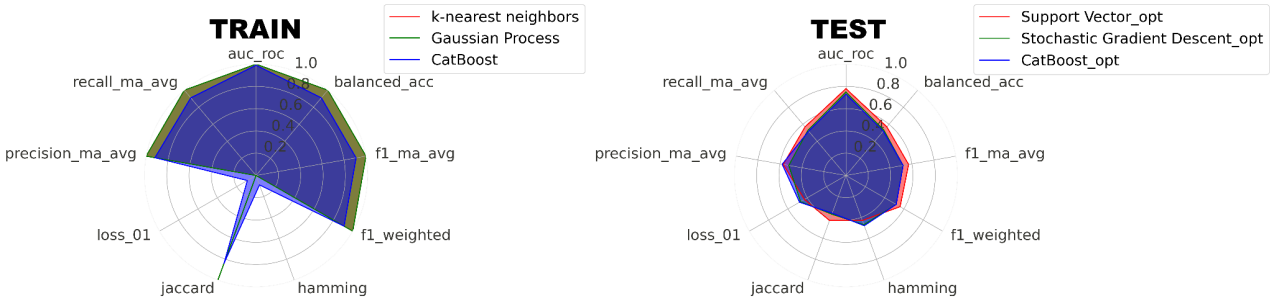

Figura 16. ***Radar plot ilustrando as métricas utilizadas para avaliar os modelos de classificação treinados com os dados de LMA do TCGA.*** À esquerda a visualização das métricas dos três melhores modelos utilizando os dados de treinamento na predição dos modelos, à direita utilizando os dados de teste. As métricas na imagem são as seguintes, a partir da posição central: AUC-ROC, acurácia balanceada, F1-score macro-média, F1-score ponderado, Perda Média de Hamming, Coeficiente de Similaridade de Jaccard, Perda de Classificação Zero-Um, Precisão, e Revocação.

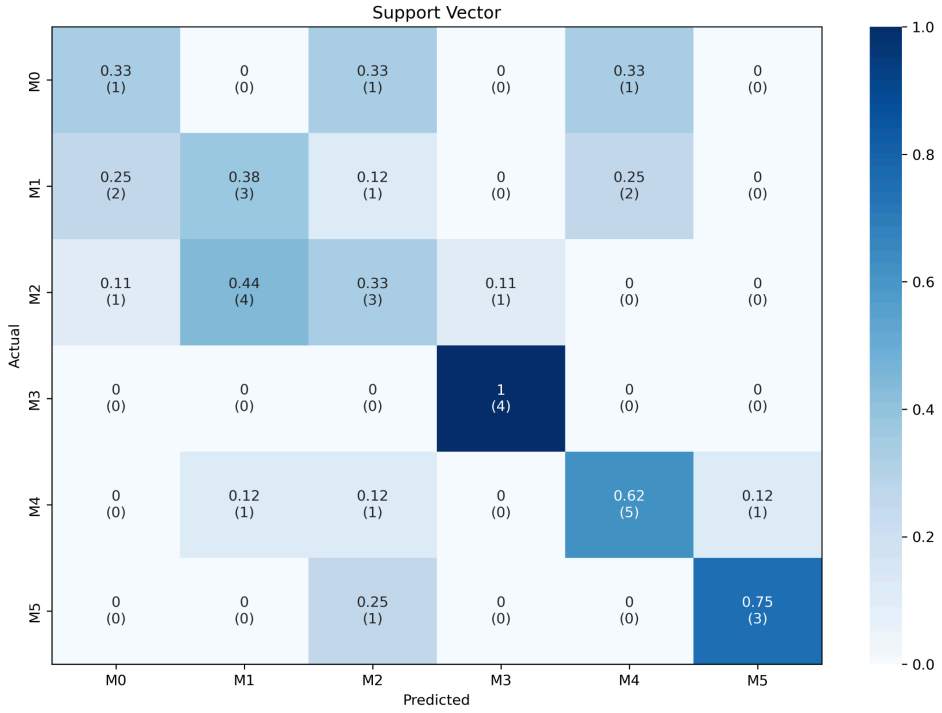

Figura 17. ***Matriz de confusão do melhor modelo de classificação treinado com os dados de LMA do TCGA.*** *Na imagem o valor acima dos parênteses simboliza o valor percentual e dentro dos parênteses o valor bruto, tanto de erros quanto de acertos.*

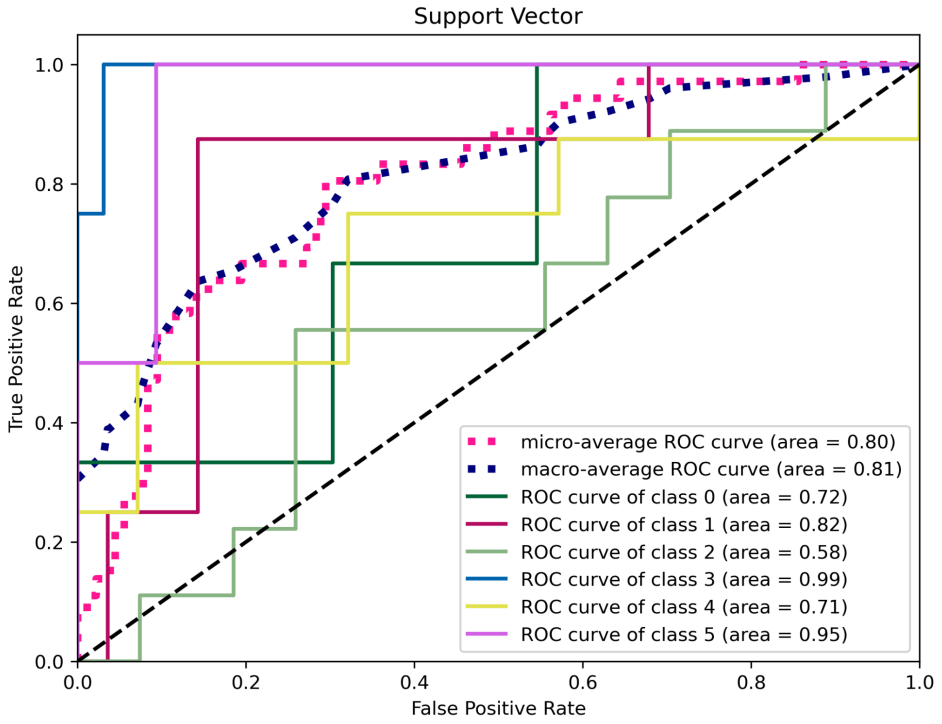

Figura 18. ***AUC-ROC do melhor modelo de classificação treinado com os dados de LMA do TCGA.*** *Na imagem o melhor desempenho foi para a classe 3 (M3) apresentando um valor de área igual à 0,99. Por outro lado, o pior desempenho foi da classe 2 (M2) com um valor de área igual à 0,58. A linha preta tracejada diagonal representa uma curva com AUC de 0.5, indicando o limiar de aleatoriedade.*

2.4.2.1.2.2. TARGET

Neste outro cenário experimental, os modelos foram treinados utilizando apenas o conjunto de dados TARGET, o qual apresenta 789 pacientes e 5 parâmetros clínicos (Tabela 3). As distribuições dos atributos no conjunto de dados foram realizadas em relação a cada subtipo de LMA segundo sistema de classificação FAB (Figura 19).

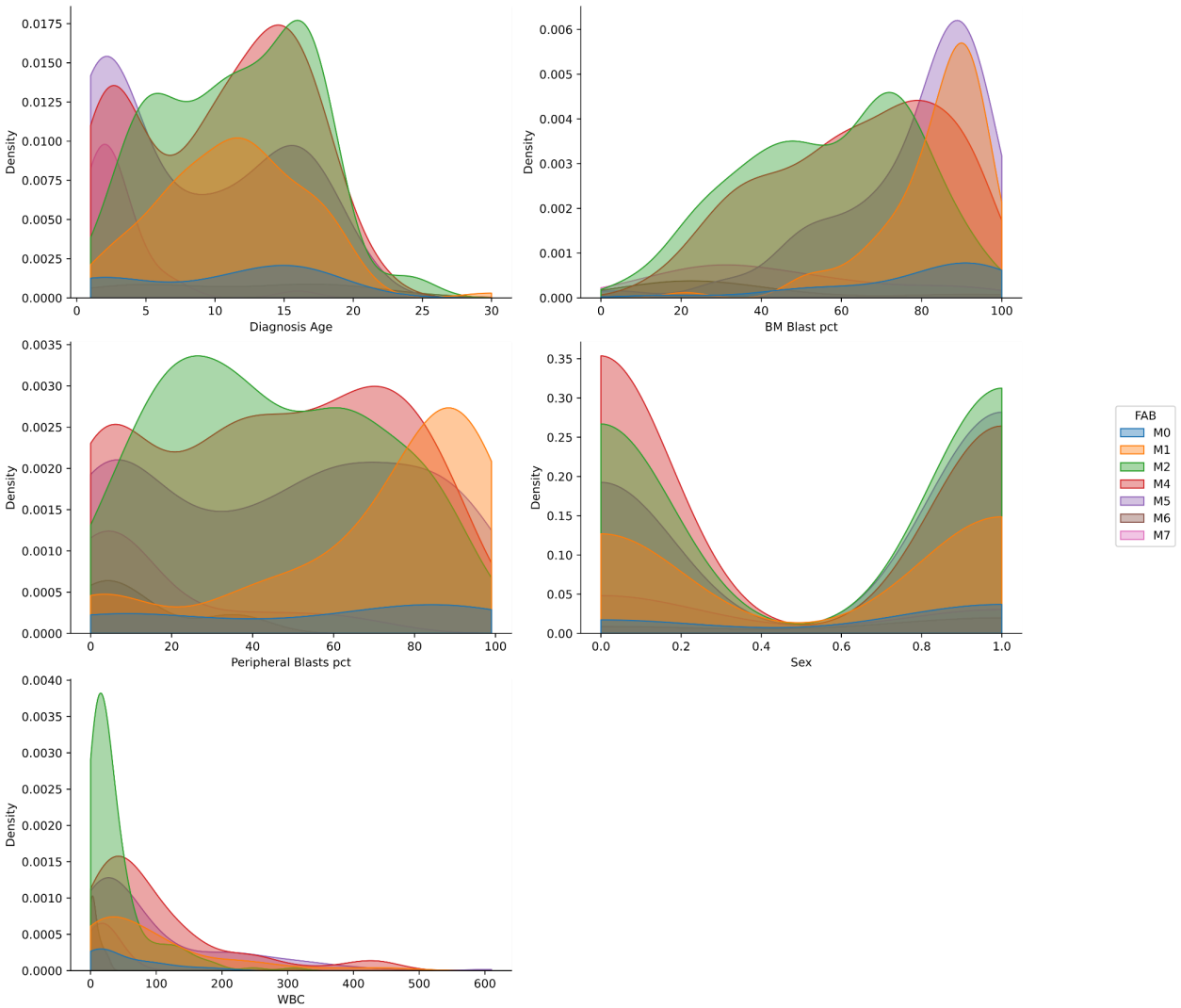

Figura 19. ***KDE plots com os valores de distribuição de cada um dos atributos utilizados no treinamento dos modelos.*** *Nesta visualização a distribuição de densidade de cada um dos atributos (antes da normalização) é separada por subtipo de LMA com o intuito de visualmente determinar se algum atributo isoladamente consegue distinguir os subtipos.*

Neste experimento, o modelo melhor ranqueado foi o CatBoost com hiperparâmetros otimizados, apresentando acurácia balanceada de 0,48 e 0,84 para média ponderada de AUC-ROC (Figura 20). As imagens para auxiliar no julgamento deste modelo estão ilustradas logo a seguir, sendo a primeira a matriz de confusão (Figura 21) e a segunda a AUC-ROC (Figura 22).

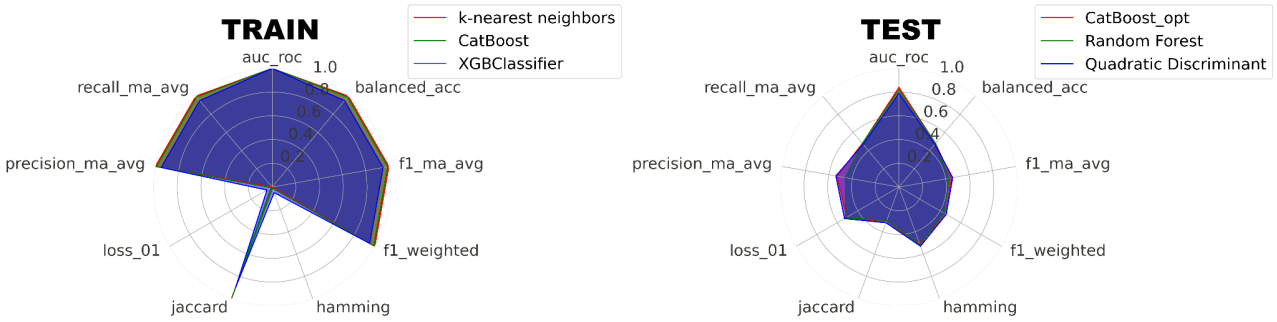

Figura 20. ***Radar plot ilustrando as métricas utilizadas para avaliar os modelos de classificação treinados com os dados de LMA do TARGET.*** À esquerda a visualização das métricas dos três melhores modelos utilizando os dados de treinamento na predição dos modelos, à direita utilizando os dados de teste. As métricas na imagem são as seguintes, a partir da posição central: AUC-ROC, acurácia balanceada, F1-score macro-média, F1-score ponderado, Perda Média de Hamming, Coeficiente de Similaridade de Jaccard, Perda de Classificação Zero-Um, Precisão, e Revocação.

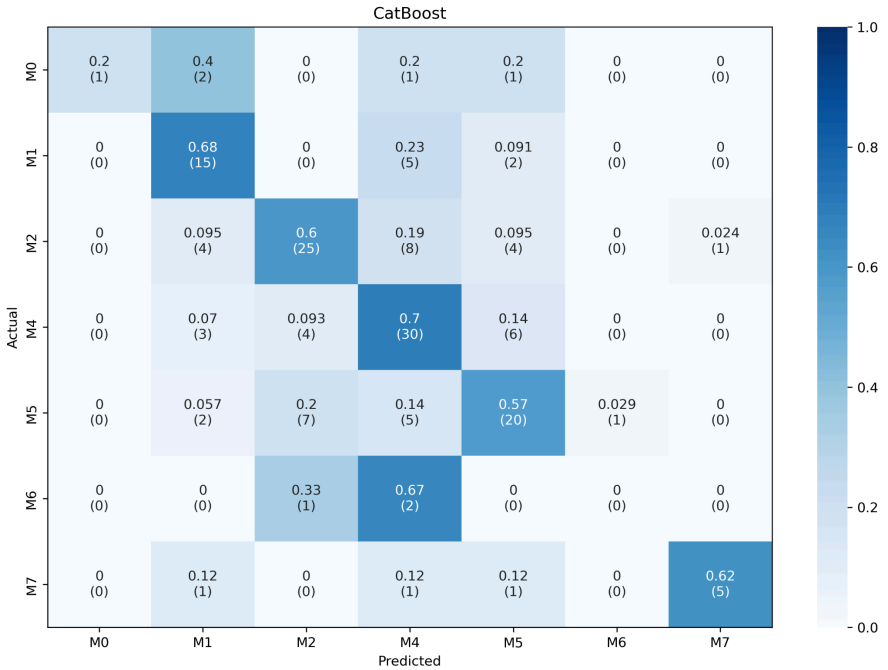

Figura 21. ***Matriz de confusão do melhor modelo de classificação treinado com os dados de LMA do TARGET.*** *Na imagem o valor acima dos parênteses simboliza o valor percentual e dentro dos parênteses o valor bruto, tanto de erros quanto de acertos.*

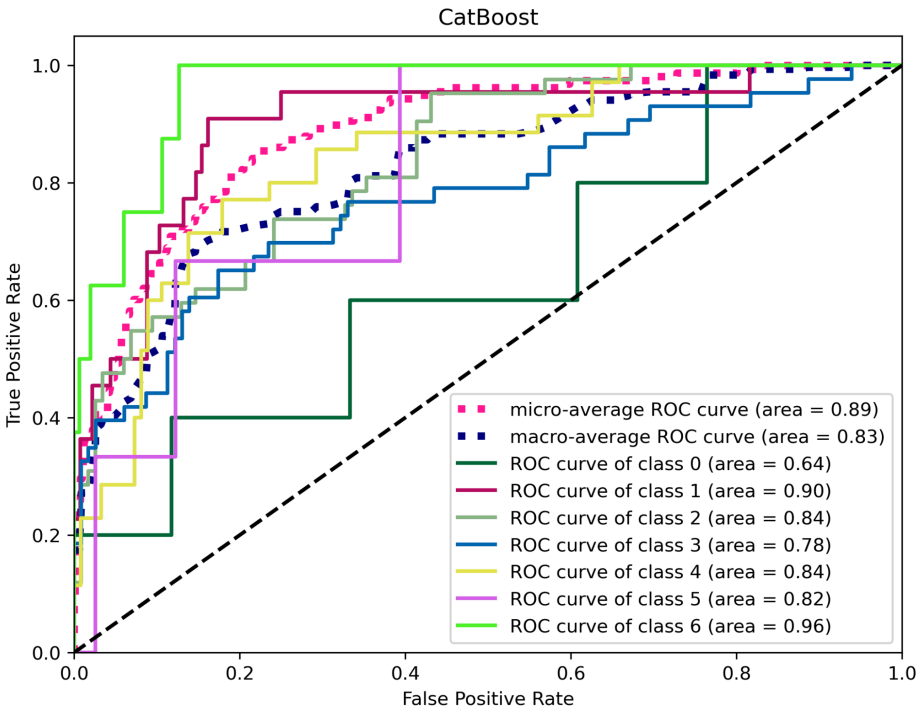

*Figura 22**. AUC-ROC do melhor modelo de classificação treinado com os dados de LMA do TARGET.** Na imagem o melhor desempenho foi para a classe 1 (M1) apresentando um valor de área igual à 0,90. Por outro lado, o pior desempenho foi da classe 0 (M0) com um valor de área igual à 0,64. A linha preta tracejada diagonal representa uma curva com AUC de 0.5, indicando o limiar de aleatoriedade.*

2.4.2.1.2.3. Hemocentro da Unicamp

O conjunto de dados do hemocentro apresenta 67 pacientes e 13 atributos clínicos, como consta na Tabela 3. As distribuições dos atributos no conjunto de dados também foram realizadas em relação a cada subtipo de LMA (Figura 23).

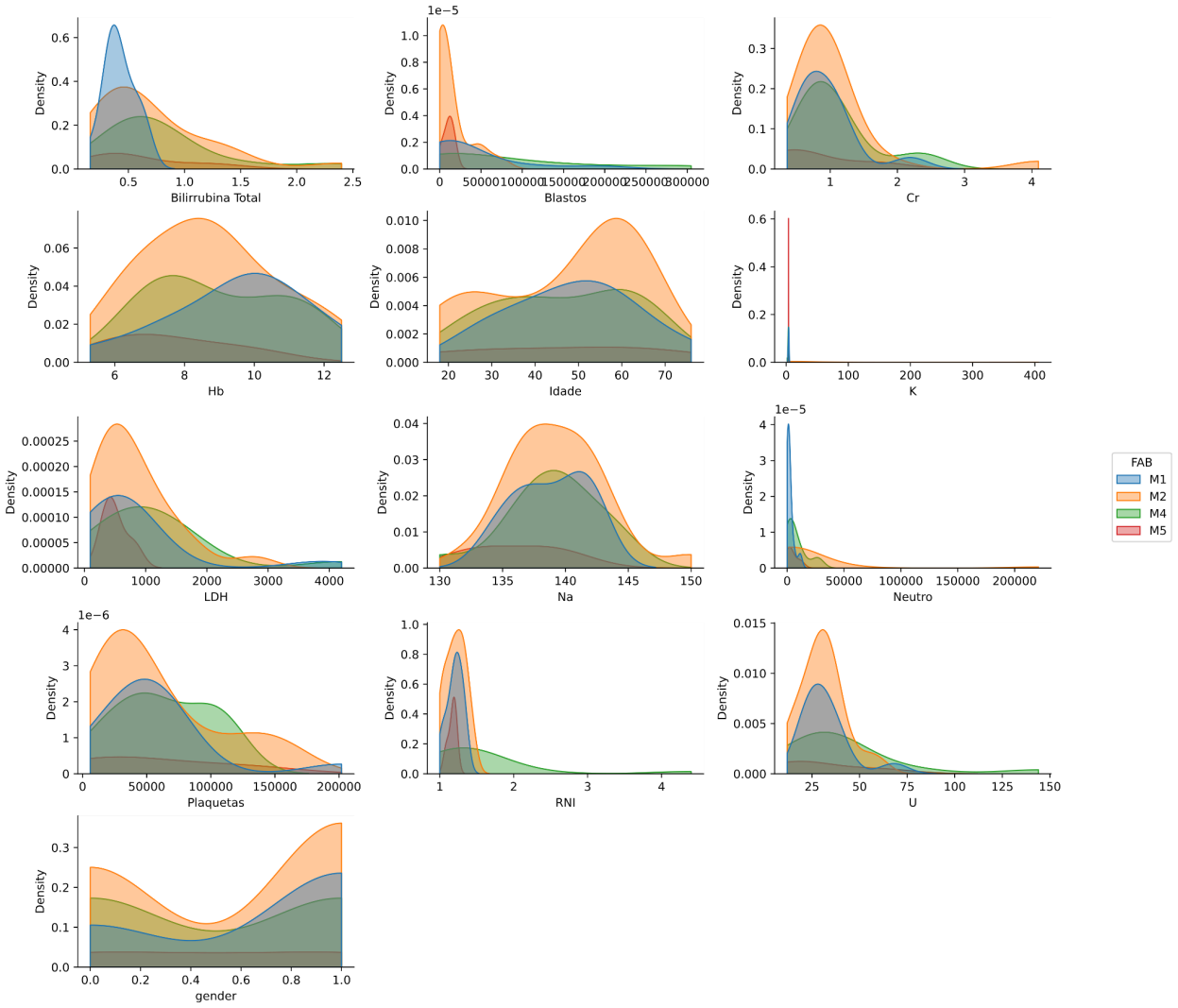

Figura 23. ***KDE plots com os valores de distribuição de cada um dos atributos utilizados no treinamento dos modelos.*** *Nesta visualização a distribuição de densidade de cada um dos atributos (antes da normalização) é separada por subtipo de LMA com o intuito de visualmente determinar se algum atributo isoladamente consegue distinguir os subtipos.*

O modelo melhor ranqueado neste cenário dentre os 34 (17 otimizados e 17 não-otimizados) foi o AdaBoost otimizado (Figura 24). O modelo apresentou 0,52 de acurácia balanceada e 0,73 para média ponderada de AUC-ROC. A matriz de confusão (Figura 25) e as curvas ROC (Figura 26) do melhor modelo treinado estão disponíveis a seguir para auxiliar na interpretação da performance dele. Apesar do valor relativamente alto desta média, o desempenho do modelo individualmente em cada classe foi medíocre, sendo o melhor resultado o valor de 0,50 (Figura 26).

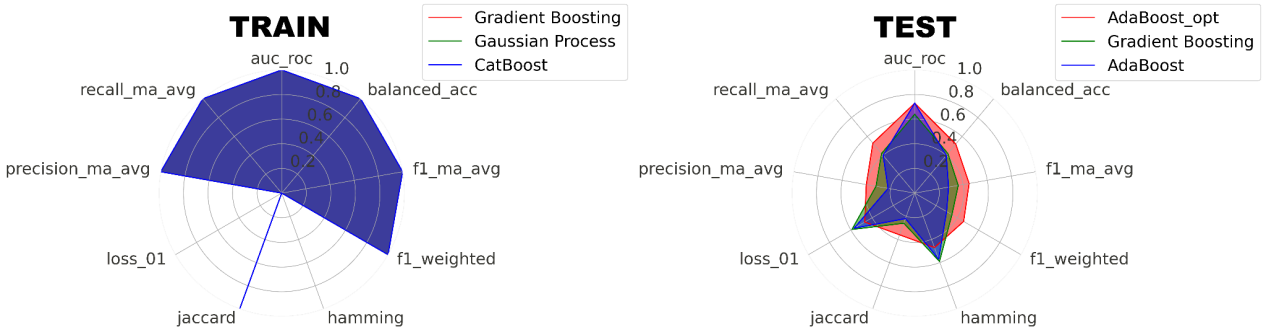

Figura 24. ***Radar plot ilustrando as métricas utilizadas para avaliar os modelos de classificação treinados com os dados de LMA do Hemocentro.*** À esquerda a visualização das métricas dos três melhores modelos utilizando os dados de treinamento na predição dos modelos, à direita utilizando os dados de teste. As métricas na imagem são as seguintes, a partir da posição central: AUC-ROC, acurácia balanceada, F1-score macro-média, F1-score ponderado, Perda Média de Hamming, Coeficiente de Similaridade de Jaccard, Perda de Classificação Zero-Um, Precisão, e Revocação.

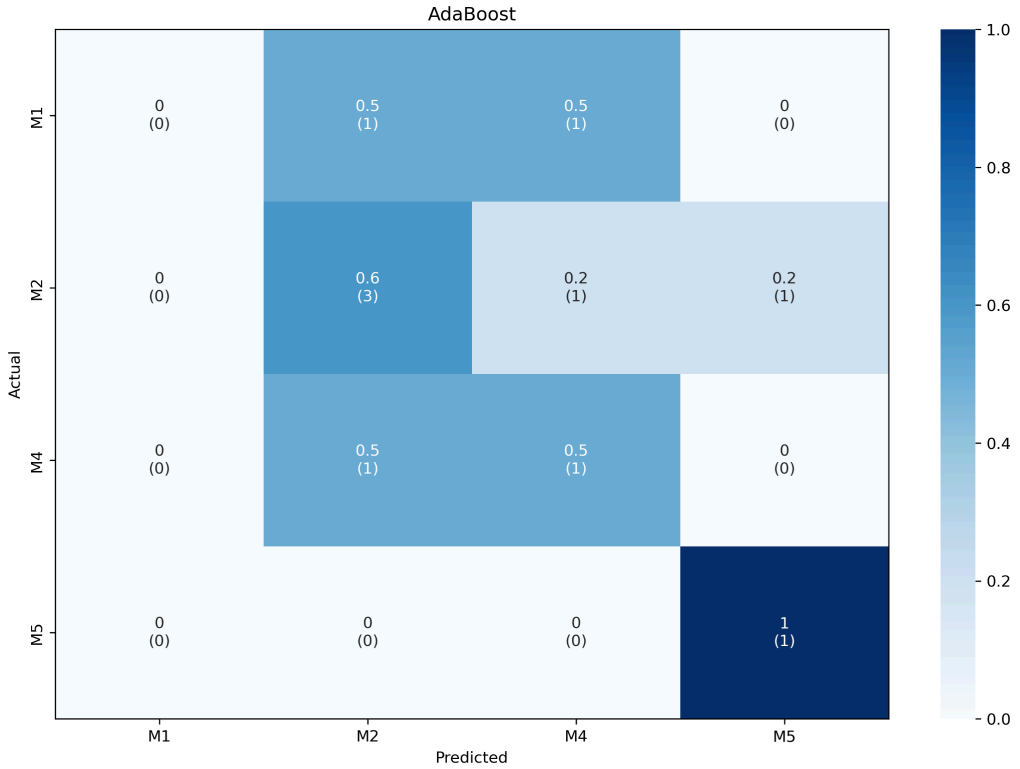

Figura 25. ***Matriz de confusão do melhor modelo de classificação treinado com os dados de LMA do Hemocentro.*** *Na imagem o valor acima dos parênteses simboliza o valor percentual e dentro dos parênteses o valor bruto, tanto de erros quanto de acertos.*

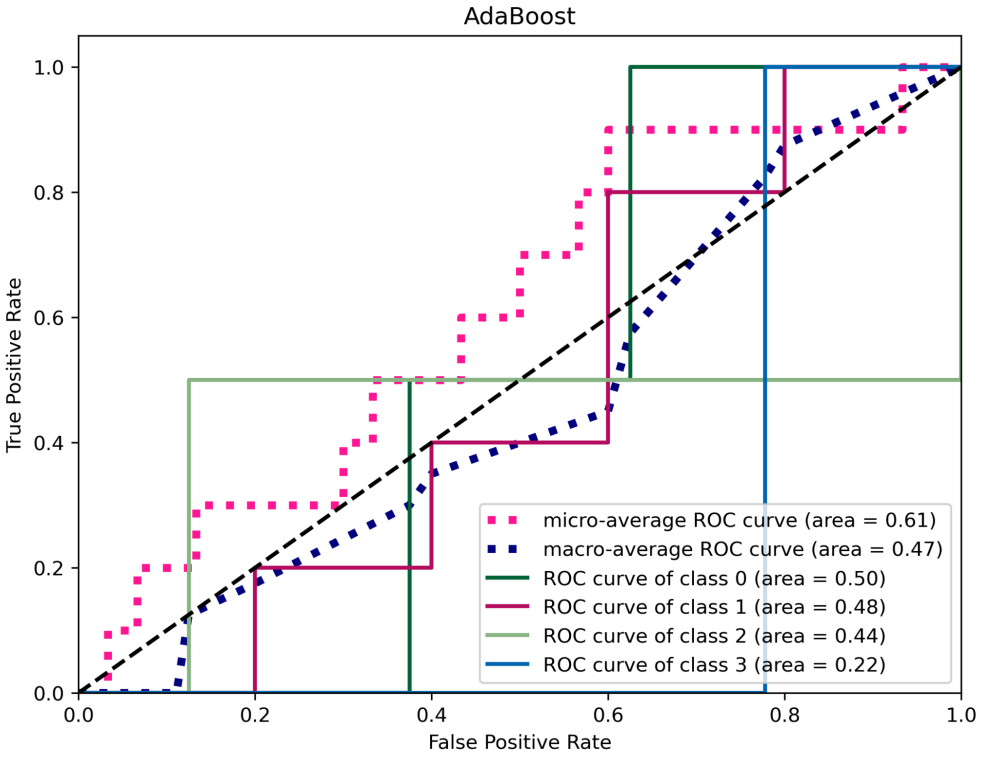

*Figura 26**. AUC-ROC do melhor modelo de classificação treinado com os dados de LMA do Hemocentro.** Na imagem o melhor desempenho foi para a classe 0 (M1) apresentando um valor de área igual à 0,50. Por outro lado, o pior desempenho foi da classe 3 (M5) com um valor de área igual à 0,22. A linha preta tracejada diagonal representa uma curva com AUC de 0.5, indicando o limiar de aleatoriedade.*

O desempenho dos modelos não apresentou ganho algum com o maior número de atributos durante o treinamento em relação ao cenário unificado. Os baixos valores das métricas indicam um baixo poder de generalização de todos os modelos treinados para LMA, em todos os cenários experimentais, dificultando desta forma a correta predição dos subtipos FAB dos pacientes de LMA.

2.4.2.2. SMD

O Apêndice B contém as tabelas com todas as métricas utilizadas para os modelos treinados com os dados de SMD (Tabelas S9-S18) e suas respectivas matrizes de confusão (Figuras S9-S18).

2.4.2.2.1. Dados unificados

2.4.2.2.1.1. Predição do sistema da OMS 2016

A dimensão resultante com a unificação foi 1943 pacientes e 5 atributos (Tabela 3). Assim como nos cenários anteriores, as visualizações das distribuições dos atributos no conjunto de dados foram realizadas em relação a cada categoria de predição, ou seja, por subtipo de SMD segundo sistema de classificação OMS 2016 (ARBER *et al.*, 2016) (Figura 27).

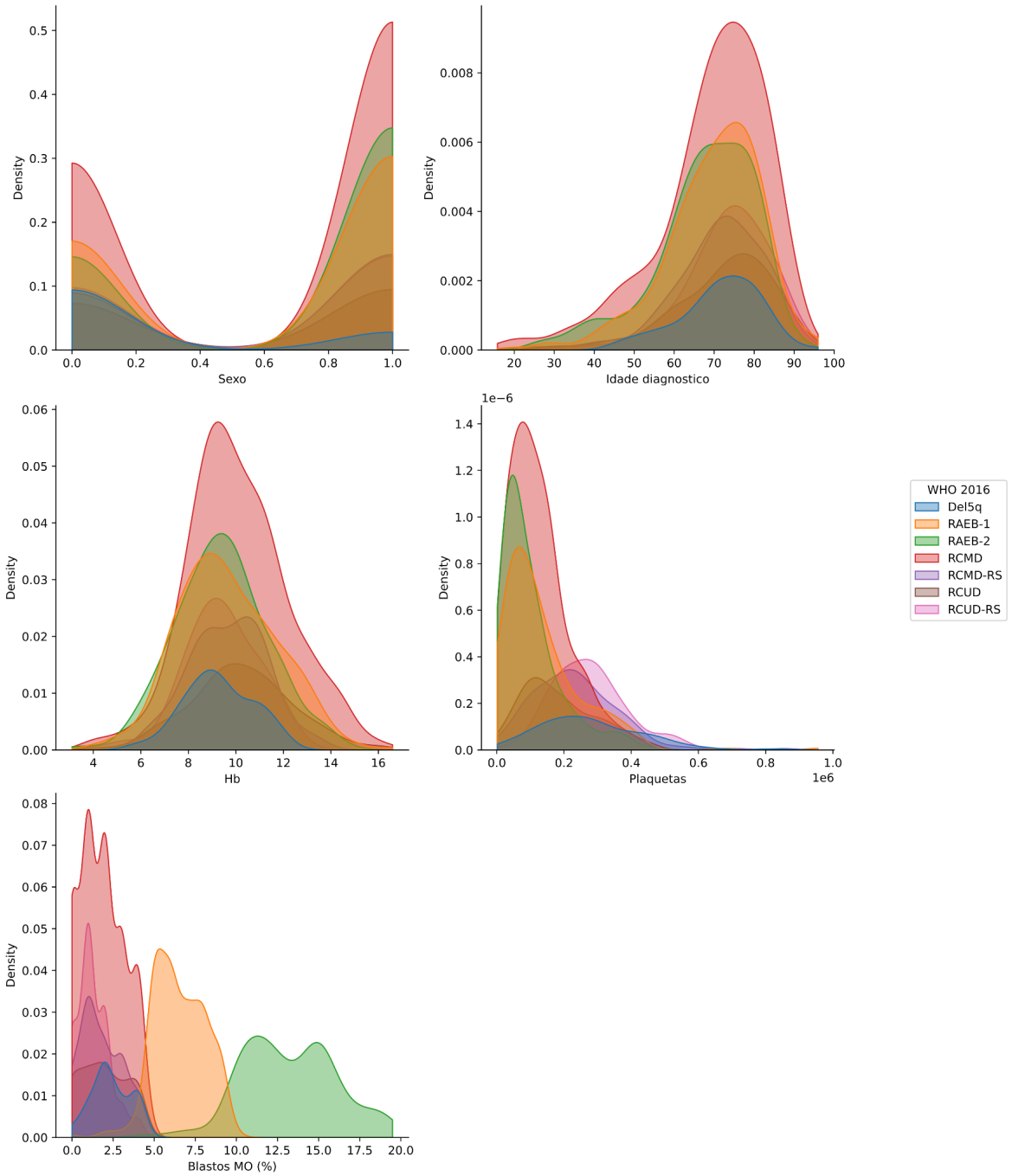

Figura 27. ***KDE plots com os valores de distribuição de cada um dos atributos utilizados no treinamento dos modelos.*** *Nesta visualização a distribuição de densidade de cada um dos atributos (antes da normalização) é separada por subtipo de SMD com o intuito de visualmente determinar se algum atributo isoladamente consegue distinguir os subtipos.*

Da mesma forma que ocorreu com os dados de LMA, os 17 modelos foram treinados sem e com otimização de hiperparâmetros. Neste experimento com os dados de SMD unificados, o modelo melhor ranqueado foi o Multi-layer Perceptron2 otimizado e treinado com os dados normalizados utilizando a técnica de

Min-Max scaler obtendo 0,53 de acurácia balanceada e 0,85 para média ponderada de AUC-ROC (Figura 28). A matriz de confusão bem como as curvas ROC do melhor modelo treinado são ilustrados pelas Figura 29 e Figura 30, respectivamente.

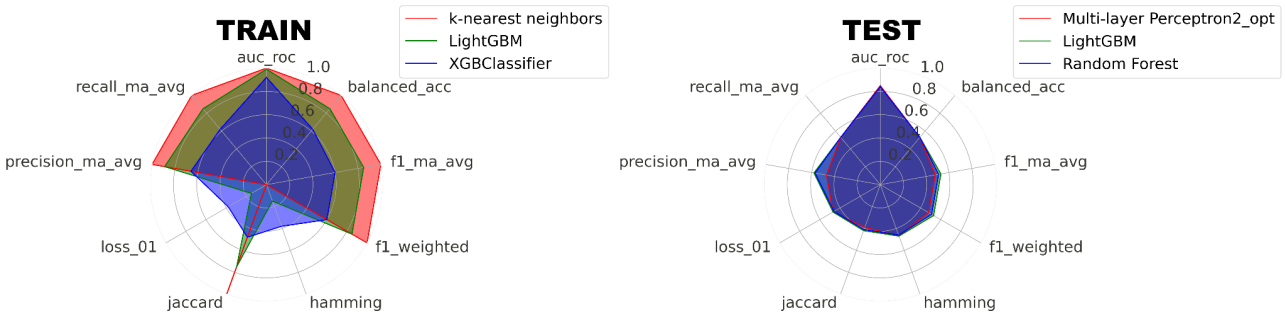

Figura 28. ***Radar plot ilustrando as métricas utilizadas para avaliar os modelos de classificação treinados com os dados de SMD reunidos.*** À esquerda a visualização das métricas dos três melhores modelos utilizando os dados de treinamento na predição dos modelos, à direita utilizando os dados de teste. As métricas na imagem são as seguintes, a partir da posição central: AUC-ROC, acurácia balanceada, F1-score macro-média, F1-score ponderado, Perda Média de Hamming, Coeficiente de Similaridade de Jaccard, Perda de Classificação Zero-Um, Precisão, e Revocação.

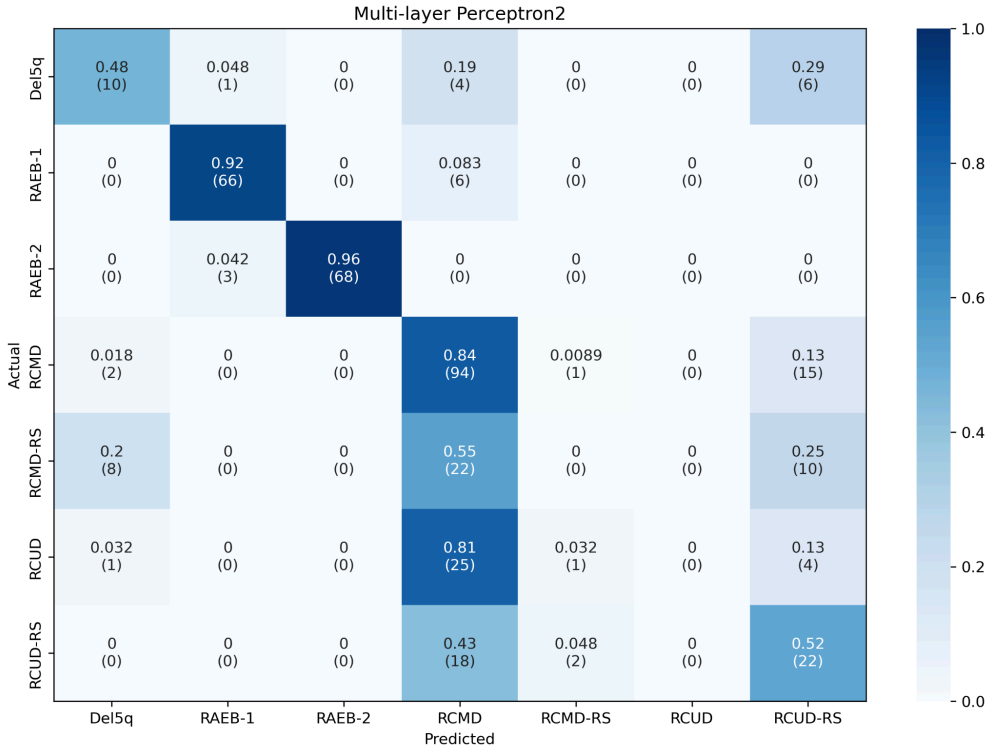

Figura 29. ***Matriz de confusão do melhor modelo de classificação treinado com os dados de SMD unificados.*** *Na imagem o valor acima dos parênteses simboliza o valor percentual e dentro dos parênteses o valor bruto, tanto de erros quanto de acertos.*

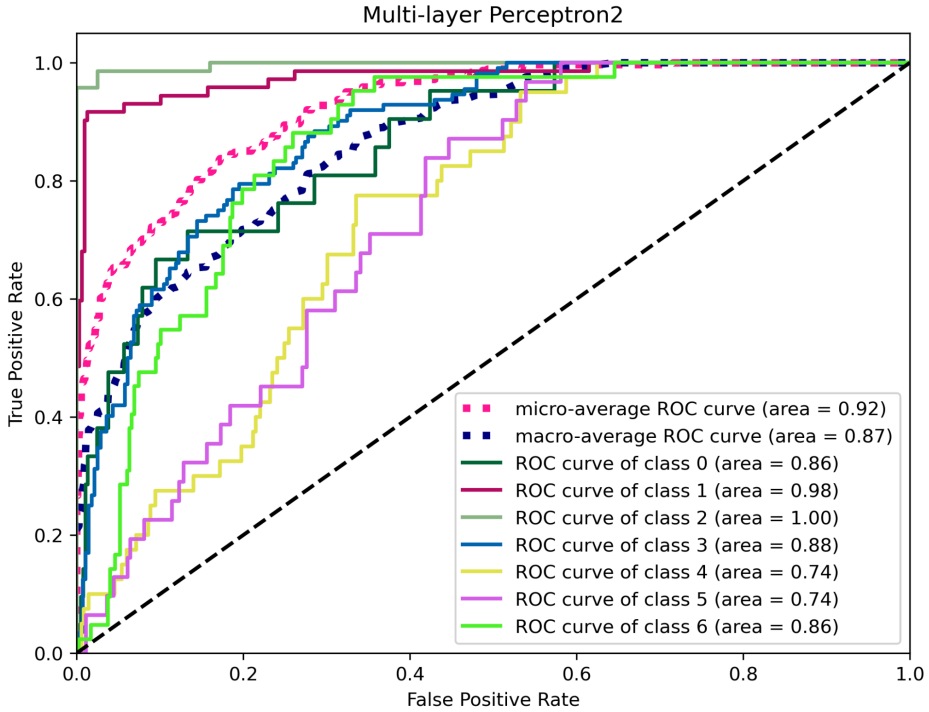

Figura 30. ***AUC-ROC do melhor modelo de classificação treinado com os dados de SMD unificados.*** *Na imagem o melhor desempenho foi para a classe 2 (RAEB-2) apresentando um valor de área igual à 1,00. Por outro lado, o pior desempenho foi compartilhado pelas classes 4 (RCMD-RS) e 5 (RCUD) com um valor de área igual à 0,74. A linha preta tracejada diagonal representa uma curva com AUC de 0.5, indicando o limiar de aleatoriedade.*

2.4.2.2.1.2. Predição do índice de prognóstico (IPSS-R)

Os modelos até o momento apresentados foram excepcionalmente treinados, neste cenário experimental, com o objetivo de predizer o IPSS-R, sem utilizar dados citogenéticos, os quais são fundamentais neste sistema de classificação. Da mesma forma que anteriormente, os modelos foram comparados antes e depois da otimização de hiperparâmetros. O conjunto de dados para este cenário foi o mesmo do anterior, com os dados de SMD reunidos, o qual apresentou a dimensão de 1943 pacientes e 5 variáveis (Tabela 3). As visualizações das distribuições dos atributos neste experimento foram desta vez realizadas em relação a cada subtipo de SMD segundo sistema de classificação IPSS-R (Figura 31).

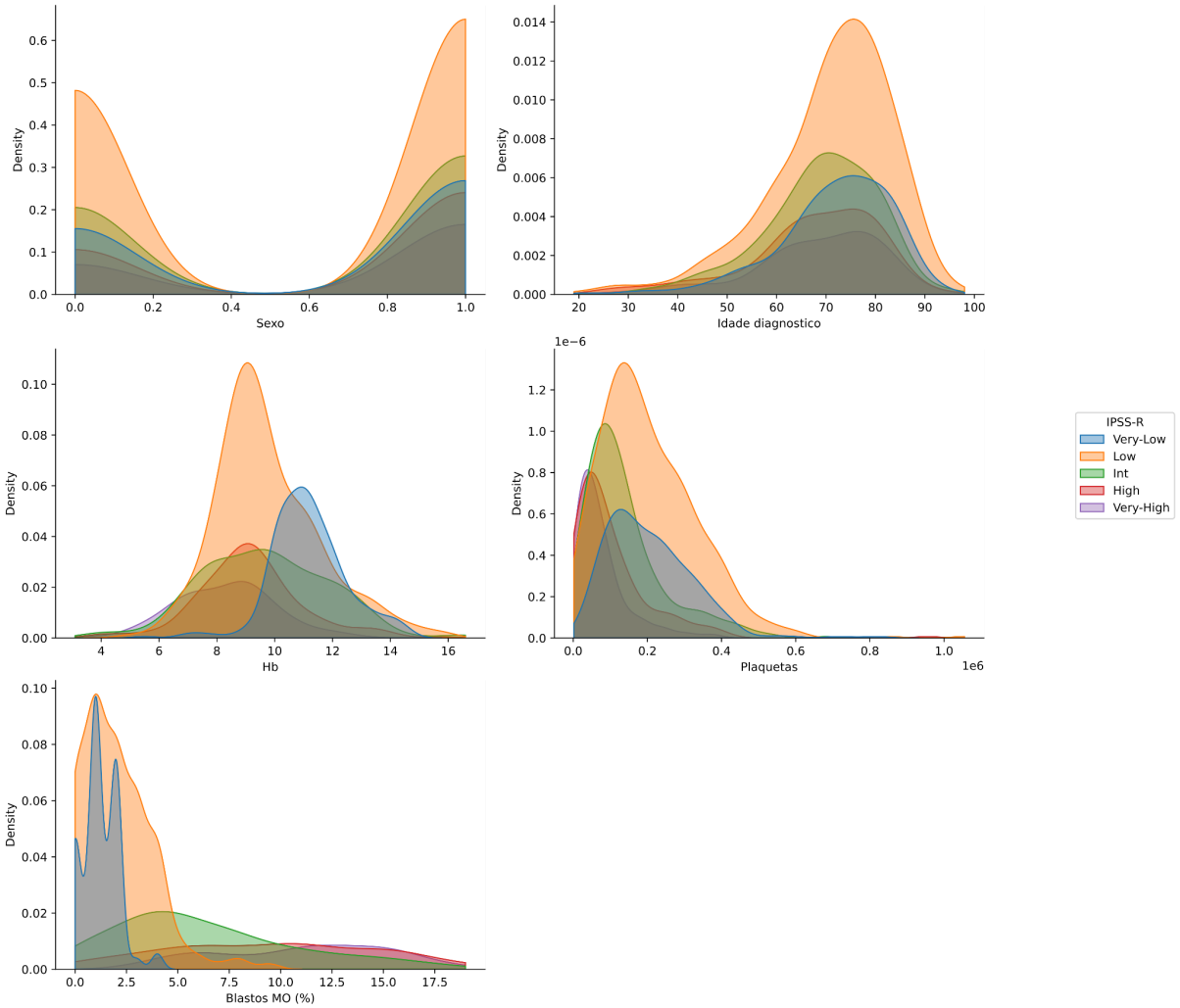

Figura 31. ***KDE plots com os valores de distribuição de cada um dos atributos utilizados no treinamento dos modelos.*** *Nesta visualização a distribuição de densidade de cada um dos atributos (antes da normalização) é separada por subtipo de SMD com o intuito de visualmente determinar se algum atributo isoladamente consegue distinguir os subtipos.*

Neste experimento, o modelo melhor ranqueado foi o CatBoost otimizado apresentando 0,64 de acurácia balanceada e 0,89 para média ponderada de AUC-ROC (Figura 32). As imagens para auxiliar no julgamento deste modelo estão ilustradas, sendo a primeira a matriz de confusão (Figura 33) e a segunda a AUC-ROC (Figura 34).

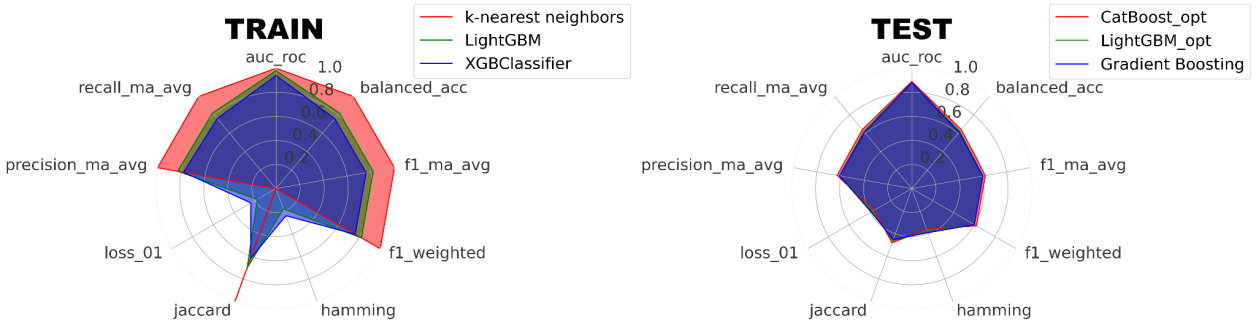

Figura 32. ***Radar plot ilustrando as métricas utilizadas para avaliar os modelos de classificação treinados para predição do IPSS-R com os dados de SMD reunidos.*** À esquerda a visualização das métricas dos três melhores modelos utilizando os dados de treinamento na predição dos modelos, à direita utilizando os dados de teste. As métricas na imagem são as seguintes, a partir da posição central: AUC-ROC, acurácia balanceada, F1-score macro-média, F1-score ponderado, Perda Média de Hamming, Coeficiente de Similaridade de Jaccard, Perda de Classificação Zero-Um, Precisão, e Revocação.

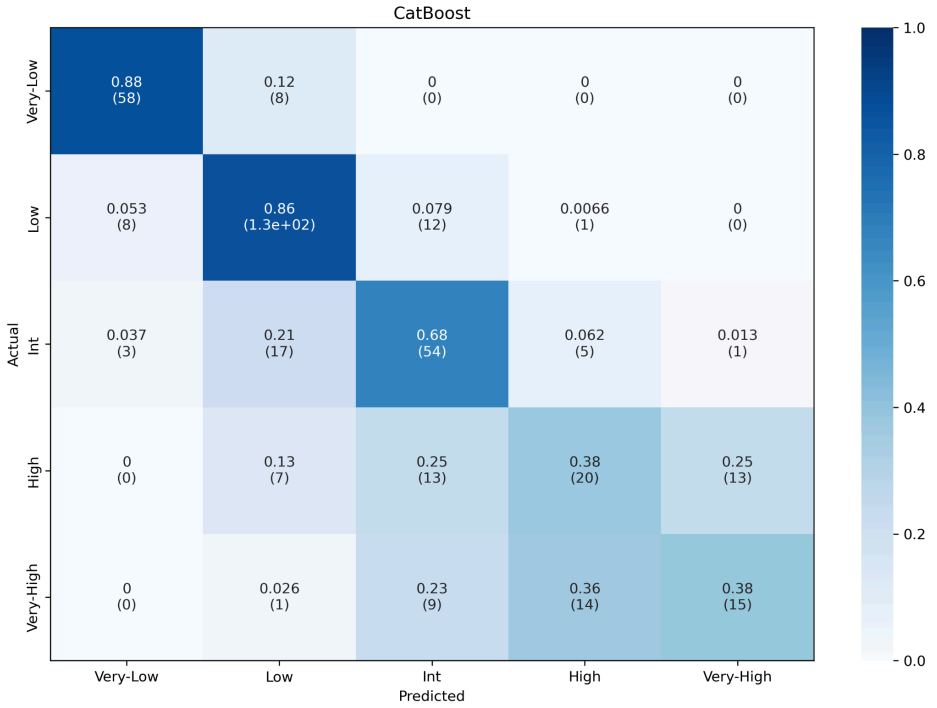

Figura 33. ***Matriz de confusão do melhor modelo de classificação treinado para predição do IPSS-R utilizando os dados de SMD unificados.*** *Na imagem o valor acima dos parênteses simboliza o valor percentual e dentro dos parênteses o valor bruto, tanto de erros quanto de acertos.*

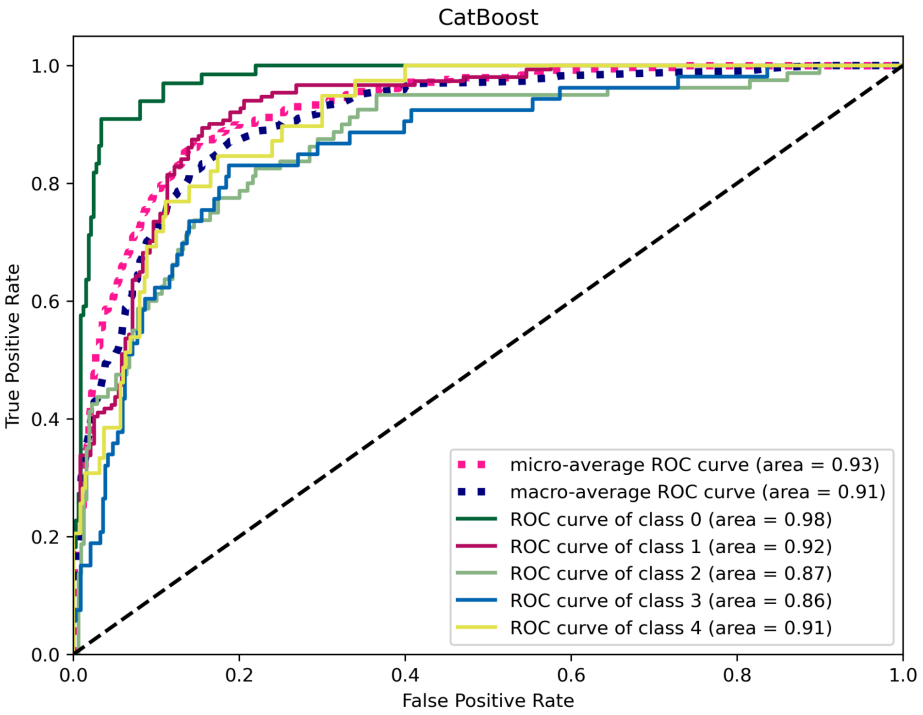

Figura 34. ***AUC-ROC do melhor modelo de classificação treinado para predição do IPSS-R utilizando os dados de SMD unificados.*** *Na imagem o melhor desempenho foi para a classe 0 (Muito Baixo) apresentando um valor de área igual à 0,98. Por outro lado, o pior desempenho foi performado para a classe 2 (Intermediário) com um valor de área igual à 0,87. A linha preta tracejada diagonal representa uma curva com AUC de 0.5, indicando o limiar de aleatoriedade.*

2.4.2.2.2. Dados individualizados

O desempenho não satisfatório do melhor modelo selecionado até o momento conduziu o trabalho para a criação de novos cenários onde cada conjunto de dados é explorado individualmente, como ocorreu após explorar os dados de LMA.

2.4.2.2.2.1. Dados UFCeará

Neste primeiro dos cenários experimentais individuais para SMD, os modelos foram treinados utilizando apenas o conjunto de dados fornecido por um de nossos colaboradores, o qual apresenta após o pré-processamento 117 pacientes e 6 parâmetros clínicos (Tabela 3). As distribuições dos atributos no conjunto de dados foram realizadas em relação a cada subtipo de SMD segundo sistema de

classificação OMS 2016 (Figura 35), assim como será realizado para todos os experimentos dentro desta seção.

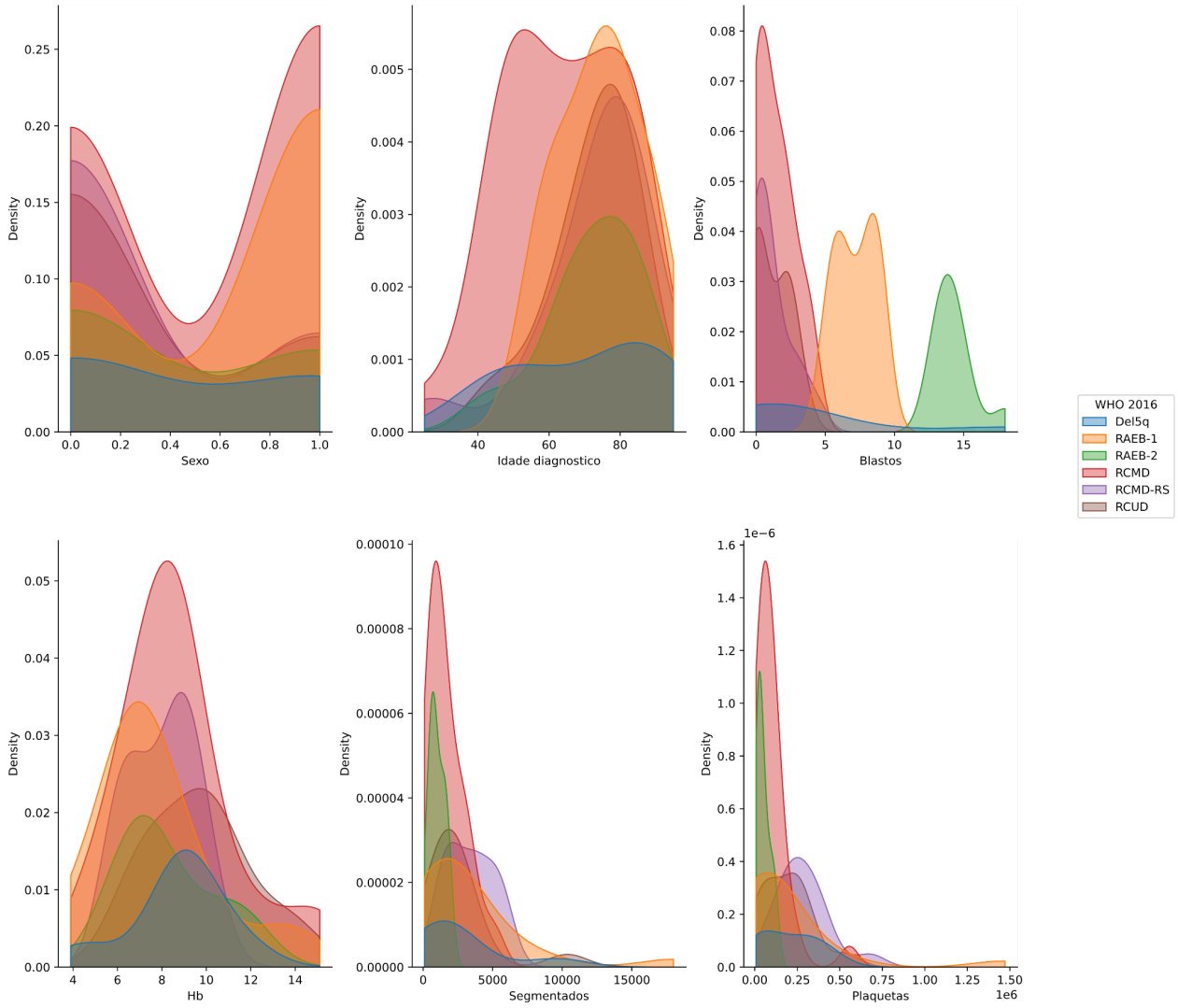

Figura 35. ***KDE plots com os valores de distribuição de cada um dos atributos utilizados no treinamento dos modelos.*** *Nesta visualização a distribuição de densidade de cada um dos atributos (antes da normalização) é separada por subtipo de SMD com o intuito de visualmente determinar se algum atributo isoladamente consegue distinguir os subtipos.*

Após os treinamentos, o modelo melhor ranqueado foi o Multi-layer Perceptron com hiperparâmetros otimizados, apresentando acurácia balanceada de 0,60 e 0,81 para média ponderada de AUC-ROC (Figura 36). As imagens para auxiliar no julgamento deste modelo estão ilustradas logo a seguir, sendo a primeira a matriz de confusão (Figura 37) e a segunda a AUC-ROC (Figura 38). Nota-se que embora o modelo tenha desempenhado bem para uma dos subtipos, ele

performou muito abaixo do esperado para outros subtipos como o da classe 5 (RCUD) (Figura 38).

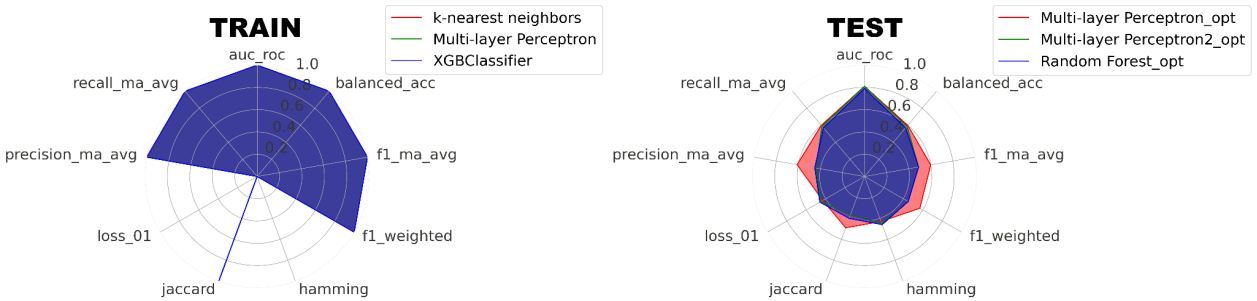

Figura 36. ***Radar plot ilustrando as métricas utilizadas para avaliar os modelos de classificação treinados.*** À esquerda a visualização das métricas dos três melhores modelos utilizando os dados de treinamento na predição dos modelos, à direita utilizando os dados de teste. As métricas na imagem são as seguintes, a partir da posição central: AUC-ROC, acurácia balanceada, F1-score macro-média, F1-score ponderado, Perda Média de Hamming, Coeficiente de Similaridade de Jaccard, Perda de Classificação Zero-Um, Precisão, e Revocação.

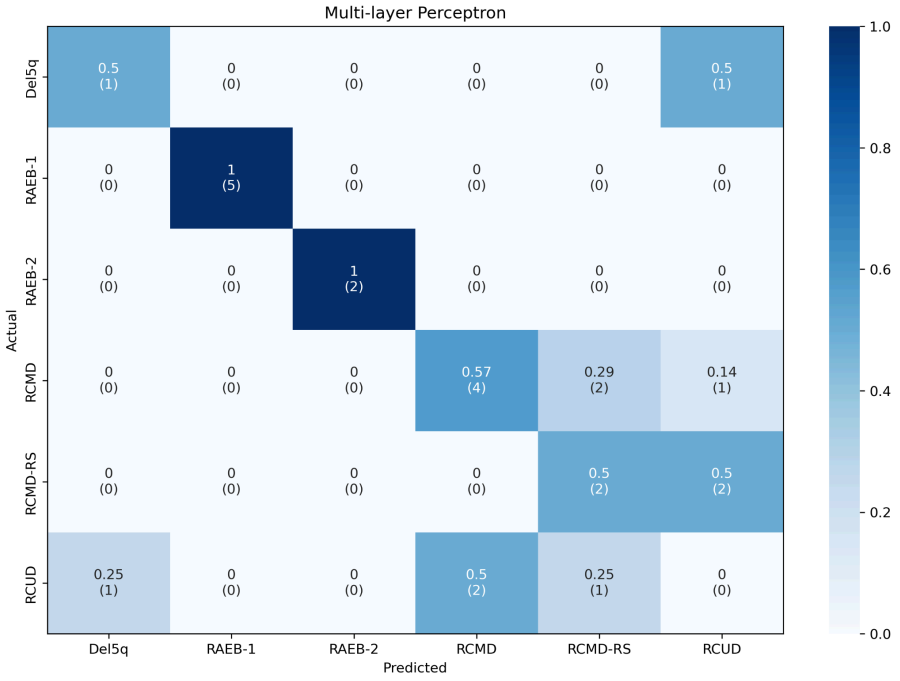

Figura 37. ***Matriz de confusão do melhor modelo de classificação treinado.*** *Na imagem o valor acima dos parênteses simboliza o valor percentual e dentro dos parênteses o valor bruto, tanto de erros quanto de acertos.*

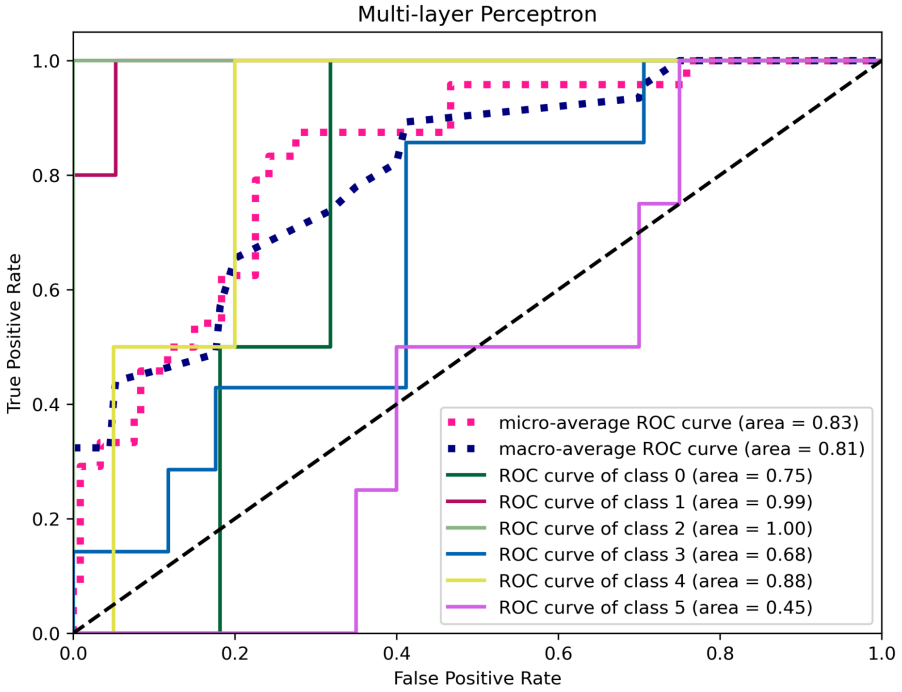

*Figura 38. **AUC-ROC do melhor modelo de classificação treinado.** Na imagem o melhor desempenho foi para a classe 2 (RAEB-2) apresentando um valor de área igual à 1,00. Por outro lado, o pior desempenho foi o da classe 5 (RCUD) com um valor de área igual à 0,45. A linha preta tracejada diagonal representa uma curva com AUC de 0.5, indicando o limiar de aleatoriedade.*

2.4.2.2.2.2. MDS-IWG

O conjunto de dados do MDS-IWG apresenta 1862 pacientes e 6 atributos clínicos, como consta na Tabela 3. As distribuições dos atributos no conjunto de dados também foram realizadas em relação a cada subtipo de SMD, pelo sistema OMS 2016 (Figura 39).

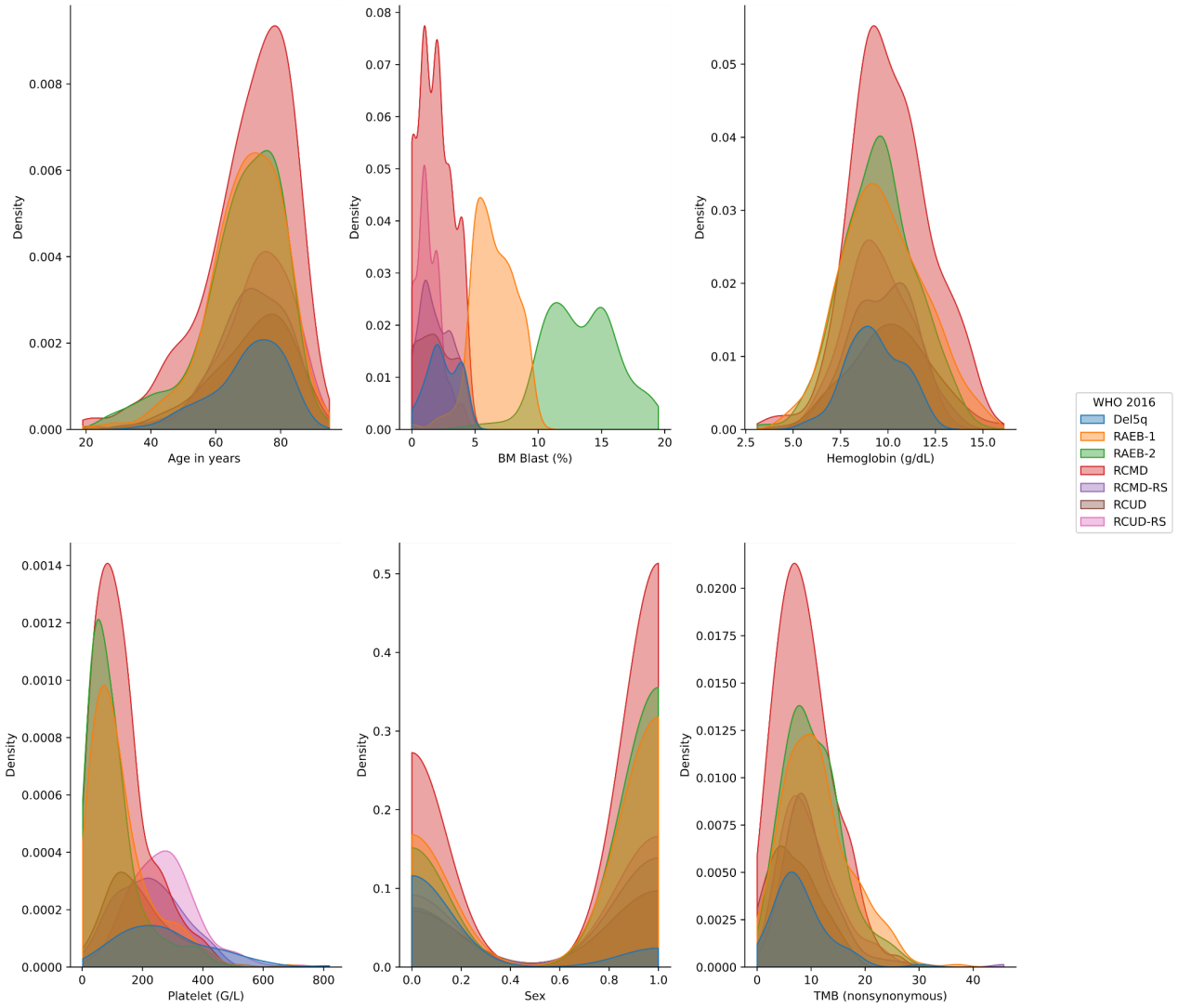

Figura 39. ***KDE plots com os valores de distribuição de cada um dos atributos utilizados no treinamento dos modelos.*** *Nesta visualização a distribuição de densidade de cada um dos atributos (antes da normalização) é separada por subtipo de SMD com o intuito de visualmente determinar se algum atributo isoladamente consegue distinguir os subtipos.*

O modelo melhor ranqueado neste cenário dentre os 34 (17 otimizados e 17 não-otimizados) foi o CatBoost otimizado (Figura 40). O modelo apresentou 0,56 de acurácia balanceada e 0,86 para média ponderada de AUC-ROC. A matriz de confusão (Figura 41) e as curvas ROC (Figura 42) deste modelo estão disponíveis a seguir para auxiliar na interpretação da performance dele.

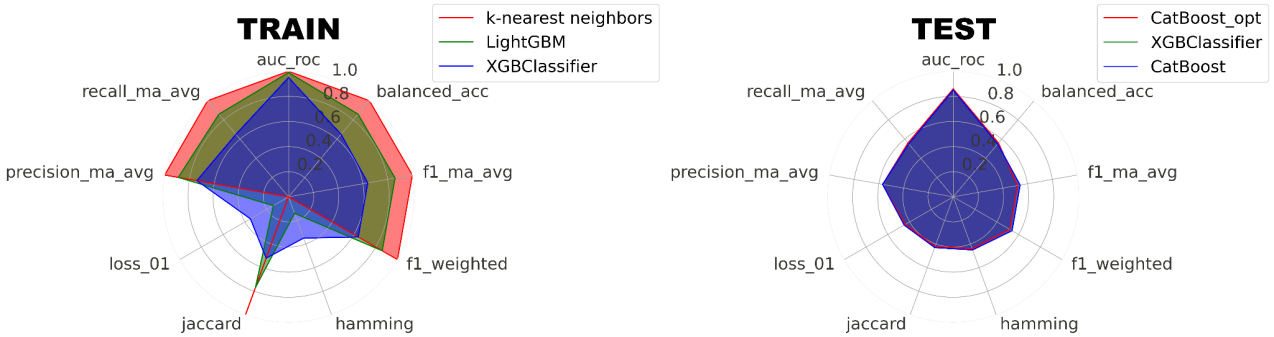

Figura 40. ***Radar plot ilustrando as métricas utilizadas para avaliar os modelos de classificação treinados com os dados de SMD oriundos do MDS-IWG.*** À esquerda a visualização das métricas dos três melhores modelos utilizando os dados de treinamento na predição dos modelos, à direita utilizando os dados de teste. As métricas na imagem são as seguintes, a partir da posição central: AUC-ROC, acurácia balanceada, F1-score macro-média, F1-score ponderado, Perda Média de Hamming, Coeficiente de Similaridade de Jaccard, Perda de Classificação Zero-Um, Precisão, e Revocação.

Figura 41. ***Matriz de confusão do melhor modelo de classificação treinado com os dados de SMD oriundos do MDS-IWG.*** *Na imagem o valor acima dos parênteses simboliza o valor percentual e dentro dos parênteses o valor bruto, tanto de erros quanto de acertos.*

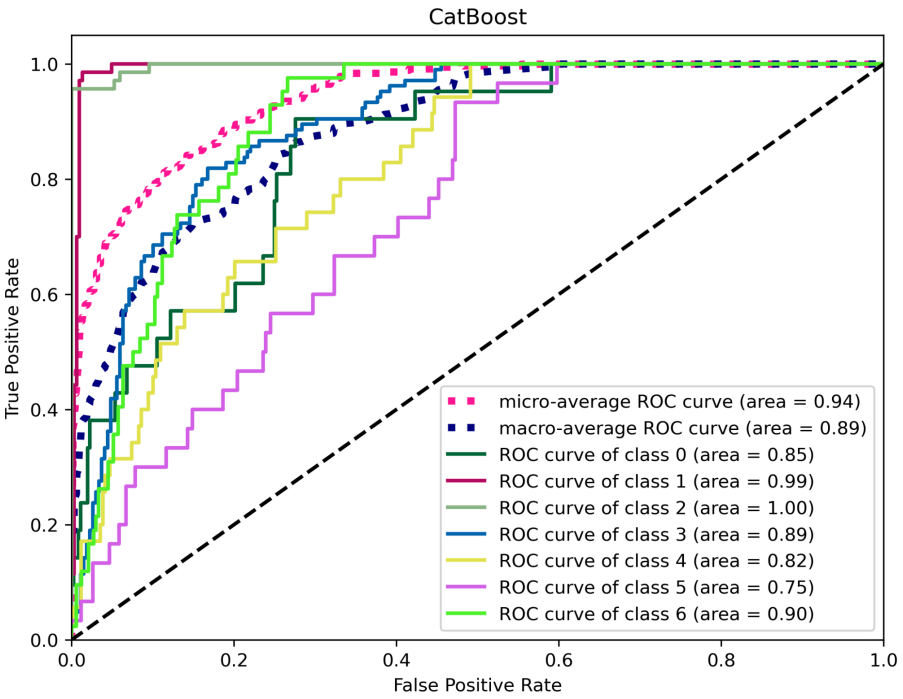

Figura 42. ***AUC-ROC do melhor modelo de classificação treinado com os dados de SMD oriundos do MDS-IWG.*** *Na imagem o melhor desempenho foi para a classe 2 (RAEB-2) apresentando um valor de área igual à 1,00. Por outro lado, o pior desempenho foi o da classe 5 (RCUD) com um valor de área igual à 0,75. A linha preta tracejada diagonal representa uma curva com AUC de 0.5, indicando o limiar de aleatoriedade.*

2.4.2.2.2.3. Hemocentro

O conjunto de dados de SMD neste cenário experimental apresentou a dimensão de 69 pacientes e 5 variáveis, após pré-processamento (Tabela 3). As distribuições dos atributos no conjunto de dados também foram realizadas em relação a cada subtipo de SMD, pelo sistema OMS 2016 (Figura 43), porém desta vez apenas 3 subtipos estão presentes nos dados.

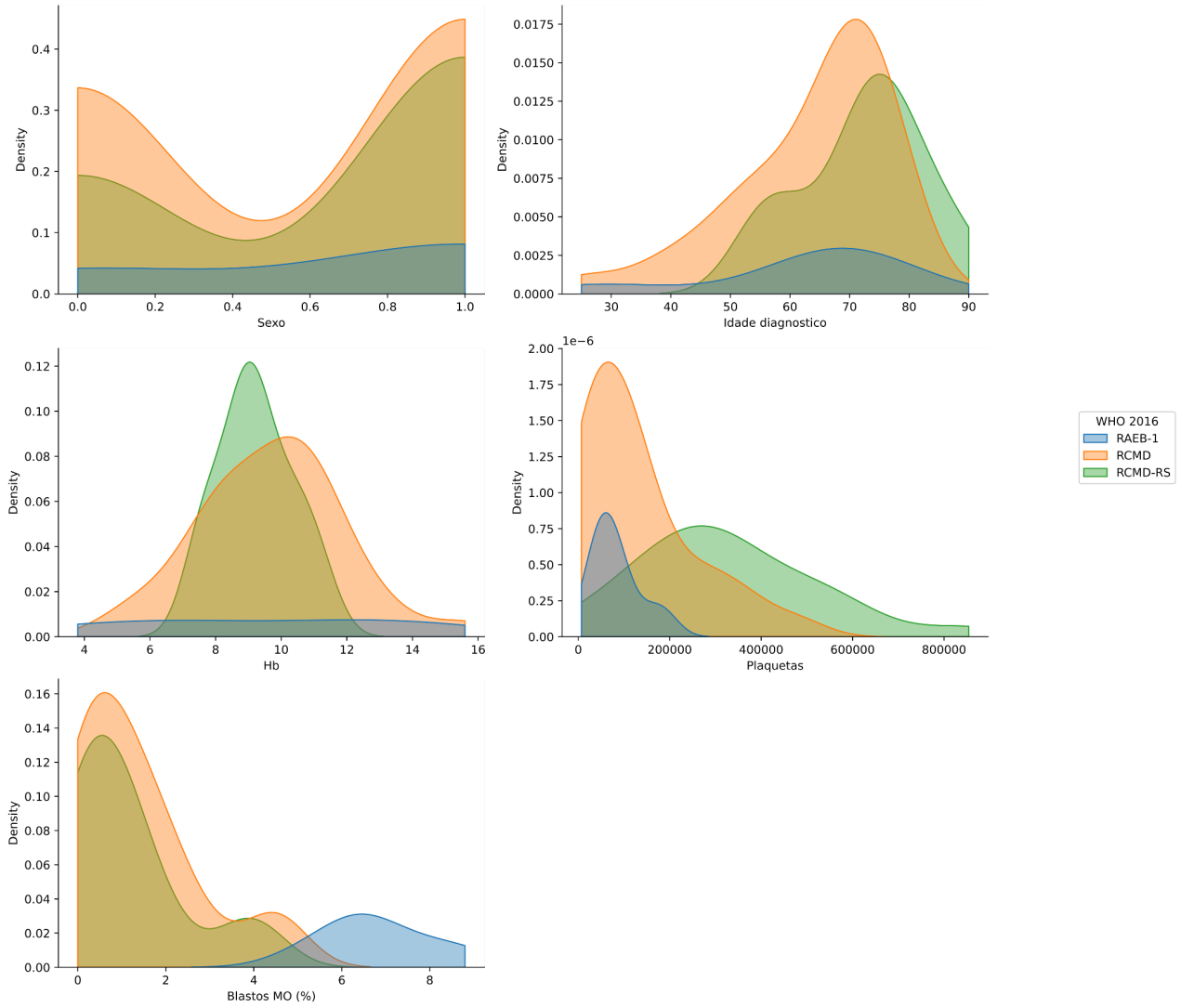

Figura 43. ***KDE plots com os valores de distribuição de cada um dos atributos utilizados no treinamento dos modelos.*** *Nesta visualização a distribuição de densidade de cada um dos atributos (antes da normalização) é separada por subtipo de SMD com o intuito de visualmente determinar se alguma feature isoladamente consegue distinguir os subtipos.*

O modelo melhor ranqueado neste cenário dentre os 34 (17 otimizados e 17 não-otimizados) foi o Gaussian Naive Bayes, empatado com sua versão otimizada (Figura 44). O modelo apresentou 0,95 de acurácia balanceada e 0,97 para média ponderada de AUC-ROC. A matriz de confusão bem como as curvas ROC do modelos treinados são ilustrados pelas Figura 45 e Figura 46, respectivamente.

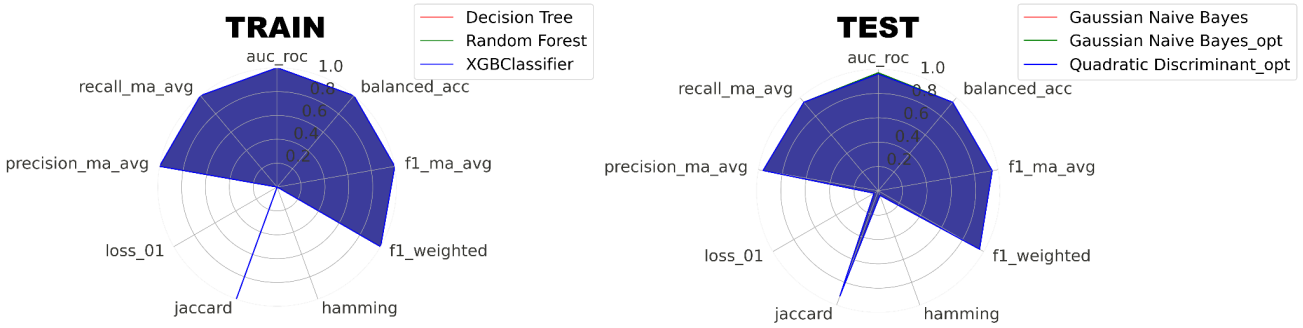

Figura 44. ***Radar plot ilustrando as métricas utilizadas para avaliar os modelos de classificação treinados com os dados de SMD do Hemocentro.*** À esquerda a visualização das métricas dos três melhores modelos utilizando os dados de treinamento na predição dos modelos, à direita utilizando os dados de teste. As métricas na imagem são as seguintes, a partir da posição central: AUC-ROC, acurácia balanceada, F1-score macro-média, F1-score ponderado, Perda Média de Hamming, Coeficiente de Similaridade de Jaccard, Perda de Classificação Zero-Um, Precisão, e Revocação.

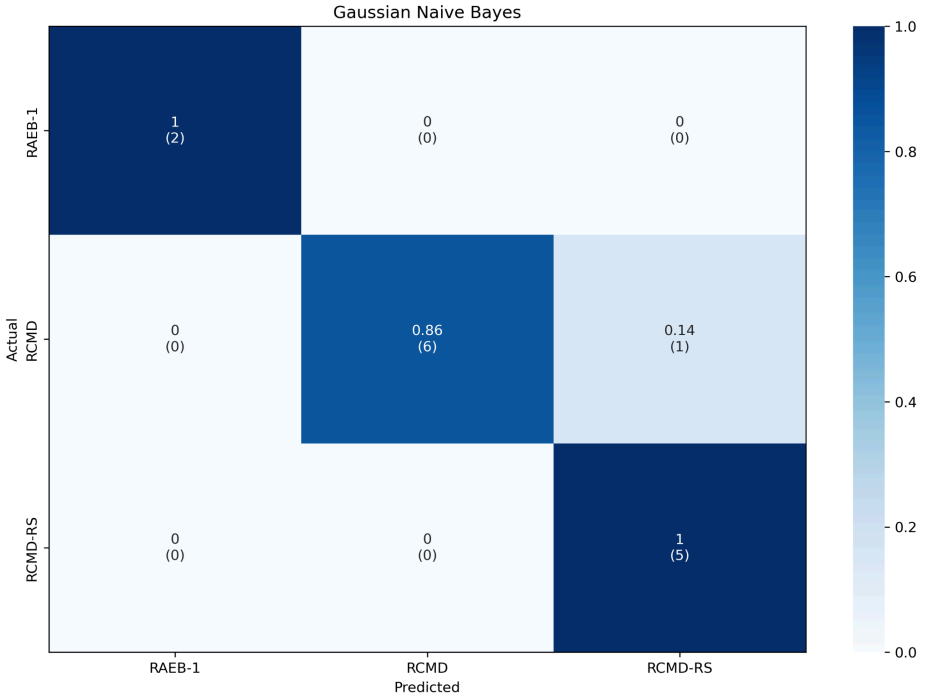

Figura 45. ***Matriz de confusão do melhor modelo de classificação treinado com os dados de SMD do Hemocentro.*** *Na imagem o valor acima dos parênteses simboliza o valor percentual e dentro dos parênteses o valor bruto, tanto de erros quanto de acertos.*

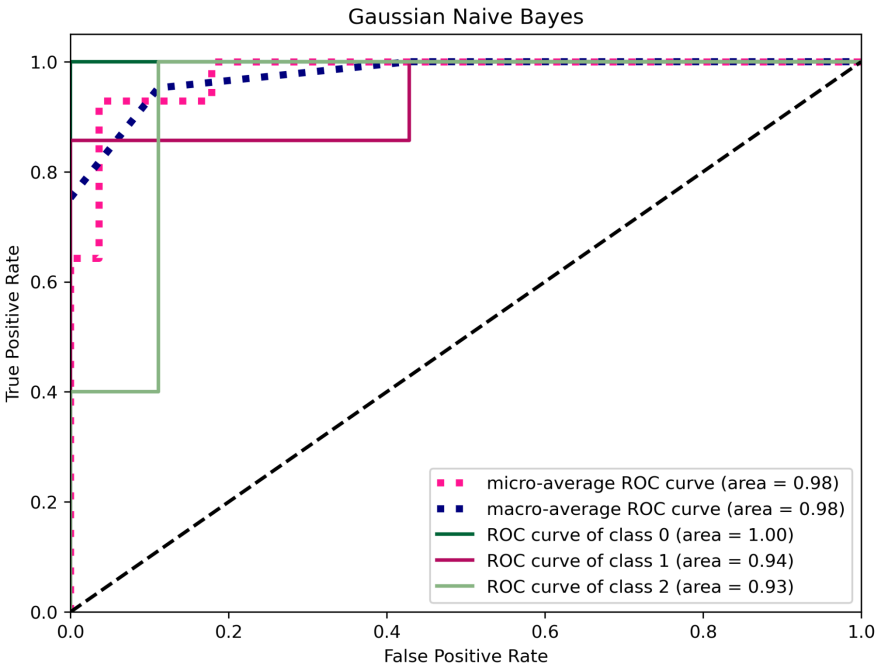

*Figura 46. **AUC-ROC do melhor modelo de classificação treinado com os dados de SMD do Hemocentro.** Na imagem o melhor desempenho foi para a classe 0 (RAEB-1) apresentando um valor de área igual à 1,00. Por outro lado, o menor desempenho foi o da classe 2 (RCMD-RS) com um valor de área igual à 0,93. A linha preta tracejada diagonal representa uma curva com AUC de 0.5, indicando o limiar de aleatoriedade.*

Este cenário apresentou o maior potencial obtido entre todos os realizados com os dados de SMD. Portanto, investigações posteriores serão realizadas para verificar este modelo mais a fundo.

2.4.3. Interpretabilidade e Explicabilidade dos melhores modelos

Os métodos ELI5, PDP, ICE, LIME e SHAP foram utilizados para acessar a interpretabilidade e explicabilidade de todos os modelos treinados. Para exemplificar a importância e relevância da interpretabilidade e explicabilidade dos modelos, o melhor modelo será utilizado, o Gaussian Naive Bayes, treinado com dados de SMD do Hemocentro.

O método SHAP é capaz de nos informar quais variáveis impactam mais as predições dos modelos e o quanto elas impactam. A Figura 47 informa que dentre os atributos utilizados, a quantidade de plaquetas é o atributo que mais impacta, seguido da porcentagem de blastos na medula óssea. Por outro lado, variáveis como o sexo são as que menos impactam na predição. Embora o impacto seja

pequeno ele ainda está presente, e talvez a performance do modelo não seria a mesma sem esse atributo. É possível testar se um dos atributos seriam o suficientes para a predição do problema em questão. Tendo isso em mente, todos os cenários apresentados na seção de “Treino e seleção dos modelos” foram testados nessa questão de peso dos atributos. Estes resultados encontram-se nos notebooks python hospedados no repositório deste trabalho, e para ilustrar o teste do cenário de SMD hemocentro será apresentado a seguir.

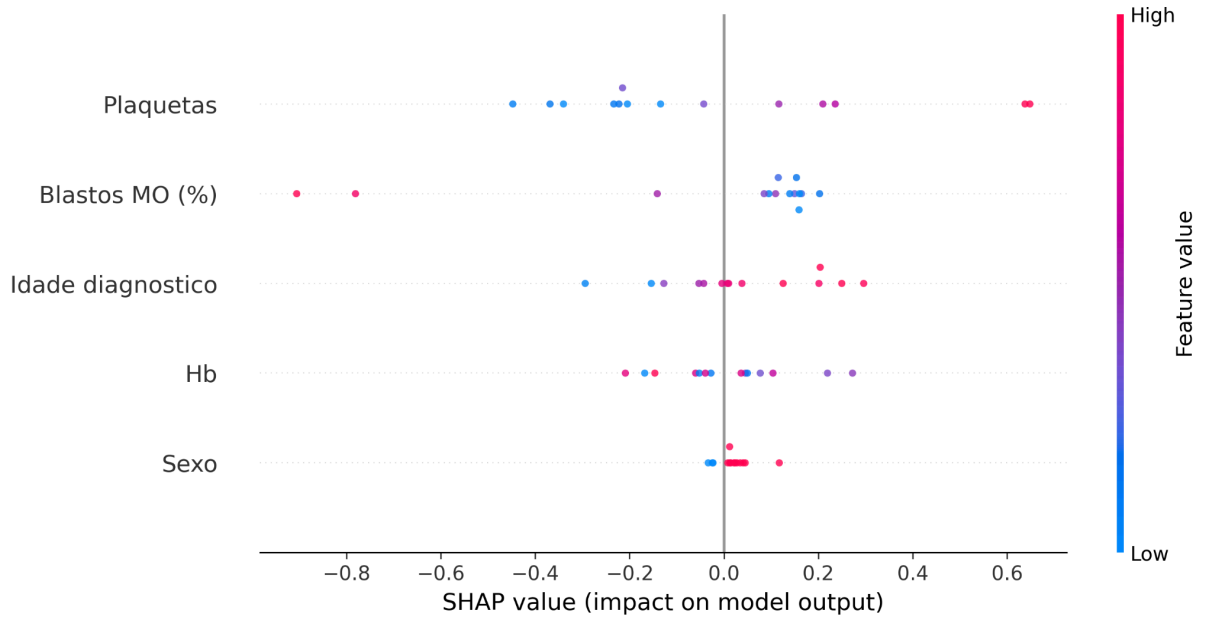

*Figura 47. **Imagem do KernelExplainer do SHAP para o modelo Gaussian Naive Bayes não-otimizado treinado com os dados de SMD do Hemocentro.** A quantidade de plaquetas é o atributo que mais impacta, seguido da porcentagem de blastos na medula óssea.*

Um modelo de regressão logística foi treinado em todos os cenários experimentais para avaliar o poder de predição de um único atributo, entretanto de maneira individual, diferentemente do apresentado anteriormente (Figura 47). Para o cenário de SMD Hemocentro a Tabela 5 apresenta os resultados gerados. Dentre os atributos testados, o único promissor é “Blastos MO (%)”. Entretanto, foi possível obter uma performance muito melhor ao se utilizar todas as variáveis juntas para treinar um único modelo. Isso confirma e justifica a utilização de métodos de aprendizado de máquina para lidar com os problemas explorados pelo presente trabalho.

Tabela 5. Modelos de classificação utilizados com suas respectivas siglas utilizadas no trabalho.

Atributos	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	jaccard	precision_ma_avg	recall_ma_avg
Sexo	0.35	0.33	0.16	0.14	0.09	0.1	0.33
Idade diagnostico	0.57	0.39	0.31	0.28	0.19	0.28	0.39
Hb	0.66	0.29	0.14	0.13	0.08	0.09	0.29
Plaquetas	0.87	0.33	0.15	0.13	0.08	0.1	0.33
Blastos MO (%)	0.89	0.67	0.56	0.61	0.56	0.5	0.67

Outra forma de se interpretar e explicar as predições de um modelo é do ponto de vista local, analisando subtipo a subtipo. Isso é possível graças a métodos como o LIME. Este método possibilita visualizar e avaliar o impacto das variáveis na predição de cada um dos alvos. Ademais, ele informa o intervalo em que essas variáveis são mais relevantes para um determinado subtipo, bem como a direção do impacto dele. Por exemplo, para predizer o subtipo RCMD (Citopenia Refratária Com Displasia Multilinhagem - *Refractory Cytopenia With Multilineage Dysplasia*) os valores de plaquetas são mais relevantes e apresentam um impacto diretamente maior na predição do modelo, por isso a barra na cor verde (Figura 48). Além disso, é reportado o intervalo em que essa informação é verdadeira, e que quando ocorre a probabilidade do paciente pertencer ao subtipo RCMD é maior.

O mesmo é verdadeiro, porém ao inverso para o valor de hemoglobina (Hb) e idade no diagnóstico. Isso pois, um valor de Hb no intervalo informado na Figura 48 aponta para um não pertencimento do paciente para o subtipo RCMD, por isso a barra na cor vermelha, já que é uma relação inversamente proporcional.

Outro ponto importante a ser destacado é que os valores que determinam os intervalos dos atributos estão na escala de z-score, uma vez que foram normalizados de tal forma para o treinamento dos modelos, com exceção do Sexo, o qual por ser uma variável dicotômica não foi normalizada, por não haver necessidade. Logo, a interpretação é imediata, e segundo a Figura 48, ser do sexo masculino (Sexo = 1) tem um impacto inversamente proporcional ao pertencimento do paciente para subtipo RCMD. Entretanto, não podemos assumir que o oposto é verdadeiro de imediato, ou seja, que o fato do paciente ser do sexo feminino ela teria maior propabilidade de pertencer ao subtipo em questão. Para tal, seria prudente investigar o impacto deste atributo (Sexo) nos outros subtipos analisados.

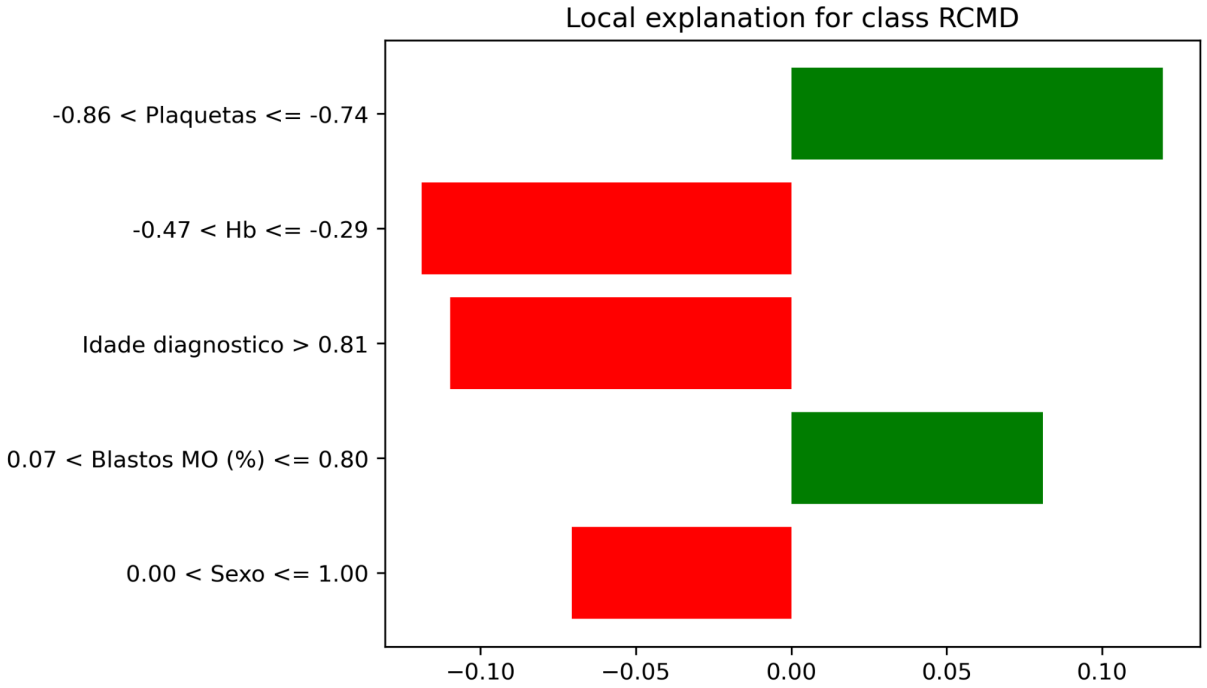

Figura 48. ***Imagem do pacote LIME para o modelo Gaussian Naive Bayes não-otimizado treinado com os dados de SMD do Hemocentro.*** *A quantidade de plaquetas é o atributo que mais impacta diretamente (em verde) na probabilidade de predição do subtipo RCMD (Citopenia Refratária Com Displasia Multilinhagem), seguido por hemoglobina e idade no diagnóstico, as quais impactam inversamente (em vermelho) na probabilidade de predição do subtipo RCMD.*

2.5. Conclusão, Próximos Passos e Doutorado Sanduíche

2.6. Conclusão e Próximos Passos

O uso do aprendizado de máquina aplicado em problemáticas ligadas às malignidades mieloides tem sido comum. Em 2019, biomarcadores genômicos foram utilizados para predizer a resistência de agentes hipometilantes em pacientes com SMD (NAZHA *et al.*, 2019). No caso específico de LMA, um trabalho recente utilizou o ML para relacionar os pacientes, com os quimioterápicos mais indicados (LEE *et al.*, 2018). A classificação de pacientes em subtipos de LMA e SMD, com dados laboratoriais e clínicos usando aprendizagem de Máquina ainda não foi explorada, abrindo assim a oportunidade deste projeto de doutorado (JAGADEV; VIRANI, 2017).

Os resultados deste trabalho demonstram e avançam um passo para a viabilidade da área de medicina personalizada, uma vez que os modelos

experimentais criados podem ser utilizados tanto na pesquisa básica quanto na aplicação direta em diagnósticos mais precisos desta doença. Em especial, o modelo treinado para predição do IPSS, o qual conseguiu realizar predições sem a utilização de dados citogenéticos, essenciais para a classificação utilizando o método convencional. Vale ressaltar que gerar dados citogenéticos é mais custoso do que um exame de hemograma. Em 2024, um exame citogenético de cariótipos em cultura de longa duração custa em média cerca de R$ 360,00 (02.02.10.001-4, 2024), já um exame de sangue tem o valor de R$ 4,11 (02.02.02.038-0, 2024), quase 88 vezes mais caro, segundo valores do DATASUS.

Outro modelo interessante foi o modelo obtido com os dados de SMD do Hemocentro da UNICAMP, o qual realizou predições para três dos oito subtipos de SMD com 0,95 de acurácia balanceada e 0,97 para média ponderada de AUC-ROC. Já para LMA, o melhor modelo criado foi utilizando os dados do TCGA apresentando acurácia balanceada de 0,57 e 0,78 para média ponderada de AUC-ROC. Infelizmente, estes valores estão aquém do ideal, embora pelo lado otimista do espectro este resultado não seja de todo perdido e aponte para uma necessidade de mudança de foco para a análise dos dados de LMA. Dentre as possibilidades, está a ideia de combinar os subtipos similares a fim de reduzir a complexidade do problema.

A área de aprendizado de máquina é uma área prática, portanto nossos modelos selecionados serão exportados e disponibilizados para a comunidade para que ele seja utilizado. Uma interface web poderá ser desenvolvida no futuro para facilitar a interação do usuário com nosso modelo preditivo. O framework web Django é um candidato óbvio a ser utilizado para essa tarefa devido a sua flexibilidade de uso, assim como ao seu suporte e documentação. Por fim, o site poderá ser hospedado em um dos servidores do LaBIS (grupo do Prof. Marcelo Brandão). A interação esperada seria a seleção do tipo tumoral (LMA ou SMD), a introdução dos dados clínicos do paciente e o pressionamento de um botão para início da predição. Dessa forma, seria retornado ao usuário a classificação e a probabilidade associada, juntamente com a explicação da predição evidenciando quais variáveis contribuem mais/menos com a predição apresentada.

Este trabalho foi financiado pelo CNPq sob o número de processo 140610/2020-0. Além disso, ele contou com a colaboração com o grupo de pesquisa da Profª. Drª Sara Saad, do Hemocentro da UNICAMP. O grupo é nacionalmente um

dos maiores no estudo de alterações celulares e moleculares do sangue. Posteriormente obtivemos a colaboração do grupo do Prof. Dr. Ronald Feitosa Pinheiro da Universidade Federal do Ceará (UFC), o qual gentilmente cedeu dados valiosos de pacientes de SMD para análise, impulsionando o bom andamento deste trabalho.

2.7. Doutorado sanduíche

Com o restabelecimento de algumas atividades após o fim da pandemia foram retomados os planos até então postergados para o doutorado sanduíche. Ele foi realizado no grupo do Prof. Dr. Jan Baumbach, diretor do *Institute for Computational Systems Biology*, Universidade de Hamburgo, Alemanha. A Modelagem Preditiva voltada à medicina de precisão é uma das áreas de pesquisa do grupo do Prof. Baumbach. Essa oportunidade foi possível graças a uma parceria com o grupo estrangeiro e ao financiamento da *Bayer Foundation Carl-Duisberg-Fellowship for Medical Sciences*, obtido através de uma seleção internacional, na qual nosso projeto foi selecionado (Anexo A). O conhecimento obtido durante essa oportunidade foi de fundamental importância para o prosseguimento desta tese, bem como no desenvolvimento de projetos futuros na área de aprendizado de máquina na saúde.

Dois projetos foram desenvolvidos durante este período no grupo do Prof. Baumbach. Um deles envolvia a colaboração no desenvolvimento de uma ferramenta de aprendizagem de máquina não supervisionada, a qual utiliza a técnica de bicluster para estratificar pacientes com base no perfil de expressão gênica. Esta ferramenta denominada UnPaSt (*Unsupervised Patient Stratification*) está em fase final de desenvolvimento, e demonstrou ser superior às outras alternativas disponíveis em nossos benchmarks realizados, o artigo correspondente está em fase de submissão. O segundo projeto envolveu a elaboração de um pacote Python para inferência de redes chave regulatórias diferenciais e candidatos de reproposição de medicamentos a partir de dados de scRNA-seq (OUBOUNYT *et al.*, 2023).

2.8. Referências

02.02.02.038-0. **SIGTAP - Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e OPM do SUS**. [*S. l.: s. n.*], 2024. Disponível em: http://sigtap.datasus.gov.br/tabela-unificada/app/sec/procedimento/exibir/020202038 0/06/2024. Acesso em: 27 jun. 2024.
02.02.10.001-4. **SIGTAP - Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e OPM do SUS**. [*S. l.: s. n.*], 2024. Disponível em: http://sigtap.datasus.gov.br/tabela-unificada/app/sec/procedimento/exibir/020210001 4/06/2024. Acesso em: 27 jun. 2024.
ARBER, D. A.; ORAZI, A.; HASSERJIAN, R. P.; BOROWITZ, M. J.; CALVO, K. R.; KVASNICKA, H.-M.; WANG, S. A.; BAGG, A.; BARBUI, T.; BRANFORD, S.; BUESO-RAMOS, C. E.; CORTES, J. E.; DAL CIN, P.; DINARDO, C. D.; DOMBRET, H.; DUNCAVAGE, E. J.; EBERT, B. L.; ESTEY, E. H.; FACCHETTI, F.; FOUCAR, K.; GANGAT, N.; GIANELLI, U.; GODLEY, L. A.; GÖKBUGET, N.; GOTLIB, J.; HELLSTRÖM-LINDBERG, E.; HOBBS, G. S.; HOFFMAN, R.; JABBOUR, E. J.; KILADJIAN, J.-J.; LARSON, R. A.; LE BEAU, M. M.; LOH, M. L.-C.; LÖWENBERG, B.; MACINTYRE, E.; MALCOVATI, L.; MULLIGHAN, C. G.; NIEMEYER, C.; ODENIKE, O. M.; OGAWA, S.; ORFAO, A.; PAPAEMMANUIL, E.; PASSAMONTI, F.; PORKKA, K.; PUI, C.-H.; RADICH, J. P.; REITER, A.; ROZMAN, M.; RUDELIUS, M.; SAVONA, M. R.; SCHIFFER, C. A.; SCHMITT-GRAEFF, A.; SHIMAMURA, A.; SIERRA, J.; STOCK, W. A.; STONE, R. M.; TALLMAN, M. S.; THIELE, J.; TIEN, H.-F.; TZANKOV, A.; VANNUCCHI, A. M.; VYAS, P.; WEI, A. H.; WEINBERG, O. K.; WIERZBOWSKA, A.; CAZZOLA, M.; DÖHNER, H.; TEFFERI, A. International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. **Blood**, [*s. l.*], v. 140, n. 11, p. 1200–1228, 2022.
ARBER, D. A.; ORAZI, A.; HASSERJIAN, R.; THIELE, J.; BOROWITZ, M. J.; LE BEAU, M. M.; BLOOMFIELD, C. D.; CAZZOLA, M.; VARDIMAN, J. W. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. **Blood**, [*s. l.*], v. 127, n. 20, p. 2391–2405, 2016.
AREOSA, I.; TORGO, L. Visual Interpretation of Regression Error. *Em*: PROGRESS IN ARTIFICIAL INTELLIGENCE, 2019, Cham. **Anais** [...]. Cham: Springer International Publishing, 2019. p. 473–485.
BENNETT, J. M.; CATOVSKY, D.; DANIEL, M.-T.; FLANDRIN, G.; GALTON, D. a. G.; GRALNICK, H. R.; SULTAN, C. Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. **British Journal of Haematology**, [*s. l.*], v. 33, n. 4, p. 451–458, 1976.
BERGSTRA, J.; BENGIO, Y. **Journal of machine learning research : JMLR.** [*S. l.*]: MIT Press, 2012.(, v. 10). v. 13. Disponível em: http://www.jmlr.org/papers/v13/bergstra12a.html.
BERNARD, E.; TUECHLER, H.; GREENBERG, P. L.; HASSERJIAN, R. P.; ARANGO OSSA, J. E.; NANNYA, Y.; DEVLIN, S. M.; CREIGNOU, M.; PINEL, P.; MONNIER, L.; GUNDEM, G.; MEDINA-MARTINEZ, J. S.; DOMENICO, D.; JÄDERSTEN, M.; GERMING, U.; SANZ, G.; VAN DE LOOSDRECHT, A. A.; KOSMIDER, O.; FOLLO, M. Y.; THOL, F.; ZAMORA, L.; PINHEIRO, R. F.; PELLAGATTI, A.; ELIAS, H. K.; HAASE, D.; GANSTER, C.; ADES, L.; TOBIASSON, M.; PALOMO, L.; DELLA PORTA, M. G.; TAKAORI-KONDO, A.; ISHIKAWA, T.; CHIBA, S.; KASAHARA, S.; MIYAZAKI, Y.; VIALE, A.; HUBERMAN, K.; FENAUX, P.; BELICKOVA, M.; SAVONA, M. R.; KLIMEK, V. M.; SANTOS, F. P. S.; BOULTWOOD, J.; KOTSIANIDIS, I.;

SANTINI, V.; SOLÉ, F.; PLATZBECKER, U.; HEUSER, M.; VALENT, P.; OHYASHIKI, K.; FINELLI, C.; VOSO, M. T.; SHIH, L.-Y.; FONTENAY, M.; JANSEN, J. H.; CERVERA, J.; GATTERMANN, N.; EBERT, B. L.; BEJAR, R.; MALCOVATI, L.; CAZZOLA, M.; OGAWA, S.; HELLSTRÖM-LINDBERG, E.; PAPAEMMANUIL, E. Molecular International Prognostic Scoring System for Myelodysplastic Syndromes. **NEJM Evidence**, [*s. l.*], v. 1, n. 7, p. EVIDoa2200008, 2022.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R. A.; STONE, C. J. **Classification and Regression Trees**. New York: Chapman and Hall/CRC, 1984. 1984.

CERAMI, E.; GAO, J.; DOGRUSOZ, U.; GROSS, B. E.; SUMER, S. O.; AKSOY, B. A.; JACOBSEN, A.; BYRNE, C. J.; HEUER, M. L.; LARSSON, E.; ANTIPIN, Y.; REVA, B.; GOLDBERG, A. P.; SANDER, C.; SCHULTZ, N. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. **Cancer Discovery**, [*s. l.*], v. 2, n. 5, p. 401–404, 2012.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, [*s. l.*], v. 16, p. 321–357, 2002.

CHEN, X.; FU, R.; SHAO, Q.; CHEN, Y.; YE, Q.; LI, S.; HE, X.; ZHU, J. Application of artificial intelligence to pancreatic adenocarcinoma. **Frontiers in Oncology**, [*s. l.*], v. 12, 2022. Disponível em: https://www.frontiersin.org/articles/10.3389/fonc.2022.960056. Acesso em: 15 ago. 2023.

CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System** *Em*: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2016, New York, NY, USA. **Anais** [...]. New York, NY, USA: Association for Computing Machinery, 2016. p. 785–794. Disponível em: https://doi.org/10.1145/2939672.2939785. Acesso em: 20 jul. 2022.

EXPLAINABLE AND INTERPRETABLE MODELS IN COMPUTER VISION AND MACHINE LEARNING. [*S. l.: s. n.*], [*s. d.*]. [*s. d.*]. Disponível em: https://link.springer.com/book/10.1007/978-3-319-98131-4. Acesso em: 20 jul. 2022.

FREUND, Y.; SCHAPIRE, R. E. A desicion-theoretic generalization of on-line learning and an application to boosting. *Em*: COMPUTATIONAL LEARNING THEORY, 1995, Berlin, Heidelberg. **Anais** [...]. Berlin, Heidelberg: Springer, 1995. p. 23–37.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, [*s. l.*], v. 29, n. 5, p. 1189–1232, 2001.

GANIE, S. M.; MALIK, M. B. An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators. **Healthcare Analytics**, [*s. l.*], v. 2, p. 100092, 2022.

GÉRON, A. **Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems**. [*S. l.: s. n.*], 2017. 2017. Disponível em: https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/.

GOLDSTEIN, A.; KAPELNER, A.; BLEICH, J.; PITKIN, E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. **Journal of Computational and Graphical Statistics**, [*s. l.*], v. 24, n. 1, p. 44–65, 2015.

GREENBERG, P. L.; TUECHLER, H.; SCHANZ, J.; SANZ, G.; GARCIA-MANERO, G.; SOLÉ, F.; BENNETT, J. M.; BOWEN, D.; FENAUX, P.; DREYFUS, F.; KANTARJIAN, H.; KUENDGEN, A.; LEVIS, A.; MALCOVATI, L.; CAZZOLA, M.; CERMAK, J.; FONATSCH, C.; LE BEAU, M. M.; SLOVAK, M. L.; KRIEGER, O.;

LUEBBERT, M.; MACIEJEWSKI, J.; MAGALHAES, S. M. M.; MIYAZAKI, Y.; PFEILSTÖCKER, M.; SEKERES, M.; SPERR, W. R.; STAUDER, R.; TAURO, S.; VALENT, P.; VALLESPI, T.; VAN DE LOOSDRECHT, A. A.; GERMING, U.; HAASE, D. Revised International Prognostic Scoring System for Myelodysplastic Syndromes. **Blood**, [*s. l.*], v. 120, n. 12, p. 2454–2465, 2012.
HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. New York, NY: Springer, 2009. 2009.(Springer Series in Statistics). Disponível em: http://link.springer.com/10.1007/978-0-387-84858-7. Acesso em: 15 ago. 2023.
HOFFBRAND, A. V.; MOSS, P. A. H. **Hoffbrand's essential haematology Seventh edition**. [*S. l.: s. n.*], 2015. 2015. Disponível em: https://www.wiley.com/en-br/Hoffbrand%27s+Essential+Haematology%2C+7th+Edition-p-9781118408674.
INCA. **Quais os efeitos colaterais da quimioterapia? | INCA - Instituto Nacional de Câncer**. , 2019. Disponível em: https://www.inca.gov.br/perguntas-frequentes/quais-os-efeitos-colaterais-da-quimioterapia. Acesso em: 23 set. 2019.
JAGADEV, P.; VIRANI, H. G. Detection of leukemia and its types using image processing and machine learning. *Em*: 2017 INTERNATIONAL CONFERENCE ON TRENDS IN ELECTRONICS AND INFORMATICS (ICEI), 2017, [*s. l.*], . **Anais** [...]. [*S. l.*]: IEEE, 2017. p. 522–526. Disponível em: http://ieeexplore.ieee.org/document/8300983/.
KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. **LightGBM: A Highly Efficient Gradient Boosting Decision Tree** *Em*: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2017, [*s. l.*], . **Anais** [...]. [*S. l.*]: Curran Associates, Inc., 2017. Disponível em: https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html. Acesso em: 28 jun. 2023.
KORN, C.; MÉNDEZ-FERRER, S. Myeloid malignancies and the microenvironment. **Blood**, [*s. l.*], v. 129, n. 7, p. 811–822, 2017.
KRATZ, A.; FERRARO, M.; SLUSS, P. M.; LEWANDROWSKI, K. B. Normal Reference Laboratory Values. **New England Journal of Medicine**, [*s. l.*], v. 351, n. 15, p. 1548–1563, 2004.
LEE, S.-I.; CELIK, S.; LOGSDON, B. A.; LUNDBERG, S. M.; MARTINS, T. J.; OEHLER, V. G.; ESTEY, E. H.; MILLER, C. P.; CHIEN, S.; DAI, J.; SAXENA, A.; BLAU, C. A.; BECKER, P. S. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. **Nature Communications**, [*s. l.*], v. 9, n. 1, p. 42, 2018.
LIAPIS, K.; PAPADOPOULOS, V.; VRACHIOLIAS, G.; GALANOPOULOS, A. G.; PAPOUTSELIS, M.; PAPAGEORGIOU, S. G.; DIAMANTOPOULOS, P. T.; PAPPA, V.; VINIOU, N.-A.; KOURAKLI, A.; TSOKANAS, D.; VASSILAKOPOULOS, T. P.; HATZIMICHAEL, E.; BOURONIKOU, E.; XIMERI, M.; PONTIKOGLOU, C.; MEGALAKAKI, A.; ZIKOS, P.; PANAYIOTIDIS, P.; DIMOU, M.; KARAKATSANIS, S.; PAPAIOANNOU, M.; VARDI, A.; KONTOPIDOU, F.; HARCHALAKIS, N.; ADAMOPOULOS, I.; SYMEONIDIS, A.; KOTSIANIDIS, I. Refinement of prognosis and the effect of azacitidine in intermediate-risk myelodysplastic syndromes. **Blood Cancer Journal**, [*s. l.*], v. 11, n. 2, p. 1–7, 2021.
LOPES, M. R.; PEREIRA, J. K. N.; DE MELO CAMPOS, P.; MACHADO-NETO, J. A.; TRAINA, F.; SAAD, S. T. O.; FAVARO, P. De novo AML exhibits greater microenvironment dysregulation compared to AML with myelodysplasia-related changes. **Scientific reports**, [*s. l.*], v. 7, p. 40707, 2017.

LUNDBERG, S.; LEE, S.-I. **A Unified Approach to Interpreting Model Predictions**. Versão arXiv:1705.07874. [*S. l.*]: arXiv, 2017. Disponível em: http://arxiv.org/abs/1705.07874. Acesso em: 20 jul. 2022.
MA, X. Epidemiology of myelodysplastic syndromes. **The American journal of medicine**, [*s. l.*], v. 125, n. 7 Suppl, p. S2-5, 2012.
MULLAINATHAN, S.; OBERMEYER, Z. Does Machine Learning Automate Moral Hazard and Error?. **The American economic review**, [*s. l.*], v. 107, n. 5, p. 476–480, 2017.
MURATI, A.; BRECQUEVILLE, M.; DEVILLIER, R.; MOZZICONACCI, M.-J.; GELSI-BOYER, V.; BIRNBAUM, D. Myeloid malignancies: mutations, models and management. **BMC Cancer**, [*s. l.*], v. 12, n. 1, p. 304, 2012.
NAZHA, A.; SEKERES, M. A.; BEJAR, R.; RAUH, M. J.; OTHUS, M.; KOMROKJI, R. S.; BARNARD, J.; HILTON, C. B.; KERR, C. M.; STEENSMA, D. P.; DEZERN, A.; ROBOZ, G.; GARCIA-MANERO, G.; ERBA, H.; EBERT, B. L.; MACIEJEWSKI, J. P. Genomic Biomarkers to Predict Resistance to Hypomethylating Agents in Patients With Myelodysplastic Syndromes Using Artificial Intelligence. **JCO precision oncology**, [*s. l.*], v. 3, 2019. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/31663066.
NIH. **Impact of NIH Research | National Institutes of Health (NIH)**. , 2019. Disponível em: https://www.nih.gov/about-nih/what-we-do/impact-nih-research. Acesso em: 29 ago. 2019.
NTZIACHRISTOS, P.; MULLENDERS, J.; TRIMARCHI, T.; AIFANTIS, I. Mechanisms of Epigenetic Regulation of Leukemia Onset and Progression. *Em*: ADVANCES IN IMMUNOLOGY. [*S. l.: s. n.*], 2013. v. 117, p. 1–38. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/23611284 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3830954 https://linkinghub.elsevier.com/retrieve/pii/B9780124105249000013.
OUBOUNYT, M.; ADLUNG, L.; PATRONI, F.; WENKE, N. K.; MAIER, A.; HARTUNG, M.; BAUMBACH, J.; ELKJAER, M. L. Inference of differential key regulatory networks and mechanistic drug repurposing candidates from scRNA-seq data with SCANet. **Bioinformatics**, [*s. l.*], v. 39, n. 11, p. btad644, 2023.
PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; MÜLLER, A.; NOTHMAN, J.; LOUPPE, G.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, É. Scikit-learn: Machine Learning in Python. [*s. l.*], 2012. Disponível em: http://arxiv.org/abs/1201.0490.
PROKHORENKOVA, L.; GUSEV, G.; VOROBEV, A.; DOROGUSH, A. V.; GULIN, A. **CatBoost: unbiased boosting with categorical features** Versão arXiv:1706.09516. [*S. l.*]: arXiv, 2019. Disponível em: http://arxiv.org/abs/1706.09516. Acesso em: 28 jun. 2023.
REILLY, B.; TANAKA, T. N.; DIEP, D.; YEERNA, H.; TAMAYO, P.; ZHANG, K.; BEJAR, R. DNA methylation identifies genetically and prognostically distinct subtypes of myelodysplastic syndromes. **Blood Advances**, [*s. l.*], v. 3, n. 19, p. 2845–2858, 2019.
RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. **"Why Should I Trust You?": Explaining the Predictions of Any Classifier** *Em*: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2016, New York, NY, USA. **Anais** [...]. New York, NY, USA: Association for Computing Machinery, 2016. p. 1135–1144. Disponível em:

https://doi.org/10.1145/2939672.2939778. Acesso em: 20 jul. 2022.
RICHARD A. DUSCHL; HEIDI A. SCHWEINGRUBER; ANDREW W. SHOUSE. **Taking Science to School: Learning and Teaching Science in Grades K-8**. [*S. l.*]: National Academies of Sciences, Engineering, and Medicine, 2007. 2007. Disponível em: https://www.nap.edu/read/11625/chapter/8. Acesso em: 15 ago. 2023.
SHAPLEY, L. S. 17. A Value for n-Person Games. *Em*: 17. A VALUE FOR N-PERSON GAMES. [*S. l.*]: Princeton University Press, 1953. p. 307–318. Disponível em: https://www.degruyter.com/document/doi/10.1515/9781400881970-018/html. Acesso em: 20 jul. 2022.
THE MYTHOS OF MODEL INTERPRETABILITY - ACM QUEUE. [*S. l.: s. n.*], [*s. d.*]. Disponível em: https://queue.acm.org/detail.cfm?id=3241340. Acesso em: 20 jul. 2022.
TOMEK, I. Two Modifications of CNN. **IEEE Transactions on Systems, Man, and Cybernetics**, [*s. l.*], v. SMC-6, n. 11, p. 769–772, 1976.
WALTER, R. B.; OTHUS, M.; BURNETT, A. K.; LÖWENBERG, B.; KANTARJIAN, H. M.; OSSENKOPPELE, G. J.; HILLS, R. K.; VAN MONTFORT, K. G. M.; RAVANDI, F.; EVANS, A.; PIERCE, S. R.; APPELBAUM, F. R.; ESTEY, E. H. Significance of FAB subclassification of “acute myeloid leukemia, NOS” in the 2008 WHO classification: analysis of 5848 newly diagnosed patients. **Blood**, [*s. l.*], v. 121, n. 13, p. 2424, 2013.
WEST, A. H.; GODLEY, L. A.; CHURPEK, J. E. Familial myelodysplastic syndrome/acute leukemia syndromes: a review and utility for translational investigations. **Annals of the New York Academy of Sciences**, [*s. l.*], v. 1310, n. 1, p. 111–118, 2014.
YAUNEY, G.; SHAH, P. **Reinforcement Learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regimen Selection**. , 2018. Disponível em: http://proceedings.mlr.press/v85/yauney18a.html.
ZITNIK, M.; NGUYEN, F.; WANG, B.; LESKOVEC, J.; GOLDENBERG, A.; HOFFMAN, M. M. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. **arXiv e-prints**, [*s. l.*], 2018. Disponível em: http://arxiv.org/abs/1807.00123 http://dx.doi.org/10.1016/j.inffus.2018.09.012.

3. Capítulo 2. Comprehensive Glycoprofiling of Oral Tumors Associates N-Glycosylation With Lymph Node Metastasis and Patient Survival

Paper published on MCP (2023) under a Creative Commons Attribution – NonCommercial license.
DOI: https://doi.org/10.1016/j.mcpro.2023.100586
Supplemental material
Published with a CC BY-NC-ND 4.0 license.

Carolina Moretto Carnielli 1, Thayná Melo de Lima Morais 2, **Fábio Malta de Sá Patroni** 3, Ana Carolina Prado Ribeiro 4,5, Thaís Bianca Brandão 4, Evandro Sobroza 4, Leandro Luongo Matos 6, Luiz Paulo Kowalski 7,8, Adriana Franco Paes Leme 1*, Rebeca Kawahara 9,10#*, Morten Thaysen-Andersen 9,10#*

Affiliations

1 Laboratório de Espectrometria de Massas, Laboratório Nacional de Biociências (LNBio), Centro Nacional de Pesquisa em Energia e Materiais (CNPEM), Campinas, 13083-970 SP, Brazil.
2 Oral Pathology, Department of Oral Diagnosis, Piracicaba Dental School, University of Campinas, Brazil.
3 Molecular Biology and Genetic Engineering Center, University of Campinas, Campinas, Brazil.
4 Serviço de Odontologia Oncológica, Instituto do Câncer do Estado de São Paulo, ICESP-FMUSP, São Paulo, 01246-000 SP, Brazil.
5 Universidade Brasil, Fernandópolis, 15600-000 SP, Brazil.
6 Serviço de Cirurgia de Cabeça e Pescoço, Instituto do Câncer do Estado de São Paulo, ICESP-FMUSP, São Paulo, 01246-000 SP, Brazil.
7 Departamento de Cirurgia de Cabeça e Pescoço e Otorrinolaringologia, A.C. Camargo Cancer Center, São Paulo, SP, 01509-900, Brazil.
8 Departamento de Cirurgia de Cabeça e Pescoço, Faculdade de Medicina, Universidade de São Paulo - USP, São Paulo, SP, 01246-903, Brazil.
9 School of Natural Sciences, Macquarie University, Sydney, NSW-2109, Australia

10 Institute for Glyco-core Research (iGCORE), Nagoya University, Nagoya, 464-8601, Japan

# These authors contributed equally to this work.
*Joint corresponding authors.

Corresponding Authors: Adriana Franco Paes Leme, Rebeca Kawahara, Morten Thaysen-Andersen.

Abstract

While altered protein glycosylation is regarded a trait of oral squamous cell carcinoma (OSCC), the heterogeneous and dynamic glycoproteome of tumor tissues from OSCC patients remain unmapped. To this end, we here employ an integrated multi-omics approach comprising unbiased and quantitative glycomics and glycoproteomics applied to a cohort of resected primary tumor tissues from OSCC patients with (n = 19) and without (n = 12) lymph node metastasis. While all tumor tissues displayed relatively uniform N-glycome profiles suggesting overall stable global N-glycosylation during disease progression, altered expression of six sialylated N-glycans was found to correlate with lymph node metastasis. Notably, glycoproteomics and advanced statistical analyses uncovered altered site-specific N-glycosylation revealing previously unknown associations with several clinicopathological features. Importantly, the glycomics and glycoproteomics data unveiled that comparatively high abundance of two core-fucosylated and sialylated N-glycans (Glycan 40a and Glycan 46a) and one N-glycopeptide from fibronectin were associated with low patient survival, while a relatively low abundance of N-glycopeptides from both afamin and CD59 were also associated with poor survival. This study provides insight into the complex OSCC tissue N-glycoproteome, thereby forming an important resource to further explore the underpinning disease mechanisms and uncover new prognostic glycomarkers for OSCC.

3.1. Introduction

Oral squamous cell carcinoma (OSCC) is the most common type of head and neck cancer (1). With more than 377,000 new cases and 170,000 deaths annually worldwide, OSCC is an aggressive disease with a dishearteningly low five-year survival rate of 50% (2, 3), mainly due to lymph node metastasis and loco-regional failures. OSCC prognostication is currently based on the clinical staging system of tumor-lymph node-metastasis (TNM system) of the disease. However, this system has several flaws, as patients assigned with the same TNM stage may present with different clinical features and experience different outcomes (4, 5). Thus, identification of molecular signatures that may assist in a more precise staging and prognosis of patients with OSCC is urgently needed.

Histopathological analysis of formalin-fixed paraffin-embedded (FFPE) tissues remains the principal method for diagnosis and prognosis of OSCC patients (6, 7). FFPE tissue slides is known to preserve the cellular morphology and molecular features of the primary tumor tissues, and have therefore been explored as a source for biomarker investigations (8, 9). We recently used sensitive mass spectrometry (MS)-based proteomics to map different areas of FFPE tissues from OSCC patients, which revealed key proteins with potential prognostic value as demonstrated by their associations with various clinicopathological parameters; the prognostic value of these proteins was also demonstrated by targeted proteomics of saliva from OSCC patients with and without lymph node metastasis (10). Our findings also revealed that specific glycoproteins i.e. ITGAV and COL6A1 from the tumor stroma, and COL1A2 from tumor cells associate with lymph node metastasis and type of disease treatment, respectively.

Evidence is emerging that not only the protein abundance, but also the glycosylation patterns of proteins and the levels of the underlying glycosylation enzymes, are altered in OSCC (11, 12). Moreover, glycosylation changes have previously been associated with altered adhesion behavior, migration and metastasis of oral cancer cells as well as OSCC disease progression (12–14). Aberrant N-glycosylation of E-cadherin, a key marker of epithelial mesenchymal transition in cancer development (15), and reduced adhesion of human salivary epidermoid carcinoma cells were found to result from the overexpression of dolichol-P-dependent N-acetylglucosamine-1-phosphate transferase (DPAGT1), an enzyme that initiates

the synthesis of the lipid-linked oligosaccharide precursor for protein N-glycosylation in the endoplasmic reticulum (16). Overexpression of DPAGT1 detected in specimens from resected OSCC tumors in the oral cavity was also found to be associated with an aberrant activation of the canonical Wnt signaling pathway that contributes to the development and progression of many human cancers (13, 17, 18). Moreover, site-specific glycoprofiling of a gingival carcinoma cell line uncovered elevated sialylation and reduced fucosylation relative to a non-cancerous oral epithelial cell line from gingiva (19). The study also revealed that B7-H3 (also known as B7 homolog 3 or CD276 isoform 1) knockdown suppresses tumor cell proliferation, and that B7-H3 restoration enhances tumor growth. Finally, elevated sialidase activity was observed in saliva of oral cancer patients as compared to healthy individuals and patients with precancerous conditions (20). While these studies collectively indicate that aberrant protein glycosylation is linked to the onset and progression of oral cancer, altered protein glycosylation has not previously been associated with patient prognosis including lymph node metastasis status and survival outcome.

Herein, we employ an integrative glyco-centric multi-omics approach to quantitatively profile the heterogenous N-glycoproteome in primary tumor tissues from OSCC patients with (N+) or without (N0) lymph node metastasis to establish the glycan fine structures, their site-specific microheterogeneity, and their disease dynamics. Our study provides a comprehensive characterisation of the OSCC N-glycoproteome and we report on novel associations between altered N-glycosylation and key clinical features including lymph node metastasis and patient survival. The quantitative glycomics and glycoproteomics data provide a resource to further explore the underpinning disease mechanisms and uncover new disease markers for OSCC.

3.2. Methods

3.2.1. OSCC primary tumor tissues

The study was approved by the Ethics Review Board of the Cancer Institute of São Paulo (ICESP), Octavio Frias de Oliveira, ICESP, São Paulo, SP, Brazil, and

Plataforma Brasil (protocol CAAE 61402116.8.0000.0065). Informed consent was obtained from all patients included in the study. The methods and experimental protocols were performed in accordance with the approved guidelines and regulations. Surgically removed oral squamous cell carcinoma (OSCC) primary tumors provided as formalin-fixed paraffin-embedded (FFPE) tissues from a 31-patient cohort collected over a two-year period (2017-18) were investigated. In total, 19 patients presented lymph node metastasis (N+) and 12 patients presented without lymph node metastasis (N0) as determined by histopathology. See Supplementary Table S1 for key clinicopathological data.

3.2.2. Tissue sample preparation

Eight histological sections (each 10 µm thick) were prepared for each of the 31 OSCC FFPE tissues. Paraffin blocks were cut using a microtome and the tumor area was assigned by a trained pathologist for manual excision. A total tumor area of 7-8 cm2 were excised from 3-8 histological sections from each primary tumor tissue sample and transferred to a microtube for deparaffinization and protein extraction using a published protocol with modifications (21). For paraffin removal, 1 mL xylene was added to each sample, mixed for 1 min and centrifuged at 12,000 x g for 15 min at room temperature. The pellet was washed with 1 mL ethanol, mixed for 30 s, centrifuged as above and dried in a vacuum concentrator. Samples were resuspended in 300 µL lysis buffer containing 0.1 M N-(2-hydroxyethyl)piperazine-N′-(3-propanesulfonic acid) (EPPS), 0.1 M dithiothreitol (DTT), 1 x protease inhibitor cocktail (cOmplete ULTRA Tablets, Mini, EDTA-free, Roche; 1:10), pH 8.5. SDS was added to a final concentration of 4% (w/v) and the samples incubated for 60 min at 99°C with agitation (400 rpm). Samples were then centrifuged at 15,000 x g for 30 min at 4°C and submitted to ice-cold acetone precipitation with three-fold (v/v) excess acetone for 16 h at -20°C. Following centrifugation at 15,000 x g for 30 min at 4°C, protein pellets were resuspended in 30 µL buffer containing 8 M urea, 50 mM EPPS, protease inhibitor cocktail, 1 mM sodium fluoride, 1 mM sodium orthovanadate, 1 mM phenylmethylsulfonyl fluoride (PMSF), 1 mM EDTA, pH 8.5. Protein concentrations were determined by a bicinchoninic acid (BCA) protein assay kit (Thermo Fisher Scientific, Waltham, MA) and stored at -20°C until further handling.

3.2.3. N-glycan release

Glycans were prepared as described (22). Briefly, the protein extracts (15 µg total protein/sample) were blotted on a primed 0.45 µm polyvinylidene fluoride (PVDF) membrane (Merck-Millipore). In total, 31 patient samples were spotted and two for bovine fetuin (Sigma-Aldrich) experimental controls onto three separate PVDF membranes, including two membranes with each 14 spots and one with five spots. Each spot was approximately 1.5 cm2 and less than 0.5 cm in diameter with ~1 cm of spacing between spots (22). Protein spots were stained using Direct Blue (Sigma-Aldrich, Australia), excised and transferred to a flat bottom polypropylene 96-well plate (Corning Life Sciences, Australia), blocked with 1% (w/v) polyvinylpyrrolidone in 50% (v/v) methanol and washed with MilliQ water. The N-glycans were then released using 10 U recombinant Elizabethkingia miricola N-glycosidase F (Promega, V4831, 10 U/µL) for 16 h at 37°C. The detached N-glycans were reduced with 1 M sodium borohydride in 50 mM aqueous potassium hydroxide for 3 h at 50°C. The reaction was stopped using glacial acetic acid and the N-glycans desalted sequentially using custom-made strong cation exchange (AG 50W X8, Bio-Rad) and porous graphitised carbon (PGC, Thermo Fisher Scientific) solid phase extraction columns both packed on a C18 disc (Empore). The desalted N-glycan samples were stored at -20°C until analysis.

3.2.4. N-glycome profiling

N-glycomics data were acquired using an established PGC-LC-MS/MS method (22, 23), see Supplementary Table S2 for experimental overview and details of all generated 'omics datasets. The N-glycans were separated on an UltiMate 3000 HPLC system (Dionex, Sunnyvale, CA, USA) interfaced with a linear ion trap quadrupole (LTQ) Velos Pro (Thermo Scientific, San Jose, CA, USA). The glycan samples (3 µL injection volume, equivalent of N-glycans detached from ~15 µg total protein extract) were loaded on a PGC HPLC capillary column (Hypercarb KAPPA, 5 µm particle size, 200 Å pore size, 180 µm inner diameter x 100 mm length, Thermo Scientific) operated at 50°C with a constant flow rate (4 µL/min) supplemented with a

post-column make-up flow supplying pure acetonitrile (ACN) at 4 µL/min delivered by the HPLC system. Aqueous ammonium bicarbonate (10 mM), pH 8.0 (solvent A) and 10 mM ammonium bicarbonate in 70% ACN (solvent B) were used as mobile phases with the following 86 min-gradient: 8 min at 2.6% B, 2.6-13.5% B over 2 min, 13.5-37.3% B over 55 min, 37-64% B over 10 min, 64-98% B over 1 min, 5 min at 98% B, 98-2.6% B over 1 min and 4 min at 2.6% B. The ESI source was operated in negative ion polarity mode with source potential of 3.6 kV. Full MS1 scans were acquired in the range m/z 570-2,000 using 1 microscan, m/z 0.25 full width half maximum (FWHM) resolution, 5 × 104 automatic gain control (AGC) and 50 ms maximum accumulation time. MS/MS data were acquired using m/z 0.35 FWHM resolution; 2 × 104 AGC, 300 ms maximum accumulation time, and 2 m/z precursor ion isolation window. The five most abundant precursors in each MS1 full scan were selected for collision-induced dissociation (CID)-based MS/MS using a normalized collision energy (NCE) of 33% with an activation Q of 0.250 and 10 ms activation time. To prevent instrument bias, the injection order of the glycan samples from the OSCC patient cohort was randomised (24), using the psych package (25) under R (v3.6.0).

3.2.5. N-glycan data analysis

Xcalibur v2.2 (Thermo Scientific) was used to browse and interrogate the raw LC-MS/MS data. Putative glycan precursor ions were extracted using RawMeat v2.1 (Vast Scientific) (26, 27). Common contaminants not matching known N-glycan compositions and redundant precursors were manually removed. The monoisotopic precursors were searched against GlycoMod (28) with a mass tolerance of 0.5 Da to identify putative monosaccharide compositions. Only compositions containing Hex, HexNAc, dHex, NeuAc, and NeuGc were considered and N-glycans already reported in the UniCarbKB database were prioritized for masses matching multiple putative compositions. The N-glycan fine structures were then manually elucidated using monoisotopic mass, absolute and relative PGC-LC retention time and MS/MS fragmentation patterns as described (29, 30). EIC-based relative quantification of the identified N-glycans was performed using Skyline v.20.1.0.31 as described (27, 31). GlycoWorkBench v2.1 (32) was used to aid the manual annotation of the glycan fragment spectra and to generate glycan cartoons. Lowly expressed N-glycans with a

relative abundance below 0.1% were excluded from the relative quantitation due to poor spectral signal-to-noise ratios.

3.2.6. Protein digestion and peptide desalting

Protein samples (35 µg/sample) were reduced with 5 mM DTT (final concentration) for 25 min at 56°C and alkylated using 14 mM iodoacetamide (IAA) for 30 min at room temperature in the dark. Urea was added to a final concentration of 1.6 M in 50 mM ammonium bicarbonate and 1 mM calcium chloride prior to digestion with 1 µg sequencing grade porcine trypsin (Promega) for 16 h at 37°C (33). The reaction was quenched with a final concentration of 0.4% (v/v) trifluoroacetic acid (TFA). The resulting peptides were desalted using Oasis HLB 1 cc (30 mg) SPE cartridges (Waters) using an established protocol with modifications (33). Briefly, columns were activated with 1 mL methanol, washed twice with 1.5 mL 1% formic acid (FA) in 70% ACN (both v/v) and equilibrated with 1.5 mL 1% (v/v) aqueous FA prior to sample loading. After washing twice with 1.5 mL 1% (v/v) aqueous FA, peptides were eluted twice with 250 µL 1% FA in 70 % ACN (both v/v) and dried.

3.2.7. TMT labeling of peptides

A reference sample was prepared to allow for quantitative comparisons across multiple TMT experiments by pooling 3 µg digested protein from each sample. In total, 25 µg from this pool and 25 µg protein extract from each sample were used for TMT labeling (27). The peptide samples were randomized before being labeled using three separate TMT-11plex sets. The reference sample was consistently labeled with the 131 Da (131C) reporter ion channel in all three sets. In total, 30 peptide samples were labeled with TMT for quantitative proteomics including 12 N0 and 18 N+ samples (one random N+ sample from the sample cohort was excluded in TMT experimental design due to limited channels available), see Supplementary Table S2 for details. For the TMT labeling, 100 µL 100 mM triethylammonium bicarbonate buffer (final concentration) was added to each peptide sample that was labeled individually with 0.4 mg TMT-11plex mass tags (Thermo) in 41 µL anhydrous ACN over a 1 h incubation period at room temperature. The labeling reactions were

quenched using 8 µL 5% (v/v) hydroxylamine for 15 min at room temperature. After TMT labeling, all peptide samples were mixed 1:1:1 (w/w/w), desalted using HLB SPE cartridges (Waters) and dried. Aliquots of the TMT-labeled peptide mixtures from the three TMT-11plex experiments were analyzed without any pre-fractionation by LC-MS/MS, see Supplementary Table S2 for details.

3.2.8. Glycopeptide enrichment

TMT-labeled peptides (from 275 µg protein extract from 10 primary tumor tissue samples + one pool) was reconstituted in 50 µL loading/washing solvent containing 1% TFA in 80% ACN (both v/v). Five microliters (~18 µg protein digest) were allocated to the proteome analysis. The remaining peptides were loaded onto primed custom-made HILIC SPE micro-columns packed with ZIC-HILIC resin (10 µm particle size, 200 Å pore size, kindly provided by Merck kGaA, Darmstadt, Germany) onto supporting C8 disks (Empore) in p10 pipette tips (34). The flow-through fractions were collected. The HILIC-SPE micro-columns were then washed with 50 µL loading/washing solvent and the wash fraction combined with the flow-through fraction for separate downstream analysis, as this fraction contained the non-glycosylated peptides. The retained N-glycopeptides were eluted in three sequential steps, firstly with 50 µL 0.1% (v/v) aqueous TFA, followed by 50 µL 25 mM aqueous ammonium bicarbonate and then 50 µL 50% (v/v) ACN. The three N-glycopeptide fractions were combined to form the enriched glycopeptide mixture, which was dried and desalted on a primed Oligo R3 reversed phase SPE micro-column. R3 resin suspended in ACN was packed into a p10 pipette tips, washed three times with 50 µL ACN, and then washed with 0.1% (v/v) aqueous TFA before sample loading. The sample loading was repeated once to ensure high recovery ahead of three washing steps with 0.1% (v/v) aqueous TFA and elution with 50 µL 0.1% TFA in 50% ACN (both v/v) and then with 50 µL 0.1% TFA in 70% ACN (both v/v). The desalted glycopeptide fractions were combined, dried and stored at -20°C (27).

3.2.9. High-pH reversed phase prefractionation

The HILIC enriched N-glycopeptides and HILIC flow-through fractions containing the non-glycosylated peptides were resuspended separately in 50 µL 25 mM aqueous ammonium bicarbonate for high pH pre-fractionation using Oligo R2 reversed phase SPE micro-columns packed on supporting C18 discs (Empore) in standard p10 pipette tips. The SPE micro-columns were primed three times with 50 µL ACN and then washed with 50 µL 25 mM aqueous ammonium bicarbonate. Samples were loaded on the columns followed by two washing steps with 50 µL 25 mM aqueous ammonium bicarbonate. Peptides were eluted in three fractions i.e. fraction 1: 25 mM ammonium bicarbonate in 10% (v/v) ACN; fraction 2: 25 mM ammonium bicarbonate in 20% (v/v) ACN, and fraction 3: 25 mM ammonium bicarbonate in 60% (v/v) ACN. Each eluted fraction was dried and resuspended in 0.1% (v/v) aqueous FA for separate LC-MS/MS analysis.

3.2.10. (Glyco)proteome profiling by LC-MS/MS

Approximately 130 ng (glyco)peptide material was injected per LC-MS/MS run (estimated). The (glyco)peptides were loaded on a trap column (2 cm length x 100 µm inner diameter) custom packed with ReproSil-Pur C18 AQ 5 µm resin (Dr. Maisch, Ammerbuch-Entringen, Germany) and separated at a constant flow rate of 250 nL/min on an analytical column (Reprosil-Pur C18-Aq, 25 cm length x 75 µm inner diameter, 3 µm particle size, Dr. Maisch, Ammerbuch-Entringen, Germany) using an UltiMate™ 3000 RSLCnano System. The mobile phases were 0.1% FA in 99.9% (both v/v) ACN (solvent B) and 0.1% (v/v) aqueous FA (solvent A). The gradient was 2-30% B over 100 min, 30-50% B over 18 min, 50-95% B over 1 min and 9 min at 95% B. The nanoLC was connected to a Q-Exactive HF-X Hybrid Quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific) operating in positive ion polarity mode. The Orbitrap acquired full MS1 scans with an AGC of 3 x 106 ions and 50 ms maximum accumulation time. Full MS1 scans were acquired at high-resolution (60,000 FWHM at m/z 200) in the m/z 350-1,800 range. The 20 most abundant precursor ions were selected from each MS1 full scan using data-dependent acquisition and were fragmented utilizing higher energy collision-induced dissociation (HCD) with a NCE of 35%. Only multicharged

precursors (z ≥ 2) were selected for fragmentation. Fragment spectra were acquired at 45,000 resolution (at m/z 200) with an AGC of 1 x 105 and 90 ms maximum accumulation time using a precursor isolation window of m/z 1.0 and a dynamic exclusion of 30 s after a single isolation and fragmentation of a given precursor ion.

3.2.11. N-glycoproteomics data analysis

The HCD-MS/MS data of intact N-glycopeptides were searched with Byonic v2.6.46 (Protein Metrics Inc, CA, USA) (35) using 10/20 ppm as the precursor/product ion mass tolerance, respectively. Cys carbamidomethylation (+57.021 Da) and TMT (+229.163 Da) at N-term and lysine (K) were considered fixed modifications. Trypsin specific cleavages were considered with a maximum of two missed cleavages allowed per peptide. The following variable modifications were considered: Met oxidation (+15.994 Da), and N-glycosylation of sequon-localized Asn with a glycomics-informed N-glycan database comprising the N-glycan compositions identified in the OSCC tissue N-glycome data (Supplementary Table S3). HexNAc1, HexNAc1Fuc1 and HexNAc2 were manually added to the database since these are typically not effectively observed with our glycomics method. A maximum of two common modifications and a maximum of one rare modification were allowed. The HCD-MS/MS data were searched against a protein database composed of all reviewed UniProtKB human proteins (20,300 sequences, released December 11, 2019). All searches were filtered to <1% false discovery rate (FDR) at the protein level and 0% at the peptide level by using a protein decoy database (36). Only N-glycopeptides confidently identified with PEP 2D scores < 0.001 were considered (27). Glycopeptides identified with lower confidence (PEP2D ≥ 0.001), and those found in the reverse database and contaminant database were excluded. The identified N-glycopeptides were quantified using the ‘Report Ion Quantifier’ available as a node in Proteome Discoverer v2.2 (Thermo Scientific) and the reporter ion intensities from the MS/MS scans were extracted from the QuantSpectra table (27). Glycopeptides were manually grouped by summing the reporter ion intensities from the glycopeptide-to-spectrum matches (glycoPSMs) belonging to the same UniProtKB identifier, same glycosylation site within the protein, and same glycan composition. The abundances of the unique glycopeptides from each channel were first normalized by dividing each unique glycopeptide reporter ion intensity by the

reference reporter intensity (from the peptide pool, see above) within that specific experiment and further normalized by the total sum intensity of each channel to correct for any inter-sample variation in the total yield during the labeling reactions.

3.2.12. Proteomics data analysis

LC-MS/MS-based proteomics data were acquired of both unenriched peptide samples not subjected to glycopeptide enrichment as well as to the HILIC flow-through fractions. All data were processed using MaxQuant v1.6.12.0 (37). The HCD-MS/MS data were searched against the reviewed UniProtKB Human Protein Database (20,300 sequences, released December 11, 2019) using the Andromeda search engine (38) with a tolerance of 4.5 ppm for precursor ions and 20 ppm for product ions. Report fragment ions “10plex TMT” were enabled in the quantification settings and the enzyme specificity was set to trypsin with a maximum of two missed cleavages permitted. Carbamidomethylation of Cys (+57.021 Da) was considered a fixed modification, and oxidation of Met (+15.994 Da) and protein N-terminal acetylation (+42.010 Da) were considered variable modifications. Both the protein and peptide identifications were filtered to 1% FDR. Processing and statistical analyses (Student’s t-test) of the resulting data table (MaxQuant output) were performed in Perseus v1.6.14.0 (37). Proteins identified using the reverse database, proteins only identified through modified peptides and proteins identified from the MaxQuant contaminant database were excluded (except human keratins, which were not excluded since they are of interest in the study of squamous tissues (39)). The identified proteins were quantified by their reporter ion intensities using at least one razor/unique peptide per protein. Protein abundances of each channel were first normalized by dividing the reporter ion intensity by the reference protein reporter ion intensity (131C channel of peptide pool) within that specific LC-MS/MS run and further normalized by dividing by the sum of total intensity of each channel to correct for any inter-sample variation in the total yield during the labeling reactions.

3.2.13. Experimental design and statistical rationale

In total, 31 OSCC patients presenting with (n = 19, N+) or without (n = 12, N0) lymph node metastasis were investigated. No technical replicates were acquired due to the considerable number of biological replicates. For the glycome profiling and glycosylation site analyses, statistical significance was assessed using unpaired two-tailed Student's t-tests in which $p < 0.05$ was used as the confidence threshold. For the glycoproteome profiling, the statistical significance of the summed intensities of groups of glycopeptides, which were grouped based on their N-glycan classes or tissue cluster, was assessed using unpaired two-tailed Student's t-tests in which $p < 0.05$ was used as the confidence threshold. For data visualization, heat maps with z-score values of normalized intensities were generated using the open-source statistical programming language R.

3.2.14. Clinical association and survival analysis

Linear regression analysis was performed using the R code to evaluate the linear relationship between glycan or glycopeptide abundance and the following clinicopathological variables: age (>63 or ≤63, mean age of patient cohorts chosen as the threshold), sex, smoking habit, alcohol consumption, tumor size, lymph node metastasis (N+, N0), clinical stage, type of treatment (surgery, surgery and radiotherapy, or a combination of surgery, radiation and chemotherapy), disease-free survival, presence of the worst pattern of invasion (40), presence of inflammatory infiltrate, and perineural invasion. Linear regressions with $p < 0.05$ were considered significant. The Pearson product−moment correlation coefficient (R) was also calculated to measure the strength of the association between variables. Only associations with $R < -0.5$ or $0.5 < R$ with at least six valid values per group and at least three valid values per clinical feature were considered (41). For cluster analysis and associations to clinical features, Fisher's Exact Test (for two group comparisons) or Pearson Chi-Square test (for comparisons of more than two groups) were performed using IBM SPSS Statistics v 28.0.0.0 (190). Furthermore, a survival analysis was calculated on GraphPad Prism v.9.1.2 using the Kaplan-Meier methodology and compared with the log-rank test, evaluating the survival probability comparing a higher and lower abundance of specific glycoproteins, glycopeptides and glycans.

3.2.15. ROC curve analysis by logistic regression and random forest models

The potential of select N-glycans and N-glycopeptides to stratify N0 and N+ patients was evaluated by the area-under-the-curve (AUC) of constructed receiver operating characteristic (ROC) curves generated using random forest and logistic regression models. An AUC-ROC with a 95% confidence interval was used for comparison. Optimal cut-off by highest sensitivity (true positive rate) as a function of the specificity (false positive rate) was calculated, and 60% was chosen as the decision threshold. For all statistical comparisons, an ANOVA $p < 0.05$ was considered as significance threshold. The data analysis was performed using the package pROC (42) and the R environment version 3.6.0.

3.2.16. Enrichment analysis of gene ontology (GO) terms

GO enrichment analysis was performed using the DAVID (43) or Enrichr (44) with corrected $p < 0.05$ chosen as the significance threshold. The entire human proteome GO annotation file was used as a reference set.

3.3. Results

3.3.1. Comprehensive N-glycoprofiling of primary tumors from OSCC patients

This study applied an integrative MS-driven multi-omics approach comprising both quantitative glycomics and glycoproteomics to comprehensively profile the protein N-glycosylation in FFPE primary tumor tissues surgically removed from 31 patients operated with curative intent due to OSCC, including patients with (N+, n = 19) and without (N0, n = 12) lymph node metastasis, Figure 1 and Supplementary Figure S1. See Supplementary Table S1-S2 for patient metadata and overview of the generated LC-MS/MS datasets, respectively.

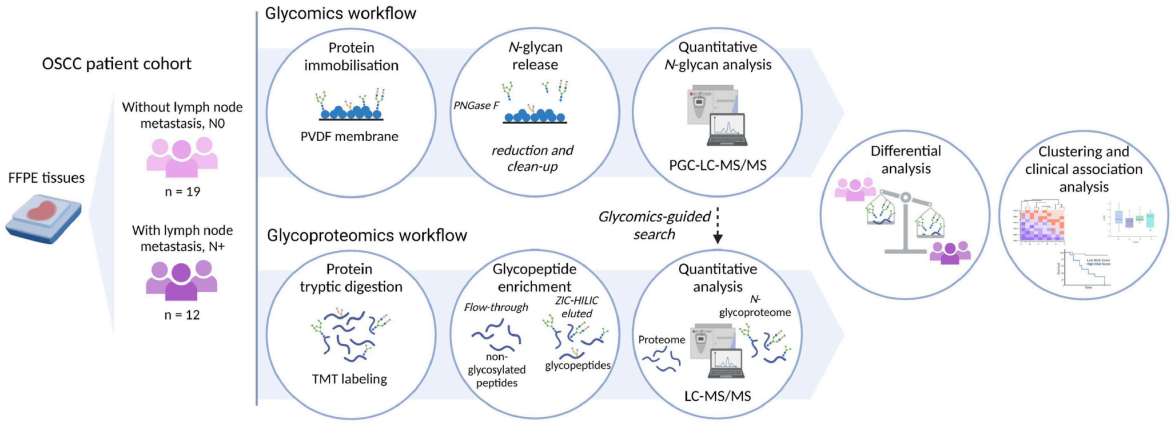

Fig. 1 Study overview. *Protein N-glycosylation was investigated using quantitative glycomics (top) and glycoproteomics (bottom) from resected primary tumor tissues from 31 OSCC patients with (N+, n = 19) or without (N0, n = 12) lymph node metastasis. The resulting N-glycome and N-glycoproteome data were quantitatively compared between patient groups using advanced statistical tests including clustering analysis, and associations to a range of clinicopathological features were explored using patient metadata. OSCC, oral squamous cell carcinoma. Created with Biorender.com.*

3.3.2. Comparative N-glycome profiling of OSCC tissues from N+ and N0 patients

State-of-the-art quantitative N-glycomics of proteins extracted from the OSCC primary tumor tissues was used to profile, with glycan fine structural resolution, a total of 83 N-glycan structures spanning 52 N-glycan compositions (denoted Glycan 1-52 with isomers denoted a, b, c…). See Supplementary Figure S2 for a map of the identified N-glycan structures and Supplementary Table S3 and Supplementary File S1 for tabulated glycomics data and annotated spectral evidence, respectively.

Mainly complex (38.7-61.5%), oligomannosidic (18.3-36.5%) and paucimannosidic (9.8-31.9%) N-glycans were identified across the OSCC tumor tissues, Figure 2A and Supplementary Figure S3A. At the global N-glycome level, prominent differences were neither observed in the N-glycan type distribution nor in the common N-glycan structural features such as the degree of branching and levels of fucosylation and sialylation (all $p \geq 0.05$), Supplementary Figure S3B. Relatively uniform global N-glycosylation across the N0 and N+ tissues was further demonstrated using pair-wise N-glycome correlation analysis, PCA and hierarchical clustering analysis, Supplementary Figure S4 and Supplementary Table S4.

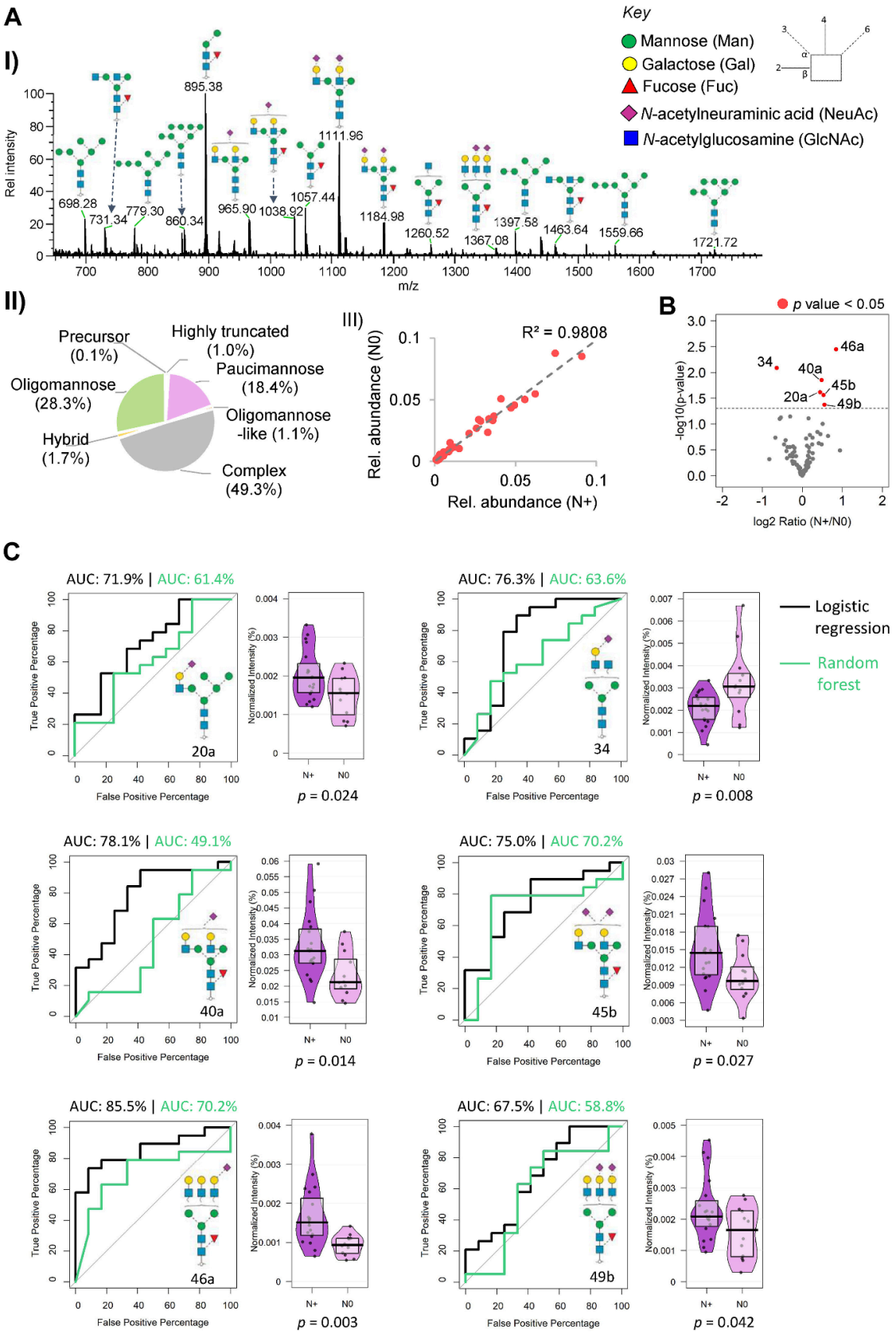

Fig. 2 Tumor tissue N-glycans show a potential for OSCC patient stratification. *A, I- The OSCC tumor tissue N-glycome predominantly comprises complex-, oligomannosidic- and paucimannosidic-type N-glycans as shown by an exemplar summed MS1 spectrum. II- Overall*

distribution of N-glycan classes across all investigated OSCC tumor tissues. III- Similar N-glycan distribution within N+ and N0 tissues as demonstrated by a high correlation coefficient (R2). B, volcano plot of the 83 N-glycan structures identified in the OSCC tissues. The six differentially abundant N-glycan structures are presented as red dots (N+ versus N0; p < 0.05; two-sided unpaired Student's t test). C, five N-glycans (Glycan 20a, 40a, 45b, 46a, and 49b) with increased expression in N+ tissues and one N-glycan (Glycan 34) with decreased level in N+ tissues (relative to N0, see pirate plots in purple) showed a potential to stratify the N0 and N+ patient groups with an AUC-ROC >60% using the logistic regression model (black trace). Four of those N-glycans were also able to stratify the N0 and N+ patient groups by using a random forest model (AUC-ROC > 60%, green trace). AUC-ROC: area-under-the-curve of the receiver operating characteristic. p < 0.05 was used as a threshold to denote statistical significance in the applied Student's t-tests. AUC, area-under-the-curve; OSCC, oral squamous cell carcinoma; ROC, receiver operating characteristic.

Despite notable N-glycome similarities between the two investigated patient groups, differential expression was observed for six individual N-glycan structures ($p < 0.05$) including one down-regulated N-glycan (Glycan 34) and five up-regulated N-glycans (Glycan 20a, Glycan 40a, Glycan 45b, Glycan 46a, Glycan 49b) in N+ compared to N0 tissues, Figure 2B and Supplementary Table S3. Interestingly, each of these differentially expressed N-glycans was able to accurately stratify the N0 and N+ patients with confidence using a logistic regression model (AUC values ranging from 67.5-85.5%), and, in part, also by a random forest model (AUC values from 49.1% to 70.2%), Figure 2C.

3.3.3. Comparative N-glycoproteome profiling of OSCC primary tumor tissues from N+ and N0 patients

Guided by the acquired N-glycomics data, we then explored the glycoproteome complexity of the OSCC tumor tissues using our recently developed glycomics-informed glycoproteomics method (27, 45). A total of 3,117 unique N-glycopeptides (i.e. glycopeptides carrying a discrete glycan composition at a discrete site) displaying 55 different N-glycan compositions from 419 different source N-glycoproteins spanning a wide dynamic range were profiled across the OSSC tissue cohort, Figure 3A and Supplementary Table S5-S6.

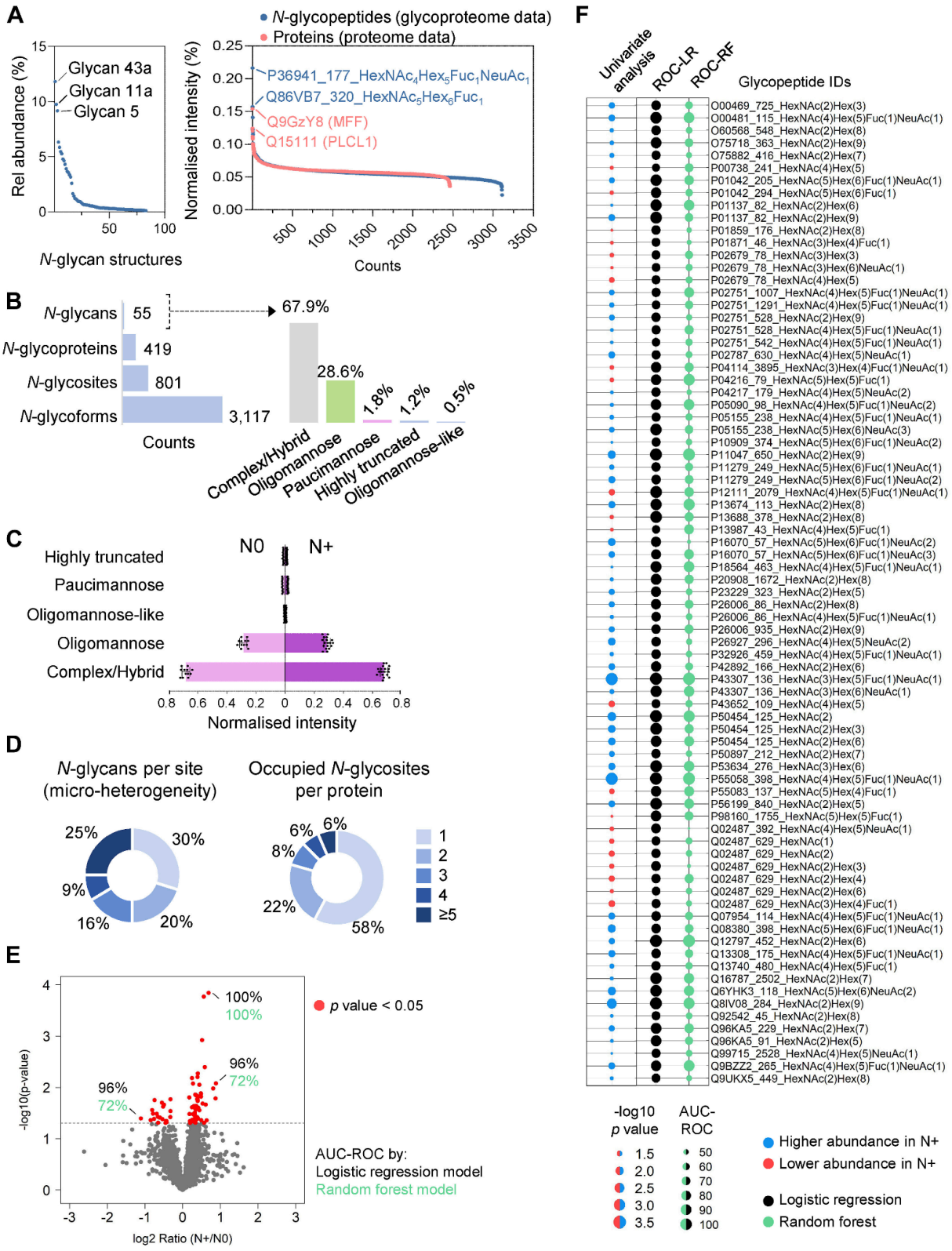

Fig. 3 Overview of the N-glycoproteome of OSCC primary tumor tissues. *A, abundance range of identified N-glycan compositions (left) and identified N-glycopeptides (both from the acquired N-glycoproteome data, blue dots) and proteins (from the proteome data, pink dots). The most abundant N-glycans, N-glycopeptides, and proteins are labeled in each graph. B, overview of the N-glycoproteome of the investigated OSCC tumor tissues including the N-glycan compositions (and their distribution across the N-glycan classes, right), source N-glycoproteins, N-glycosites, and unique*

N-glycoforms (unique protein + unique site + unique glycan) identified in the ZIC-HILIC–enriched fractions. C, comparison between the distribution of N-glycan classes in the N0 and N+ tumor tissues based on the N-glycoproteome data. D, site-specific N-glycan micro-heterogeneity (top) and occupied sites per protein (bottom) in the OSCC tumor tissues. E, volcano plot of the 3117 N-glycopeptides identified in the OSCC tissues. The 79 differentially abundant N-glycopeptides are presented as red dots (N+ versus N0; $p < 0.05$; two-sided unpaired Student's t test), including 54 with higher abundance in N+ and 22 with lower abundance in N+ (relative to N0). AUC-ROC values (in percentages) are indicated for the N-glycopeptides with the highest fold-change (N+/N0) or p values, by logistic regression (black) or by random forest model (green) (see supplemental Table S7). F, all 79 quantitatively altered N-glycopeptides were able to stratify the N0 and N+ patient groups by using a logistic regression (black dots) or by random forest model (green dots) (AUC-ROC > 60%). AUC-ROC: area-under-the-curve of the receiver operating characteristic. AUC, area-under-the-curve; C/H, complex and hybrid (grouped since these glycan types cannot reliably be distinguished through glycoproteomics); HT, highly truncated; O, oligomannose; O-L, oligomannose-like; OSCC, oral squamous cell carcinoma; P, paucimannose; ROC, receiver operating characteristic; ZIC-HILIC, zwitterionic hydrophilic interaction liquid chromatography.

In line with the N-glycome data, most N-glycopeptides carried complex/hybrid (67.9%) or oligomannosidic (28.6%) N-glycans, Figure 3B. The under-representation of paucimannosidic N-glycopeptides in the glycoproteomics data (1.8%) relative to the considerable levels found in the N-glycome (18.4%) can likely be attributed to an inefficient HILIC-SPE enrichment of these less hydrophilic N-glycopeptides (unpublished observation). Recapitulating findings from the N-glycome data, the glycoproteomics data did not reveal any consistent differences in the N-glycan type distribution, Figure 3C, and the global N-glycoproteome, Supplementary Figure S5, between the N0 and N+ patients. Providing further insights into the OSCC tumor tissue N-glycoproteome complexity, the glycoproteomics data also revealed that most N-glycosites (70%) carried more than one discrete N-glycan composition (25% of sites were decorated with ≥5 glycan compositions), while most N-glycoproteins (58%) were identified with only one occupied N-glycosylation site, Figure 3D. In our previous study of fresh prostate tissues samples (27), we used the same glycoprofiling approach to characterize protein N-glycosylation and identified a similar microheterogeneity, with ~30% of the glycosites carrying at least five different glycan compositions. Extensive glycosylation site microheterogeneity is a well-known phenomenon exhibited by most glycoproteins studied to date with more than 5-20 structures often observed at each site (46–49). Therefore, the microheterogeneity reported in this work, albeit extensive, may be an underestimation. Given the complexity of the OSCC tumor samples, it is likely that we only covered the most abundant structures carried by each glycosylation site.

Despite the relatively uniform N-glycoproteome across the N0 and N+ patients, a total of 79 N-glycopeptides from 56 source N-glycoproteins were found to be quantitatively altered between the two patient groups ($p < 0.05$) including 57 N-glycopeptides from 43 source N-glycoproteins that were elevated and 22 N-glycopeptides from 14 source N-glycoproteins of lower abundance in N+ relative to N0 tissues, Figure 3E and Supplementary Table S5. Excitingly, all 79 N-glycopeptides demonstrated a notable potential to stratify N0 and N+ OSCC patients (AUC-ROC >60% by logistic regression model), Figure 3F and Supplementary Table S7. In support, a high proportion of these N-glycopeptides (50 N-glycopeptides, 63.3%) was also found to stratify the patient groups with an AUC-ROC >60% using a random forest model.

3.3.4. N-glycosylation-guided clustering of OSCC patients

To further investigate the relatively subtle yet consistent differences in N-glycosylation found within the investigated OSCC primary tumor tissue cohort, the obtained N-glycome and N-glycoproteome data were then interrogated using advanced data analyses and visualization methods.

Unsupervised hierarchical clustering of the N-glycome data revealed two major tumor clusters (T-C1 and T-C2) and two major N-glycan clusters (NG-C1 and NG-C2), Figure 4A and Supplementary Figure S6. Within the NG-C1 cluster, altered distribution of all N-glycan classes were found between T-C1 and T-C2, while in NG-C2 differences in the distribution of paucimannosidic and complex N-glycans, but not oligomannosidic N-glycans, were observed between the two tumor clusters ($p < 0.05$) (note that the other glycan classes were not observed in NG-C2), Figure 4A.

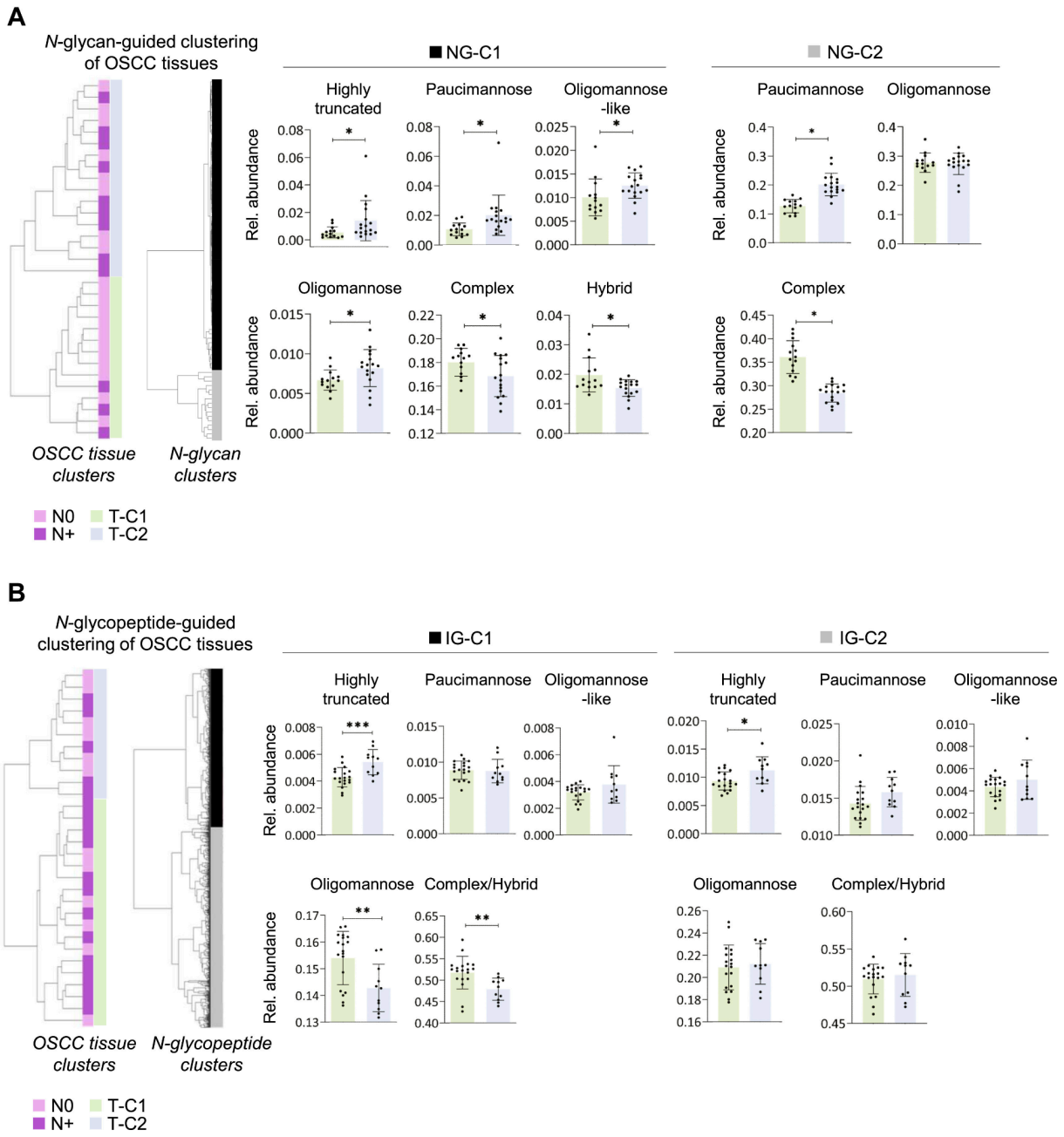

Fig. 4 N-glycome– and N-glycoproteome–guided clustering of OSCC primary tumor tissue. *A, unsupervised hierarchical clustering of the OSCC tumor tissue informed by the N-glycome data from N0 and N+ patients, performed with the 'cluster map' function in the Seaborne package under Python using Euclidean distance and Ward linkage. Relative abundance of the N-glycan types between the two major tumor clusters, T-C1 (green bars) and T-C2 (blue bars), in the two major N-glycan clusters, NG-C1 (black) and NG-C2 (gray). B, unsupervised hierarchical clustering of the OSCC tumor tissue guided by the N-glycoproteome data from N0 and N+ patients, performed with the 'cluster map' function in the Seaborne package under Python using Canberra distance and Ward linkage. Relative abundance of the N-glycan types between the two major tumor clusters, T-C1 (green bars) and T-C2 (blue bars), in the two major intact N-glycopeptide clusters, IG-C1 (black) and IG-C2 (gray). Student's t test, two-sided, unpaired analysis; α = 0.05, ∗p < 0.05, ∗∗p < 0.01, ∗∗∗p < 0.001. OSCC, oral squamous cell carcinoma.*

Similarly, unsupervised hierarchical clustering of the N-glycoproteome data revealed two main tumor clusters (T-C1 and T-C2) and two N-glycopeptide clusters (IG-C1 and IG-C2), Figure 4B and Supplementary Figure S7. Differences in the expression of highly truncated, oligomannosidic and complex/hybrid N-glycans were identified between T-C1 and T-C2 in IG-C1, while only the highly truncated N-glycans exhibited differences between the two tumor clusters in IG-C2, Figure 4B.

To explore potential biological roles of the source N-glycoproteins identified within each N-glycopeptide cluster, pathway enrichment analyses were performed. Extracellular matrix organization was the most enriched biological process in both N-glycopeptide clusters (FDR adjusted p = 9.4 x 10-36 in IG-C1 and p = 3.6 x 10-45 in IG-C2), followed by platelet degranulation in IG-C1 (FDR adjusted p = 1.1 x 10-20) and neutrophil degranulation in IG-C2 (FDR adjusted p = 2.0 x 10-21) when comparing T-C1 and T-C2, Supplementary Figure S8A. We then tested for differential N-glycan class distribution within these highly enriched pathways, which demonstrated altered expression of highly truncated N-glycans decorating proteins involved with platelet degranulation within IG-C1, and showed altered pauci- and oligomannosidic N-glycosylation on proteins involved in neutrophil degranulation within IG-C2, Supplementary Figure S8B.

3.3.5. Associations between N-glycoproteome components, clinical features and biological processes

The six N-glycans and 79 N-glycopeptides found to be differentially expressed in N0 and N+ patients were then evaluated for associations with a range of clinical features.

Unsupervised hierarchical clustering of the expression data of the six N-glycan structures revealed two new distinct tumor groups that displayed differences in vascular invasion (p = 0.049, Fisher's Exact Test) and lymph node (N) status (p = 0.039, Pearson Chi-Square test), Figure 5A-B. In contrast, the two tumor groups formed by the 79 N-glycopeptides differed in terms of N status (p = 0.0078), tumor size (p = 0.0169, both Fisher's Exact Test) and type of treatment received by the OSCC patients (p = 0.0354, Pearson Chi-Square), Figure 5C-D.

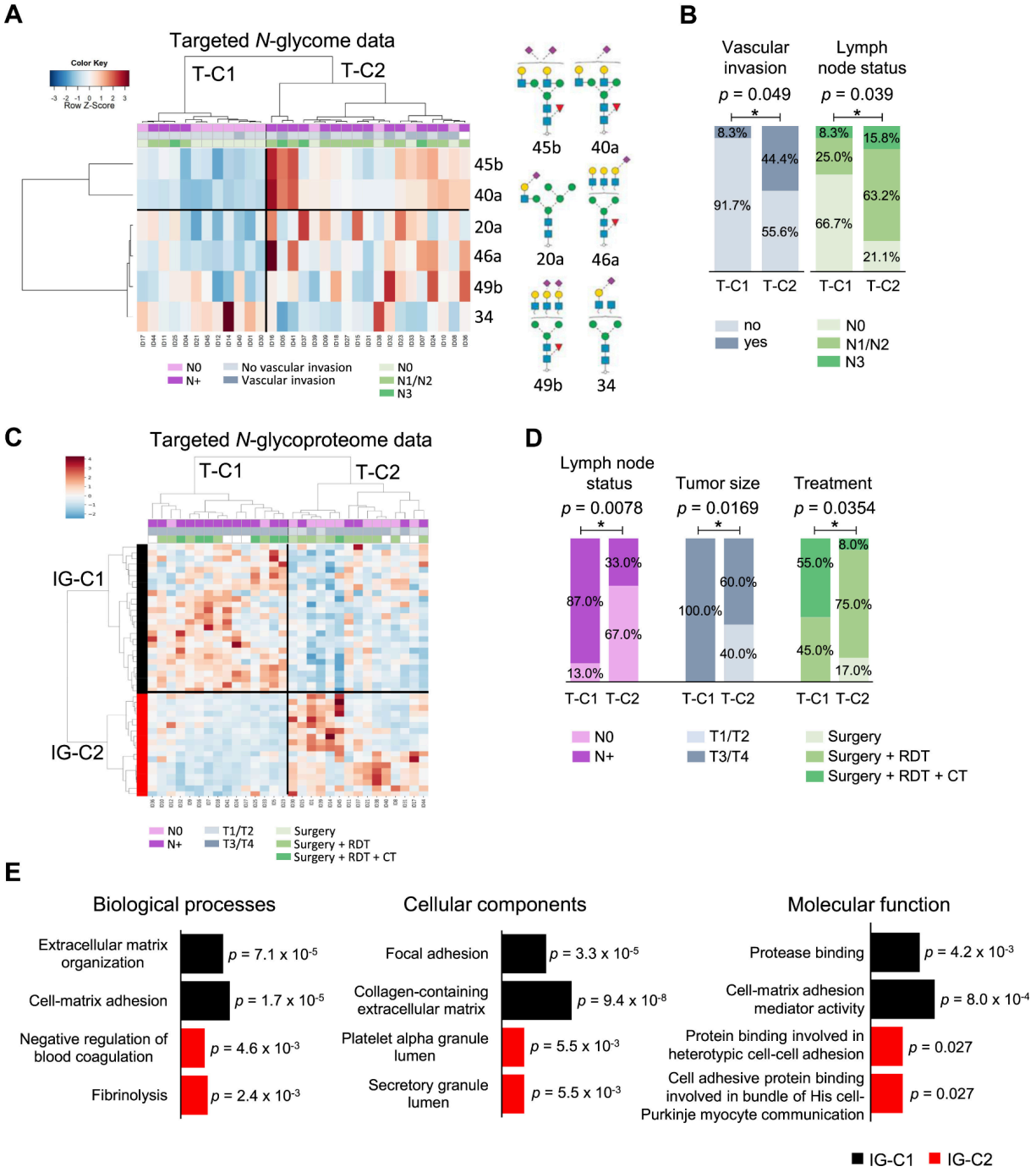

Fig. 5 Targeted N-glycan– and N-glycopeptide–guided OSCC tumor clusters associate with clinical features. *A, unsupervised hierarchical clustering analysis of six N-glycans found to be differentially abundant in OSCC tumor tissues from N0 and N+ patients. Plots are made with the 'heatmap.3' function under the R environment using Euclidean distance and Ward linkage. B, cluster association analysis with clinical features ($p < 0.05$, Fisher's Exact Test for two group comparisons; or Pearson Chi-Square test for more than two group comparisons) performed using IBM SPSS Statistics. C, unsupervised hierarchical clustering analysis of 79 N-glycopeptides found to be differentially expressed between the N0 and N+ tumor tissues. Plots were performed with the 'cluster map' function in the Seaborne package under Python using Ward and correlation for clustering. D, cluster association analysis with clinical features. E, the two most enriched gene ontology (GO) biological processes, cellular components, and molecular function in the IG-C1 (black) and IG-C2 (red) glycopeptide clusters. An adjusted p value is indicated for each term in the enrichment analysis performed using the Enrichr tool. Fisher's Exact Test or Pearson Chi-Square test, $*p < 0.05$. CT, chemotherapy; OSCC, oral squamous cell carcinoma; RDT, radiotherapy.*

To investigate whether N-glycosylation was associated with specific biological processes, cellular compartments and molecular functions, we searched for enriched gene ontology (GO) terms within the IG-C1 and IG-C2 N-glycopeptide clusters. This analysis revealed significant enrichment of extracellular matrix organization and cell-matrix adhesion in IG-C1, while ‘negative regulation of blood coagulation’ and ‘fibrinolysis’ were enriched in IG-C2, Figure 5E. Proteins from IG-C1 were enriched in ‘focal adhesion’ and ‘collagen-containing extracellular matrix’ compartments, and were involved in ‘protease binding’ and ‘cell-matrix adhesion mediator activity’, whereas proteins from IG-C2 showed trends of being of platelet and secretory granule origins, and involved in protein binding.

3.3.6. Subsets of OSCC N-glycans and N-glycopeptides associate with key clinical outcomes

In total 25 of the 79 N-glycopeptides that displayed differential abundance between N0 and N+ patients were found to correlate with distinct clinicopathological features, Supplementary Figure S9. Amongst them were five N-glycopeptides all carrying Glycan 40a (but arising from five different source glycoproteins), which associated with lymph node status and surgical margin involvement (50), Figure 6A. Surprisingly, three of the five source N-glycoproteins (BTN3A1, SIGLEC1 and PLTP) were not identified in the global proteomics data, Supplementary Table S6. Moreover, while the N-glycopeptide from COL6A3 was less abundant in tumor tissues from N+ patients, the source glycoprotein was found in higher levels in these individuals. Conversely, ITGB6, whose glycopeptide was more abundant in N+ patients, displayed reduced protein abundance in this condition, observations which collectively indicate that both protein- and glycan-level regulation occur in the OSCC tumor tissues, Figure 6B.

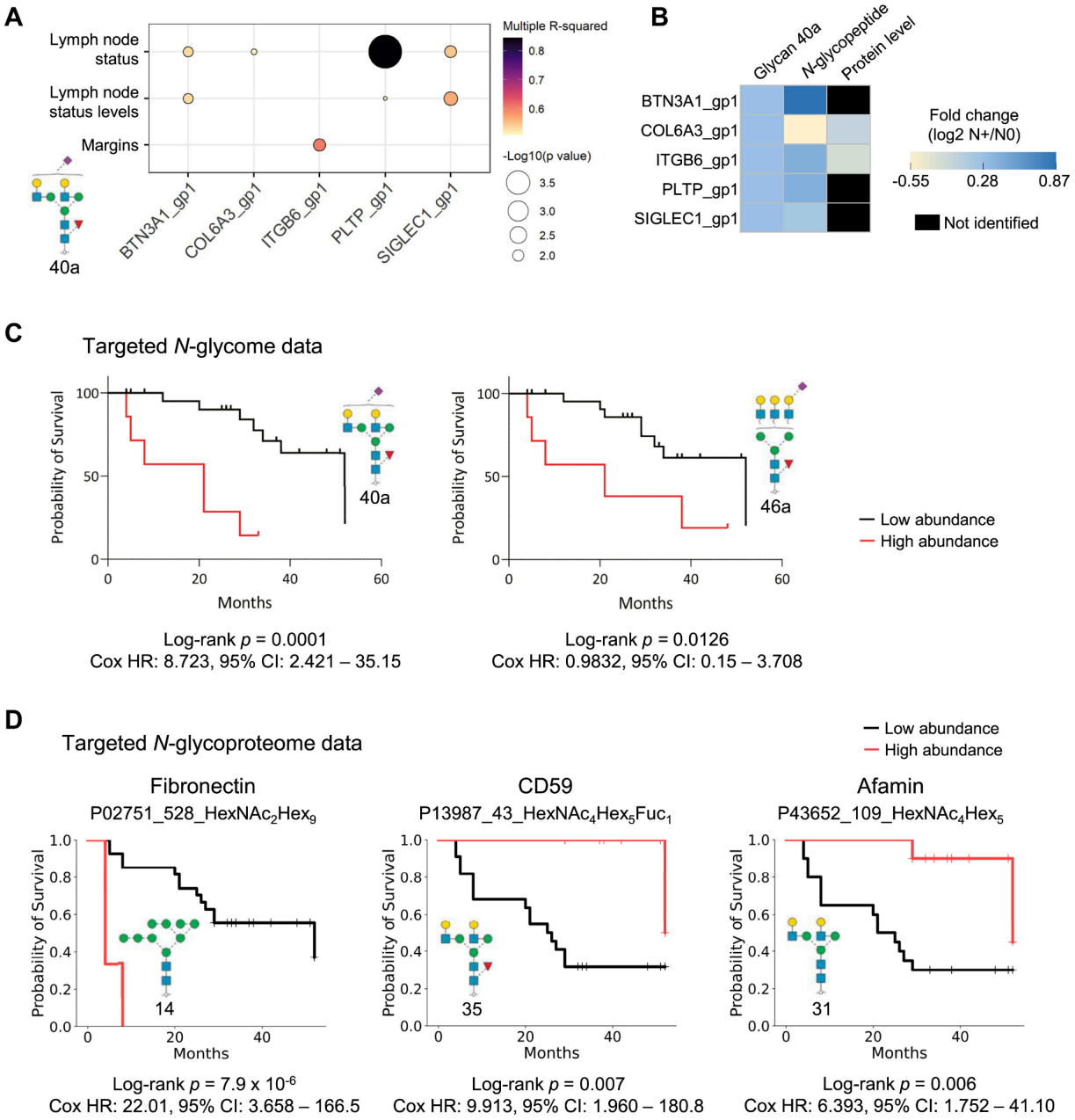

Fig. 6 The expression level of specific N-glycans and N-glycopeptides associates with clinical outcomes in OSCC. *A, five out of twenty five N-glycopeptides that associated with clinicopathological features were found to carry Glycan 40a. B, distribution of fold change values (log2 N+/N0) of Glycan 40a, glycopeptides carrying Glycan 40a, and their source glycoproteins in the OSCC patient cohort. C, glycan-focused Kaplan–Meier survival analysis. Patients with relative high levels of Glycan 40a and Glycan 46b presented worse overall survival ($p < 0.05$, log-rank test). D, glycopeptide-focused Kaplan–Meier survival analysis. Left: Relative high abundance of a HexNAc2Hex9 glycopeptide from fibronectin was associated with relatively low patient survival (elevated glycopeptide levels found in N+ samples). Middle-right: Relative low abundance of N-glycopeptides carrying a related neutral glycan from CD59 and afamin were associated with a relatively low patient survival (reduced expression levels in N+ samples). Lymph node status levels: N0, N1, N2, N3. OSCC, oral squamous cell carcinoma.*

Notably, Kaplan-Meier plots showed that relatively high levels of Glycan 40a and Glycan 46a were associated with a relatively poor patient survival, Figure 6C. The potential prognostic value of Glycan 40a and 46a was supported by their ability to stratify N0 and N+ patients as demonstrated by high ROC-AUC values (AUC 78.1 for Glycan 40a and AUC 85.5 for Glycan 46a, both by logistic regression model), Figure 2C.

Similarly, three N-glycopeptides that displayed abundance differences between N0 and N+ showed an association with patient survival, Figure 6D. Specifically, higher abundance of a fibronectin peptide (P02751) carrying an M9 oligomannosidic N-glycan (Glycan 14) was found to associate with lower patient survival, while reduced levels of a CD59 peptide (P13987) carrying a fucosylated complex N-glycan (Glycan 35) and an afamin peptide (P43652) carrying a related non-fucosylated N-glycan variant (Glycan 31) were associated with lower patient survival. These three N-glycopeptides also demonstrated an ability to stratify N0 and N+ patients as demonstrated by ROC analysis, Supplementary Figure S10.

Taken together, unbiased and global as well as targeted interrogation of our quantitative N-glycome and N-glycoproteome data pointed to previously unknown associations between OSCC tumor N-glycosylation and key clinical outcomes including, most importantly, lymph node metastasis and patient survival.

3.4. Discussion

Previous studies have investigated the glycosylation patterns in oral cancer within diverse biological specimens including in cell lines (11, 19), tumor tissues (51), saliva (12, 20, 52), blood (12), plasma (53) and serum (54, 55) in attempts to uncover changes related to cell transformation (11) and with disease progression (12). Most of those studies evaluated the total levels of specific glycans or glycan features (12, 20, 51, 53) using relatively simple analytical techniques such as lectin-based strategies (51, 53, 54) and MALDI-MS profiling (11, 19, 54, 55), which are not able to survey the glycoproteome in a comprehensive, unbiased and quantitative manner with glycan fine structure and site-specific information (56). Our study is the first to use integrated glycomics and glycoproteomics to provide detailed and quantitative insights into the heterogenous N-glycoproteome of OSCC primary tumor tissues obtained from a

valuable cohort of patients with and without lymph node metastasis, a principal prognostic factor in OSCC (3, 57, 58).

To examine the N-glycoproteome of OSCC tumor tissues, we firstly established the N-glycome using quantitative N-glycomics, which served to inform a customized N-glycan database to aid the challenging downstream glycoproteomics data analysis using our glycomics-informed glycoproteomics approach (27, 45). Intuitively, tailoring the glycan database reduces the search space resulting in lower glycopeptide FDRs and reduced search times (59) and ultimately more accurate quantitative information of the protein carriers, site occupancy (macroheterogeneity), and site distribution (microheterogeneity) of N-glycosylation (60).

We found surprisingly few global differences in the N-glycome when comparing between the N+ and N0 patient groups including in the N-glycan class distribution and N-glycosylation features (core fucosylation, sialylation and branching) suggesting a relative stable glycosylation machinery during lymph node metastasis. Related studies have reported more dramatic glycan differences including altered sialylation, fucosylation and branching when comparing OSCC patients with healthy donors (51, 55) suggesting that disease onset rather than disease progression (involving lymph node metastasis) has a stronger influence on glycan remodeling.

Comparative glycomics analysis indicated that six sialic acid-capped N-glycans (displayed in Figure 2C) exhibit an altered abundance between N+ and N0 tumor tissues (N+ vs N0; $p < 0.05$; two-sided unpaired Student's t-test). Except for the shared sialic acid capping spanning both α2,3- and α2,6-sialylation and core fucosylation on four of the six regulated N-glycans, these six N-glycans appeared structurally (and thus biosynthetically) unrelated, suggesting that alteration in the abundance and/or activity of multiple glycosylation enzymes from a diversity of cellular/tissue origins may contribute to the subtle glycan remodeling observed upon lymph node metastasis. Our quantitative proteomics and glycoproteomics data did not identify any glyco-enzymes that were consistently altered in expression between the two patient groups, likely reflecting that the six altered N-glycans arise from a complex tumor microenvironment in which multiple cell types each with dysregulated glycosylation machineries may contribute to the seemingly unrelated molecular perturbations observed in tumor tissues from OSCC patients with and without lymph node metastasis. The altered expression patterns of these six N-glycans therefore remain mechanistically unexplained.

Bisecting GlcNAcylated N-glycans were consistently observed in the OSCC tumor tissue N-glycome (albeit at relatively low abundance) as confirmed by their early PGC-LC elution and prominent D- and D-GlcNAc fragment ions (29). Aberrant expression of bisecting N-glycans has repeatedly been reported as a feature of human cancers (61) including in colorectal cancer (62), prostate cancer (27), breast cancer (63) and, importantly, in oral cancer (64). Contrasting these reports, our N-glycome and N-glycoproteome data did not reveal any differences in bisecting GlcNAcylation across the N0 and N+ tumor tissues indicating that bisecting GlcNAc is not involved directly in processes related to lymph node metastasis in OSCC.

With a total of 3,117 identified N-glycopeptides from 419 source glycoproteins, the OSCC tumor tissue N-glycoproteome displayed a high complexity and spanned a considerable dynamic range. Recapitulating the stable N-glycome, only relatively modest global N-glycoproteome differences were observed between the N0 and N+ tumor tissues. Importantly, however, 79 N-glycopeptides (~2.5%) were found to be altered between the N0 and N+ patient groups. Similarly to the regulated N-glycans, this subset of altered N-glycopeptides appeared biosynthetically unrelated spanning all N-glycan classes and belonging to different source proteins supporting that lymph node metastasis is a complex process that is not limited to the dysregulation of only a single glycosylation enzyme from a single cellular origin. Importantly, the 79 N-glycopeptides demonstrated a notable potential to accurately stratify N+ from N0 patients. From those, 25 N-glycopeptides belonging mainly to ECM source proteins including several collagens (COL5A1, COL6A3) and laminins (LAMC1, LAMA3) carrying predominantly oligomannosidic or complex/hybrid compositions were associated with key clinicopathological features such as lymph node status and tumor size.

Seeking to mine the information-rich N-glycome and N-glycoproteome data further, unsupervised clustering revealed distinct tumor clusters that displayed interesting differences in the N-glycan class distribution including altered expression of pauci- and oligomannosidic N-glycans. While less explored relative to the more widely recognised cancer-associated changes in the complex/hybrid-type N-glycans and related features such as branching, fucosylation and sialylation (14, 65, 66), we and others have recently shown that both pauci- and oligomannosylation are important N-glycan features that are altered in a wide range of human cancers (67, 68). The aberrant expression of these mannose-terminating N-glycans observed

across the tumor clusters implies regulation, and possibly functional involvement, of the family of α-mannosidases (67, 69) and β-hexosaminidases (70, 71) in OSCC, speculations which require further exploration.

Through enrichment analysis, we investigated putative biological functions of the glycoproteins localizing to each glycopeptide cluster, and observed extracellular matrix (ECM) organization, and platelet and neutrophil degranulation being enriched processes. The cell surface and ECM are composed of hundreds of glycoproteins (72), which are consequently expected to be overrepresented GO terms when analyzing enriched glycopeptides. Moreover, ECM remodeling has repeatedly been linked to cancer development and progression, including changes in the abundance and composition of ECM components, and post-translational modifications (e.g. glycosylation and proteolytic activity) (65, 73), with several studies pointing to tumor-driven changes in the ECM supporting tumor growth, tumor cell migration and metastatic processes (72, 74). Further, platelet degranulation has been associated with cancer metastasis, possibly through the release of platelet factors that can enhance vascular permeability, and consequently promote tumor cell extravasation (75, 76). Finally, elevated neutrophil degranulation observed in a subset of the OSCC primary tumor tissues not only aligns with robust literature pointing to the involvement of neutrophils in tumor progression (77), but also provide a possible mechanistic explanation for the N-glycan differences observed between the tumor clusters including altered expression of paucimannosidic and highly truncated (chitobiose core) N-glycans. Neutrophils are namely known to store and upon activation secrete bioactive proteins carrying peculiar N-glycans including paucimannosidic and chitobiose core type N-glycans as we have demonstrated (78–81). Infiltration of neutrophils in tumors (tumor-associated neutrophils) has been described across cancer types (82), including in oral cancer, in which high neutrophil counts have been associated with poor clinical outcomes (83). In an early study, Wang et al. (84) identified that tongue squamous cell carcinomas exhibiting high neutrophil infiltration displayed increased lymph node metastasis, more advanced clinical stage and increased risk of tumor recurrence.

Interesting associations between key clinical features and two well-separated N-glycan- and N-glycopeptide-guided tumor clusters were identified through clustering analysis. In addition to the association with the lymph node status, the N-glycan-guided tumor clusters informed by Glycan 20a, Glycan 34, Glycan 40a,

Glycan 45b, Glycan 46a, and Glycan 49b were associated with vascular invasion. Likewise, the two N-glycopeptide-informed tumor clusters contained patients that presented with different tumor size and that had received different treatments. Moreover, correlation analysis of the 79 differentially expressed N-glycopeptides revealed associations with 1) the presence of perineural invasion, an important indicator of poor prognosis in oral cancer (85), 2) extracapsular extension, which is associated with poor prognosis and increased risk of recurrence and survival in head and neck cancers (1, 86), 3) the worst pattern of invasion (WPOI) score used for histological risk assessment in oral cancer (40), amongst other key clinical patient features.

Amongst the 25 N-glycopeptides associated with clinicopathological features, five were identified to carry Glycan 40a (a mono-α2,6-sialylated core fucosylated complex-type N-glycan). One of the five Glycan 40a-carrying glycopeptides belongs to butyrophilin subfamily 3 member A1 (BTN3A1 or CD277) and was identified to associate with lymph node status. BTN3A1 plays a well-known role in T-cell activation and in the adaptive immune response, and altered BTN3A1 expression has been reported in different human cancers including in head and neck squamous cell carcinoma and in breast cancer (87). Remarkably, BTN3A1 was not detected in our OSCC proteome profile, but only identified in the HILIC-enriched glycoproteome analysis, highlighting the importance of employing multi-omics approaches to obtain a comprehensive coverage and holistic insights in the molecular complexity and dynamics of the heterogeneous OSCC tumor tissue. Another glycopeptide carrying Glycan 40a, which also associated with lymph node status, belongs to the collagen alpha-3(VI) chain (COL6A3). We have previously reported that COL6A3 is expressed in higher levels in the invasive tumor front compared to the inner tumor, in both tumor stroma and neoplastic islands from OSCC patients (10). All integrin and collagens proteins identified in the previous OSCC proteome (10) were also identified in the present study, some with differential protein abundance in N0 and N+ tissues, including ITGAV, a protein previously identified with higher intensity in patients with lymph node relapse (10).

Interestingly, focused analysis of the N-glycome and N-glycoproteome data demonstrated an association between two N-glycans (Glycan 40a and 46a, both mono-α2,6-sialylated core fucosylated complex-type N-glycans) and three seemingly unrelated neutral N-glycopeptides and low patient survival rate. Powered by our

glycomics-assisted glycoproteomics method, the identification of survival-associated glycans and glycopeptides expands on similar findings made in colorectal cancer (88), in which 12 serum N-glycan structures, aberrantly expressed in the diseased condition compared to non-diseased controls, were associated with overall patient survival, including seven glycan compositions, which were also identified in our N-glycome dataset, albeit with no association with OSCC patient survival. In a gastric cancer study (89), a team of investigators explored the N-glycosylation profile of a Fc fragment of serum IgG from healthy controls, patients with benign stomach diseases and patients with gastric cancer, and observed a higher patient survival in those with comparatively higher levels of G2 glycoforms. Similarly, we observed of higher levels of afamin and CD59 glycopeptides carrying G2 and G2F-type compositions (Glycan 31 and Glycan 35, respectively) in patients with higher OSCC survival. Our discovery of survival-associated N-glycans in FFPE primary tumor tissues is exciting as it opens for previously unexplored avenues to better prognosticate, monitor and manage patients diagnosed with OSCC.

In summary, we report on the complexity and dynamics of the OSCC tumor tissue N-glycoproteome in patients with and without lymph node metastasis. Our comprehensive and quantitative N-glycomics and N-glycoproteomics data have unveiled a range of previously unknown associations between protein N-glycosylation in OSCC tumor tissues and key patient clinical features and survival outcome, and provide an important publicly-available resource to explore the underpinning disease mechanisms and to uncover potential prognostic glyco-markers for OSCC.

3.5. Author Contributions

C. M. C., A. F. P. L., R. K., and M. T.-A. conceptualization; T. M. d. L. M., F. M. d. S. P., A. C. P. R., T. B. B., E. S., and L. L. M., and L. P. K. investigation; C. M. C., A. F. P. L., R. K., and M. T.-A. methodology; C. M. C., A. F. P. L., R. K., and M. T.-A. formal analysis; C. M. C., A. F. P. L., R. K., and M. T.-A. data curation; C. M. C., A. F. P. L., R. K., and M. T.-A. writing–original draft.

3.6. References

1. D. E. Johnson, B. Burtness, C. R. Leemans, V. W. Y. Lui, J. E. Bauman, J. R. Grandis, Head and neck squamous cell carcinoma. Nat. Rev. Dis. Prim. 2020 61. 6, 1–22 (2020).

2. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA. Cancer J. Clin. 71, 209–249 (2021).

3. A. S. Ho, S. Kim, M. Tighiouart, C. Gudino, A. Mita, K. S. Scher, A. Laury, R. Prasad, S. L. Shiao, J. E. Van Eyk, Z. S. Zumsteg, Metastatic lymph node burden and survival in oral cavity cancer. J. Clin. Oncol. 35, 3601–3609 (2017).

4. A. Almangush, A. A. Mäkitie, A. Triantafyllou, R. de Bree, P. Strojan, A. Rinaldo, J. C. Hernandez-Prera, C. Suárez, L. P. Kowalski, A. Ferlito, I. Leivo, Staging and grading of oral squamous cell carcinoma: An update. Oral Oncol. 107, 104799 (2020).

5. M. Sharma, P. Sah, S. Sharma, R. Radhakrishnan, Molecular changes in invasive front of oral cancer. J. Oral Maxillofac. Pathol. 17, 240 (2013).

6. A. Almangush, R. D. Coletta, I. O. Bello, C. Bitu, H. Keski-Säntti, L. K. Mäkinen, J. H. Kauppila, M. Pukkila, J. Hagström, J. Laranne, S. Tommola, Y. Soini, V.-M. Kosma, P. Koivunen, L. P. Kowalski, P. Nieminen, R. Grénman, I. Leivo, T. Salo, A simple novel prognostic model for early stage oral tongue cancer. Int. J. Oral Maxillofac. Surg. 44, 143–150 (2015).

7. F. Chasma, R. Pedr King, S. Y. Ker, Are there diagnostic alternatives to histopathology in detecting oral cancer? Evidence-Based Dent. 2022 231. 23, 24–25 (2022).

8. P. Ostasiewicz, D. F. Zielinska, M. Mann, J. R. Wisniewski, Proteome, phosphoproteome, and N-glycoproteome are quantitatively preserved in formalin-fixed paraffin-embedded tissue and analyzable by high-resolution mass spectrometry. J. Proteome Res. 9, 3688–3700 (2010).

9. A. Mangé, P. Chaurand, H. Perrochia, P. Roger, R. M. Caprioli, J. Solassol, Liquid chromatography-tandem and MALDI imaging mass spectrometry analyses of RCL2/CS100-fixed, paraffin-embedded tissues: Proteomics evaluation of an alternate fixative for biomarker discovery. J. Proteome Res. 8, 5619–5628 (2009).

10. C. M. Carnielli, C. C. S. Macedo, T. De Rossi, D. C. Granato, C. Rivera, R. R. Domingues, B. A. Pauletti, S. Yokoo, H. Heberle, A. F. Busso-Lopes, N. K. Cervigne, I. Sawazaki-Calone, G. V. Meirelles, F. A. Marchi, G. P. Telles, R. Minghim, A. C. P. Ribeiro, T. B. Brandão, G. de Castro, W. A. González-Arriagada, A. Gomes, F. Penteado, A. R. Santos-Silva, M. A. Lopes, P. C. Rodrigues, E. Sundquist, T. Salo, S. D. da Silva, M. A. Alaoui-Jamali, E. Graner, J. W. Fox, R. Della Coletta, A. F. Paes Leme, Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer. Nat. Commun. 9, 3598 (2018).

11. Y. T. Chen, Y. M. Chong, C. W. Cheng, C. L. Ho, H. W. Tsai, F. H. Kasten, Y. L. Chen, C. F. Chang, Identification of novel tumor markers for oral squamous cell carcinoma using glycoproteomic analysis. Clin. Chim. Acta. 420, 45–53 (2013).

12. B. N. Vajaria, K. A. Patel, P. S. Patel, Role of aberrant glycosylation enzymes in oral cancer progression. J. Carcinog. 17 (2018), doi:10.4103/JCAR.JCAR_7_18.

13. G. Liu, P. K. Sengupta, B. Jamal, H. Y. Yang, M. P. Bouchie, V. Lindner, X. Varelas, M. A. Kukuruzinska, N-glycosylation induces the CTHRC1 protein and drives oral cancer cell migration. J. Biol. Chem. (2013), doi:10.1074/jbc.M113.473785.

14. C. Reily, T. J. Stewart, M. B. Renfrow, J. Novak, Glycosylation in health and disease. Nat. Rev. Nephrol. (2019), , doi:10.1038/s41581-019-0129-4.

15. M. Iwatsuki, K. Mimori, T. Yokobori, H. Ishi, T. Beppu, S. Nakamori, H. Baba, M. Mori, Epithelial–mesenchymal transition in cancer development and its clinical significance. Cancer Sci. 101, 293–299 (2010).

16. M. Nita-Lazar, V. Noonan, I. Rebustini, J. Walker, A. S. Menko, M. A. Kukuruzinska, Overexpression of DPAGT1 leads to aberrant N-glycosylation of E-cadherin and cellular discohesion in oral cancer. Cancer Res. (2009), doi:10.1158/0008-5472.CAN-08-4512.

17. B. Jamal, P. K. Sengupta, Z. N. Gao, M. Nita-Lazar, B. Amin, S. Jalisi, M. P. Bouchie, M. A. Kukuruzinska, Aberrant amplification of the crosstalk between canonical Wnt signaling and N-glycosylation gene DPAGT1 promotes oral cancer. Oral Oncol. (2012), doi:10.1016/j.oraloncology.2012.01.010.

18. T. Zhan, N. Rindtorff, M. Boutros, Wnt signaling in cancer. Oncogene (2017), , doi:10.1038/onc.2016.304.

19. J.-T. Chen, C.-H. Chen, K.-L. Ku, M. Hsiao, C.-P. Chiang, T.-L. Hsu, M.-H. Chen, C.-H. Wong, Glycoprotein B7-H3 overexpression and aberrant glycosylation in

oral cancer and immune response. Proc. Natl. Acad. Sci. (2015), doi:10.1073/pnas.1516991112.

20. B. N. Vajaria, K. R. Patel, R. Begum, J. B. Patel, F. D. Shah, G. M. Joshi, P. S. Patel, Salivary Glyco-sialylation changes monitors oral carcinogenesis. Glycoconj. J. (2014), doi:10.1007/s10719-014-9561-7.

21. T. Zhao, C. Zhang, W. Ma, Y. Xiong, J. Yao, G. Yan, G. Chen, H. Lu, A practical approach to enrich intact tryptic N-glycopeptides through size exclusion chromatography and hydrophilicity (SELIC) using an acrylamide-agarose composite gel system. Anal. Chim. Acta (2019), doi:10.1016/j.aca.2019.01.044.

22. P. H. Jensen, N. G. Karlsson, D. Kolarich, N. H. Packer, Structural analysis of N- and O-glycans released from glycoproteins. Nat. Protoc. (2012), doi:10.1038/nprot.2012.063.

23. H. Hinneburg, S. Chatterjee, F. Schirmeister, T. Nguyen-Khuong, N. H. Packer, E. Rapp, M. Thaysen-Andersen, Post-Column Make-Up Flow (PCMF) Enhances the Performance of Capillary-Flow PGC-LC-MS/MS-Based Glycomics. Anal. Chem. (2019), doi:10.1021/acs.analchem.8b05720.

24. A. L. Oberg, O. Vitek, Statistical design of quantitative mass spectrometry-based proteomic experiments. J. Proteome Res. 8 (2009), pp. 2144–2156.

25. W. Revelle, Package “psych” - Procedures for Psychological, Psychometric and Personality Research. R Packag. (2015).

26. C. Ashwood, C. H. Lin, M. Thaysen-Andersen, N. H. Packer, Discrimination of Isomers of Released N- and O-Glycans Using Diagnostic Product Ions in Negative Ion PGC-LC-ESI-MS/MS. J. Am. Soc. Mass Spectrom. (2018), doi:10.1007/s13361-018-1932-z.

27. R. Kawahara, S. Recuero, M. Srougi, K. R. M. Leite, M. Thaysen-Andersen, G. Palmisano, The complexity and dynamics of the tissue glycoproteome associated with prostate cancer progression. Mol. Cell. Proteomics (2020), doi:10.1074/mcp.ra120.002320.

28. C. A. Cooper, E. Gasteiger, N. H. Packer, GlycoMod - A software tool for determining glycosylation compositions from mass spectrometric data. Proteomics (2001), doi:10.1002/1615-9861(200102)1:2<340::AID-PROT340>3.0.CO;2-B.

29. J. L. Abrahams, M. P. Campbell, N. H. Packer, Building a PGC-LC-MS N-glycan retention library and elution mapping resource. Glycoconj. J. 35, 15–29 (2018).

30. A. V. Everest-Dass, D. Kolarich, M. P. Campbell, N. H. Packer, Tandem mass spectra of glycan substructures enable the multistage mass spectrometric identification of determinants on oligosaccharides. Rapid Commun. Mass Spectrom. 27, 931–939 (2013).

31. K. J. Adams, B. Pratt, N. Bose, L. G. Dubois, L. St. John-Williams, K. M. Perrott, K. Ky, P. Kapahi, V. Sharma, M. J. Maccoss, M. A. Moseley, C. A. Colton, B. X. Maclean, B. Schilling, J. W. Thompson, Skyline for Small Molecules: A Unifying Software Package for Quantitative Metabolomics. J. Proteome Res. (2020), doi:10.1021/acs.jproteome.9b00640.

32. A. Ceroni, K. Maass, H. Geyer, R. Geyer, A. Dell, S. M. Haslam, GlycoWorkbench: A tool for the computer-assisted annotation of mass spectra of glycans. J. Proteome Res. (2008), doi:10.1021/pr7008252.

33. J. Villén, S. P. Gygi, The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. Nat. Protoc. 3, 1630–1638 (2008).

34. S. Mysling, G. Palmisano, P. Hojrup, M. Thaysen-Andersen, Utilizing ion-pairing hydrophilic interaction chromatography solid phase extraction for efficient glycopeptide enrichment in glycoproteomics. Anal. Chem. (2010), doi:10.1021/ac100530w.

35. M. Bern, Y. J. Kil, C. Becker, Byonic: Advanced peptide and protein identification software. Curr. Protoc. Bioinforma. (2012), doi:10.1002/0471250953.bi1320s40.

36. M. W. Bern, Y. J. Kil, Two-dimensional target decoy strategy for shotgun proteomics. J. Proteome Res. (2011), doi:10.1021/pr200780j.

37. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372 (2008).

38. J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, M. Mann, Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. J. Proteome Res. 10, 1794–1805 (2011).

39. M. Vaidya, C. Dmello, S. Mogre, Utility of Keratins as Biomarkers for Human Oral Precancer and Cancer. Life 2022, Vol. 12, Page 343. 12, 343 (2022).

40. M. Brandwein-Gensler, M. S. Teixeira, C. M. Lewis, B. Lee, L. Rolnitzky, J. J. Hille, E. Genden, M. L. Urken, B. Y. Wang, Oral squamous cell carcinoma: histologic risk assessment, but not margin status, is strongly predictive of local disease-free and overall survival. Am. J. Surg. Pathol. 29, 167–78 (2005).

41. L. X. Neves, D. C. Granato, A. F. Busso-Lopes, C. M. Carnielli, F. M. de Sá Patroni, T. de Rossi, A. K. Oliveira, A. C. P. Ribeiro, T. B. Brandão, A. N. Rodrigues, P. A. Lacerda, M. Uno, N. K. Cervigne, A. R. Santos-Silva, L. P. Kowalski, M. A. Lopes, A. F. Paes Leme, Peptidomics-driven strategy reveals peptides and predicted proteases associated with oral cancer prognosis. Mol. Cell. Proteomics. 20 (2021), doi:10.1074/MCP.RA120.002227.

42. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez, M. Müller, pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics (2011), doi:10.1186/1471-2105-12-77.

43. D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57 (2009).

44. E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. Meirelles, N. R. Clark, A. Ma'ayan, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 14, 128 (2013).

45. T. H. Chau, A. Chernykh, J. Ugonotti, B. L. Parker, R. Kawahara, M. Thaysen-Andersen, Glycomics-Assisted Glycoproteomics Enables Deep and Unbiased N-Glycoproteome Profiling of Complex Biological Specimens. Methods Mol. Biol. 2628, 235–263 (2023).

46. Z. Sumer-Bayraktar, O. C. Grant, V. Venkatakrishnan, R. J. Woods, N. H. Packer, M. Thaysen-Andersen, Asn347 Glycosylation of Corticosteroid-binding Globulin Fine-tunes the Host Immune Response by Modulating Proteolysis by Pseudomonas aeruginosa and Neutrophil Elastase. J. Biol. Chem. 291, 17727–17742 (2016).

47. M. Thaysen-Andersen, E. Chertova, C. Bergamaschi, E. S. X. Moh, O. Chertov, J. Roser, R. Sowder, J. Bear, J. Lifson, N. H. Packer, B. K. Felber, G. N. Pavlakis, Recombinant human heterodimeric IL-15 complex displays extensive and reproducible N- and O-linked glycosylation. Glycoconj. J. 33, 417–433 (2016).

48. Z. Sumer-Bayraktar, T. Nguyen-Khuong, R. Jayo, D. D. Y. Chen, S. Ali, N. H. Packer, M. Thaysen-Andersen, Micro- and macroheterogeneity of N-glycosylation

yields size and charge isoforms of human sex hormone binding globulin circulating in serum. Proteomics. 12, 3315–3327 (2012).

49. Y. Watanabe, J. D. Allen, D. Wrapp, J. S. McLellan, M. Crispin, Site-specific glycan analysis of the SARS-CoV-2 spike. Science. 369, 330–333 (2020).

50. D. A. Mitchell, A. Kanatas, C. Murphy, P. Chengot, A. B. Smith, T. K. Ong, Margins and survival in oral cancer. Br. J. Oral Maxillofac. Surg. 56, 820–829 (2018).

51. M. H. Shah, S. D. Telang, P. M. Shah, P. S. Patel, Tissue and serum α2-3- and α2-6-linkage specific sialylation changes in oral carcinogenesis. Glycoconj. J. 25, 279–290 (2008).

52. N. Sinevici, S. Mittermayr, G. P. Davey, J. Bones, J. O'Sullivan, Salivary N-glycosylation as a biomarker of oral cancer: A pilot study. Glycobiology (2019), doi:10.1093/glycob/cwz046.

53. S. C. Chang, W. L. Lin, Y. F. Chang, C. T. Lee, J. S. Wu, P. H. Hsu, C. F. Chang, Glycoproteomic identification of novel plasma biomarkers for oral cancer. J. Food Drug Anal. 27, 483–493 (2019).

54. Y.-L. Wong, R. Anand, K. M. Yuen, W. M. W. Mustafa, M. T. Abraham, K. K. Tay, Z. A. A. Rahman, Y. Chen, Identification of potential glycoprotein biomarkers in oral squamous cell carcinoma using sweet strategies. Glycoconj. J. (2021), doi:10.1007/s10719-021-09973-z.

55. S.-Y. Guu, T.-H. Lin, S.-C. Chang, R.-J. Wang, L.-Y. Hung, P.-J. Fang, W.-C. Tang, P. Yu, C.-F. Chang, Serum N-glycome characterization and anti-carbohydrate antibody profiling in oral squamous cell carcinoma patients. PLoS One. 12, e0178927 (2017).

56. M. Thaysen-Andersen, D. Kolarich, N. H. Packer, Glycomics & Glycoproteomics: From Analytics to Function. Mol. Omi. 17, 8–10 (2021).

57. C. Grandi, M. Alloisio, D. Moglia, S. Podrecca, L. Sala, P. Salvatori, R. Molinari, Prognostic significance of lymphatic spread in head and neck carcinomas: therapeutic implications. Head Neck Surg. 8, 67–73 (1985).

58. L. P. Kowalski, A. Sanabria, Elective neck dissection in oral carcinoma: a critical review of the evidence. Acta Otorhinolaryngol. Ital. 27, 113 (2007).

59. R. Kawahara, A. Chernykh, K. Alagesan, M. Bern, W. Cao, R. J. Chalkley, K. Cheng, M. S. Choo, N. Edwards, R. Goldman, M. Hoffmann, Y. Hu, Y. Huang, J. Y. Kim, D. Kletter, B. Liquet, M. Liu, Y. Mechref, B. Meng, S. Neelamegham, T. Nguyen-Khuong, J. Nilsson, A. Pap, G. W. Park, B. L. Parker, C. L. Pegg, J. M.

Penninger, T. K. Phung, M. Pioch, E. Rapp, E. Sakalli, M. Sanda, B. L. Schulz, N. E. Scott, G. Sofronov, J. Stadlmann, S. Y. Vakhrushev, C. M. Woo, H. Y. Wu, P. Yang, W. Ying, H. Zhang, Y. Zhang, J. Zhao, J. Zaia, S. M. Haslam, G. Palmisano, J. S. Yoo, G. Larson, K. H. Khoo, K. F. Medzihradszky, D. Kolarich, N. H. Packer, M. Thaysen-Andersen, Community evaluation of glycoproteomics informatics solutions reveals high-performance search strategies for serum glycopeptide analysis. Nat. Methods 2021 1811. 18, 1304–1316 (2021).

60. T. H. Chau, A. Chernykh, R. Kawahara, M. Thaysen-Andersen, Critical considerations in N-glycoproteomics. Curr. Opin. Chem. Biol. 73 (2023), doi:10.1016/J.CBPA.2023.102272.

61. H. E. Miwa, Y. Song, R. Alvarez, R. D. Cummings, P. Stanley, The bisecting GlcNAc in cell growth control and tumor progression. Glycoconj. J. (2012), doi:10.1007/s10719-012-9373-6.

62. M. K. Sethi, M. Thaysen-Andersen, J. T. Smith, M. S. Baker, N. H. Packer, W. S. Hancock, S. Fanayan, Comparative N-glycan profiling of colorectal cancer cell lines reveals unique bisecting GlcNAc and α-2,3-linked sialic acid determinants are associated with membrane proteins of the more metastatic/aggressive cell lines. J. Proteome Res. 13, 277–288 (2014).

63. L. Cheng, L. Cao, Y. Wu, W. Xie, J. Li, F. Guan, Z. Tan, Bisecting N-Acetylglucosamine on EGFR Inhibits Malignant Phenotype of Breast Cancer via Down-Regulation of EGFR/Erk Signaling. Front. Oncol. 10 (2020), doi:10.3389/FONC.2020.00929.

64. Y. Wang, Q. Li, L. Niu, L. Xu, Y. Guo, L. Wang, C. Guo, Suppression of G6PD induces the expression and bisecting GlcNAc-branched N-glycosylation of E-Cadherin to block epithelial-mesenchymal transition and lymphatic metastasis. Br. J. Cancer (2020), doi:10.1038/s41416-020-1007-3.

65. A. Magalhães, H. O. Duarte, C. A. Reis, Aberrant Glycosylation in Cancer: A Novel Molecular Mechanism Controlling Metastasis. Cancer Cell. 31, 733–735 (2017).

66. O. M. T. Pearce, H. Läubli, Sialic acids in cancer biology and immunity. Glycobiology. 26, 111–128 (2016).

67. S. Chatterjee, J. Ugonotti, L. Y. Lee, A. Everest-Dass, R. Kawahara, M. Thaysen-Andersen, Trends in oligomannosylation and α1,2-mannosidase expression in human cancers. Oncotarget. 12, 2188–2205 (2021).

68. S. Chatterjee, L. Y. Lee, R. Kawahara, J. L. Abrahams, B. Adamczyk, M. Anugraham, C. Ashwood, Z. Sumer-Bayraktar, M. T. Briggs, J. H. L. Chik, A. Everest-Dass, S. Förster, H. Hinneburg, K. R. M. Leite, I. Loke, U. Möginger, E. S. X. Moh, M. Nakano, S. Recuero, M. K. Sethi, M. Srougi, K. Stavenhagen, V. Venkatakrishnan, K. Wongtrakul-Kish, S. Diestel, P. Hoffmann, N. G. Karlsson, D. Kolarich, M. P. Molloy, M. H. Muders, M. K. Oehler, N. H. Packer, G. Palmisano, M. Thaysen-Andersen, Protein Paucimannosylation Is an Enriched N-Glycosylation Signature of Human Cancers. Proteomics. 19, 1900010 (2019).

69. Y. Xiang, K. Karaveg, K. W. Moremen, Substrate recognition and catalysis by GH47 α-mannosidases involved in Asn-linked glycan maturation in the mammalian secretory pathway. Proc. Natl. Acad. Sci. U. S. A. 113, E7890–E7899 (2016).

70. J. Ugonotti, R. Kawahara, I. Loke, Y. Zhu, S. Chatterjee, H. C. Tjondro, Z. Sumer-Bayraktar, S. Neelamegham, M. Thaysen-Andersen, N-acetyl-β-D-hexosaminidases mediate the generation of paucimannosidic proteins via a putative noncanonical truncation pathway in human neutrophils. Glycobiology. 32, 218–229 (2022).

71. K. T. Schjoldager, Y. Narimatsu, H. J. Joshi, H. Clausen, Global view of human protein glycosylation pathways and functions. Nat. Rev. Mol. Cell Biol. 2020 2112. 21, 729–749 (2020).

72. C. Bonnans, J. Chou, Z. Werb, Remodelling the extracellular matrix in development and disease. Nat. Rev. Mol. Cell Biol. 15, 786–801 (2014).

73. G. Marsico, L. Russo, F. Quondamatteo, A. Pandit, Glycosylation and Integrin Regulation in Cancer. Trends in Cancer. 4, 537–552 (2018).

74. J. Winkler, A. Abisoye-Ogunniyan, K. J. Metcalf, Z. Werb, Concepts of extracellular matrix remodelling in tumour progression and metastasis. Nat. Commun. 2020 111. 11, 1–19 (2020).

75. A. Braun, H. J. Anders, T. Gudermann, E. Mammadova-Bach, Platelet-Cancer Interplay: Molecular Mechanisms and New Therapeutic Avenues. Front. Oncol. 11, 2576 (2021).

76. M. Schlesinger, Role of platelets and platelet receptors in cancer metastasis. J. Hematol. Oncol. 2018 111. 11, 1–15 (2018).

77. F. Mollinedo, Neutrophil Degranulation, Plasticity, and Cancer Metastasis. Trends Immunol. 40, 228–242 (2019).

78. H. C. Tjondro, J. Ugonotti, R. Kawahara, S. Chatterjee, S. Chen, F. Soltermann, H. Hinneburg, B. L. Parker, V. Venkatakrishnan, I. Loke, R. Dieckmann, O. C. Grant, J. Bylund, A. Rodger, R. J. Woods, A. Karlsson-Bengtsson, W. B. Struwe, M. Thaysen-Andersen, Hyper-truncated Asn355- and Asn391-glycans modulate the activity of neutrophil granule myeloperoxidase. J. Biol. Chem. 296 (2021), doi:10.1074/JBC.RA120.016342.

79. V. Venkatakrishnan, R. Dieckmann, I. Loke, H. C. Tjondro, S. Chatterjee, J. Bylund, M. Thaysen-Andersen, N. G. Karlsson, A. Karlsson-Bengtsson, Glycan analysis of human neutrophil granules implicates a maturation-dependent glycosylation machinery. J. Biol. Chem. 295, 12648–12660 (2020).

80. I. Loke, N. H. Packer, M. Thaysen-Andersen, Complementary LC-MS/MS-Based N-Glycan, N-Glycopeptide, and Intact N-Glycoprotein Profiling Reveals Unconventional Asn71-Glycosylation of Human Neutrophil Cathepsin G. Biomolecules. 5, 1832–1854 (2015).

81. I. Loke, O. Østergaard, N. H. H. Heegaard, N. H. Packer, M. Thaysen-Andersen, Paucimannose-Rich N -glycosylation of Spatiotemporally Regulated Human Neutrophil Elastase Modulates Its Immune Functions . Mol. Cell. Proteomics (2017), doi:10.1074/mcp.m116.066746.

82. M. E. Shaul, Z. G. Fridlender, Tumour-associated neutrophils in patients with cancer. Nat. Rev. Clin. Oncol. 2019 1610. 16, 601–620 (2019).

83. M. Domnich, J. Riedesel, E. Pylaeva, C. H. L. Kürten, J. Buer, S. Lang, J. Jablonska, Oral Neutrophils: Underestimated Players in Oral Cancer. Front. Immunol. 0, 2529 (2020).

84. N. Wang, Y. Feng, Q. Wang, S. Liu, L. Xiang, M. Sun, X. Zhang, G. Liu, X. Qu, F. Wei, Neutrophils infiltration in the tongue squamous cell carcinoma and its correlation with CEACAM1 expression on tumor cells. PLoS One. 9 (2014), doi:10.1371/JOURNAL.PONE.0089991.

85. I. Chatzistefanou, J. Lubek, K. Markou, R. A. Ord, The role of perineural invasion in treatment decisions for oral cancer patients: A review of the literature. J. Craniomaxillofac. Surg. 45, 821–825 (2017).

86. S. K. Rajappa, U. Maheshwari, D. Ram, V. P. B. Koyyala, G. Mandal, R. Kumar, A. K. Dewan, G. Vishwakarma, Extracapsular extension in oral cavity cancers-predictive factors and impact on recurrence pattern and survival. Int. J. Oral Maxillofac. Surg. 48, 989–994 (2019).

87. F. Liang, C. Zhang, H. Guo, S. H. Gao, F. Y. Yang, G. B. Zhou, G. Z. Wang, Comprehensive analysis of BTN3A1 in cancers: mining of omics data and validation in patient samples and cellular models. FEBS Open Bio. 11, 2586–2599 (2021).

88. S. W. de Vroome, S. Holst, M. R. Girondo, Y. E. M. van der Burgt, W. E. Mesker, R. A. E. M. Tollenaar, M. Wuhrer, Serum N-glycome alterations in colorectal cancer associate with survival. Oncotarget. 9, 30610 (2018).

89. K. Kodar, J. Stadlmann, K. Klaamas, B. Sergeyev, O. Kurtenkov, Immunoglobulin G Fc N-glycan profiling in patients with gastric cancer by LC-ESI-MS: relation to tumor progression and survival. Glycoconjugate J. 2011 291. 29, 57–66 (2011).

90. J. A. Vizcaíno, E. W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Ríos, J. A. Dianes, Z. Sun, T. Farrah, N. Bandeira, P.-A. Binz, I. Xenarios, M. Eisenacher, G. Mayer, L. Gatto, A. Campos, R. J. Chalkley, H.-J. Kraus, J. P. Albar, S. Martinez-Bartolomé, R. Apweiler, G. S. Omenn, L. Martens, A. R. Jones, H. Hermjakob, ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol. 32, 223–226 (2014).

91. P. R. Baker, R. J. Chalkley, MS-Viewer: A Web-based Spectral Viewer for Proteomics Results. Mol. Cell. Proteomics. 13, 1392–1396 (2014).

92. Y. Watanabe, K. F. Aoki-Kinoshita, Y. Ishihama, S. Okuda, GlycoPOST realizes FAIR principles for glycomics mass spectrometry data. Nucleic Acids Res. 49, D1523–D1528 (2021).

4. Referências Bibliografias Gerais

02.02.02.038-0. **SIGTAP - Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e OPM do SUS**. [*S. l.: s. n.*], 2024. Disponível em: http://sigtap.datasus.gov.br/tabela-unificada/app/sec/procedimento/exibir/020202038 0/06/2024. Acesso em: 27 jun. 2024.
02.02.10.001-4. **SIGTAP - Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e OPM do SUS**. [*S. l.: s. n.*], 2024. Disponível em: http://sigtap.datasus.gov.br/tabela-unificada/app/sec/procedimento/exibir/020210001 4/06/2024. Acesso em: 27 jun. 2024.
ARBER, D. A.; ORAZI, A.; HASSERJIAN, R. P.; BOROWITZ, M. J.; CALVO, K. R.; KVASNICKA, H.-M.; WANG, S. A.; BAGG, A.; BARBUI, T.; BRANFORD, S.; BUESO-RAMOS, C. E.; CORTES, J. E.; DAL CIN, P.; DINARDO, C. D.; DOMBRET, H.; DUNCAVAGE, E. J.; EBERT, B. L.; ESTEY, E. H.; FACCHETTI, F.; FOUCAR, K.; GANGAT, N.; GIANELLI, U.; GODLEY, L. A.; GÖKBUGET, N.; GOTLIB, J.; HELLSTRÖM-LINDBERG, E.; HOBBS, G. S.; HOFFMAN, R.; JABBOUR, E. J.; KILADJIAN, J.-J.; LARSON, R. A.; LE BEAU, M. M.; LOH, M. L.-C.; LÖWENBERG, B.; MACINTYRE, E.; MALCOVATI, L.; MULLIGHAN, C. G.; NIEMEYER, C.; ODENIKE, O. M.; OGAWA, S.; ORFAO, A.; PAPAEMMANUIL, E.; PASSAMONTI, F.; PORKKA, K.; PUI, C.-H.; RADICH, J. P.; REITER, A.; ROZMAN, M.; RUDELIUS, M.; SAVONA, M. R.; SCHIFFER, C. A.; SCHMITT-GRAEFF, A.; SHIMAMURA, A.; SIERRA, J.; STOCK, W. A.; STONE, R. M.; TALLMAN, M. S.; THIELE, J.; TIEN, H.-F.; TZANKOV, A.; VANNUCCHI, A. M.; VYAS, P.; WEI, A. H.; WEINBERG, O. K.; WIERZBOWSKA, A.; CAZZOLA, M.; DÖHNER, H.; TEFFERI, A. International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. **Blood**, [*s. l.*], v. 140, n. 11, p. 1200–1228, 2022.
ARBER, D. A.; ORAZI, A.; HASSERJIAN, R.; THIELE, J.; BOROWITZ, M. J.; LE BEAU, M. M.; BLOOMFIELD, C. D.; CAZZOLA, M.; VARDIMAN, J. W. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. **Blood**, [*s. l.*], v. 127, n. 20, p. 2391–2405, 2016.
AREOSA, I.; TORGO, L. Visual Interpretation of Regression Error. *Em*: PROGRESS IN ARTIFICIAL INTELLIGENCE, 2019, Cham. **Anais** [...]. Cham: Springer International Publishing, 2019. p. 473–485.
BENNETT, J. M.; CATOVSKY, D.; DANIEL, M.-T.; FLANDRIN, G.; GALTON, D. a. G.; GRALNICK, H. R.; SULTAN, C. Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. **British Journal of Haematology**, [*s. l.*], v. 33, n. 4, p. 451–458, 1976.
BERGSTRA, J.; BENGIO, Y. **Journal of machine learning research : JMLR.** [*S. l.*]: MIT Press, 2012.(, v. 10). v. 13. Disponível em: http://www.jmlr.org/papers/v13/bergstra12a.html.
BERNARD, E.; TUECHLER, H.; GREENBERG, P. L.; HASSERJIAN, R. P.; ARANGO OSSA, J. E.; NANNYA, Y.; DEVLIN, S. M.; CREIGNOU, M.; PINEL, P.; MONNIER, L.; GUNDEM, G.; MEDINA-MARTINEZ, J. S.; DOMENICO, D.; JÄDERSTEN, M.; GERMING, U.; SANZ, G.; VAN DE LOOSDRECHT, A. A.; KOSMIDER, O.; FOLLO, M. Y.; THOL, F.; ZAMORA, L.; PINHEIRO, R. F.; PELLAGATTI, A.; ELIAS, H. K.; HAASE, D.; GANSTER, C.; ADES, L.; TOBIASSON, M.; PALOMO, L.; DELLA PORTA, M. G.; TAKAORI-KONDO, A.; ISHIKAWA, T.; CHIBA, S.; KASAHARA, S.; MIYAZAKI, Y.; VIALE, A.; HUBERMAN, K.; FENAUX, P.; BELICKOVA, M.; SAVONA,

M. R.; KLIMEK, V. M.; SANTOS, F. P. S.; BOULTWOOD, J.; KOTSIANIDIS, I.; SANTINI, V.; SOLÉ, F.; PLATZBECKER, U.; HEUSER, M.; VALENT, P.; OHYASHIKI, K.; FINELLI, C.; VOSO, M. T.; SHIH, L.-Y.; FONTENAY, M.; JANSEN, J. H.; CERVERA, J.; GATTERMANN, N.; EBERT, B. L.; BEJAR, R.; MALCOVATI, L.; CAZZOLA, M.; OGAWA, S.; HELLSTRÖM-LINDBERG, E.; PAPAEMMANUIL, E. Molecular International Prognostic Scoring System for Myelodysplastic Syndromes. **NEJM Evidence**, [*s. l.*], v. 1, n. 7, p. EVIDoa2200008, 2022.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R. A.; STONE, C. J. **Classification and Regression Trees**. New York: Chapman and Hall/CRC, 1984. 1984.

CERAMI, E.; GAO, J.; DOGRUSOZ, U.; GROSS, B. E.; SUMER, S. O.; AKSOY, B. A.; JACOBSEN, A.; BYRNE, C. J.; HEUER, M. L.; LARSSON, E.; ANTIPIN, Y.; REVA, B.; GOLDBERG, A. P.; SANDER, C.; SCHULTZ, N. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. **Cancer Discovery**, [*s. l.*], v. 2, n. 5, p. 401–404, 2012.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, [*s. l.*], v. 16, p. 321–357, 2002.

CHEN, X.; FU, R.; SHAO, Q.; CHEN, Y.; YE, Q.; LI, S.; HE, X.; ZHU, J. Application of artificial intelligence to pancreatic adenocarcinoma. **Frontiers in Oncology**, [*s. l.*], v. 12, 2022. Disponível em: https://www.frontiersin.org/articles/10.3389/fonc.2022.960056. Acesso em: 15 ago. 2023.

CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System** *Em*: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2016, New York, NY, USA. **Anais** [...]. New York, NY, USA: Association for Computing Machinery, 2016. p. 785–794. Disponível em: https://doi.org/10.1145/2939672.2939785. Acesso em: 20 jul. 2022.

EXPLAINABLE AND INTERPRETABLE MODELS IN COMPUTER VISION AND MACHINE LEARNING. [*S. l.: s. n.*], [*s. d.*]. [*s. d.*]. Disponível em: https://link.springer.com/book/10.1007/978-3-319-98131-4. Acesso em: 20 jul. 2022.

FREUND, Y.; SCHAPIRE, R. E. A desicion-theoretic generalization of on-line learning and an application to boosting. *Em*: COMPUTATIONAL LEARNING THEORY, 1995, Berlin, Heidelberg. **Anais** [...]. Berlin, Heidelberg: Springer, 1995. p. 23–37.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, [*s. l.*], v. 29, n. 5, p. 1189–1232, 2001.

GANIE, S. M.; MALIK, M. B. An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators. **Healthcare Analytics**, [*s. l.*], v. 2, p. 100092, 2022.

GÉRON, A. **Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems**. [*S. l.: s. n.*], 2017. 2017. Disponível em: https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/.

GOLDSTEIN, A.; KAPELNER, A.; BLEICH, J.; PITKIN, E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. **Journal of Computational and Graphical Statistics**, [*s. l.*], v. 24, n. 1, p. 44–65, 2015.

GREENBERG, P. L.; TUECHLER, H.; SCHANZ, J.; SANZ, G.; GARCIA-MANERO, G.; SOLÉ, F.; BENNETT, J. M.; BOWEN, D.; FENAUX, P.; DREYFUS, F.; KANTARJIAN, H.; KUENDGEN, A.; LEVIS, A.; MALCOVATI, L.; CAZZOLA, M.;

CERMAK, J.; FONATSCH, C.; LE BEAU, M. M.; SLOVAK, M. L.; KRIEGER, O.; LUEBBERT, M.; MACIEJEWSKI, J.; MAGALHAES, S. M. M.; MIYAZAKI, Y.; PFEILSTÖCKER, M.; SEKERES, M.; SPERR, W. R.; STAUDER, R.; TAURO, S.; VALENT, P.; VALLESPI, T.; VAN DE LOOSDRECHT, A. A.; GERMING, U.; HAASE, D. Revised International Prognostic Scoring System for Myelodysplastic Syndromes. **Blood**, [*s. l.*], v. 120, n. 12, p. 2454–2465, 2012.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. New York, NY: Springer, 2009. 2009.(Springer Series in Statistics). Disponível em: http://link.springer.com/10.1007/978-0-387-84858-7. Acesso em: 15 ago. 2023.

HOFFBRAND, A. V.; MOSS, P. A. H. **Hoffbrand's essential haematology Seventh edition**. [*S. l.: s. n.*], 2015. 2015. Disponível em: https://www.wiley.com/en-br/Hoffbrand%27s+Essential+Haematology%2C+7th+Edition-p-9781118408674.

INCA. **Quais os efeitos colaterais da quimioterapia? | INCA - Instituto Nacional de Câncer**. , 2019. Disponível em: https://www.inca.gov.br/perguntas-frequentes/quais-os-efeitos-colaterais-da-quimioterapia. Acesso em: 23 set. 2019.

JAGADEV, P.; VIRANI, H. G. Detection of leukemia and its types using image processing and machine learning. *Em*: 2017 INTERNATIONAL CONFERENCE ON TRENDS IN ELECTRONICS AND INFORMATICS (ICEI), 2017, [*s. l.*], . **Anais** [...]. [*S. l.*]: IEEE, 2017. p. 522–526. Disponível em: http://ieeexplore.ieee.org/document/8300983/.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. **LightGBM: A Highly Efficient Gradient Boosting Decision Tree** *Em*: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2017, [*s. l.*], . **Anais** [...]. [*S. l.*]: Curran Associates, Inc., 2017. Disponível em: https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html. Acesso em: 28 jun. 2023.

KORN, C.; MÉNDEZ-FERRER, S. Myeloid malignancies and the microenvironment. **Blood**, [*s. l.*], v. 129, n. 7, p. 811–822, 2017.

KRATZ, A.; FERRARO, M.; SLUSS, P. M.; LEWANDROWSKI, K. B. Normal Reference Laboratory Values. **New England Journal of Medicine**, [*s. l.*], v. 351, n. 15, p. 1548–1563, 2004.

LEE, S.-I.; CELIK, S.; LOGSDON, B. A.; LUNDBERG, S. M.; MARTINS, T. J.; OEHLER, V. G.; ESTEY, E. H.; MILLER, C. P.; CHIEN, S.; DAI, J.; SAXENA, A.; BLAU, C. A.; BECKER, P. S. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. **Nature Communications**, [*s. l.*], v. 9, n. 1, p. 42, 2018.

LIAPIS, K.; PAPADOPOULOS, V.; VRACHIOLIAS, G.; GALANOPOULOS, A. G.; PAPOUTSELIS, M.; PAPAGEORGIOU, S. G.; DIAMANTOPOULOS, P. T.; PAPPA, V.; VINIOU, N.-A.; KOURAKLI, A.; TSOKANAS, D.; VASSILAKOPOULOS, T. P.; HATZIMICHAEL, E.; BOURONIKOU, E.; XIMERI, M.; PONTIKOGLOU, C.; MEGALAKAKI, A.; ZIKOS, P.; PANAYIOTIDIS, P.; DIMOU, M.; KARAKATSANIS, S.; PAPAIOANNOU, M.; VARDI, A.; KONTOPIDOU, F.; HARCHALAKIS, N.; ADAMOPOULOS, I.; SYMEONIDIS, A.; KOTSIANIDIS, I. Refinement of prognosis and the effect of azacitidine in intermediate-risk myelodysplastic syndromes. **Blood Cancer Journal**, [*s. l.*], v. 11, n. 2, p. 1–7, 2021.

LOPES, M. R.; PEREIRA, J. K. N.; DE MELO CAMPOS, P.; MACHADO-NETO, J. A.; TRAINA, F.; SAAD, S. T. O.; FAVARO, P. De novo AML exhibits greater microenvironment dysregulation compared to AML with myelodysplasia-related

changes. **Scientific reports**, [*s. l.*], v. 7, p. 40707, 2017.
LUNDBERG, S.; LEE, S.-I. **A Unified Approach to Interpreting Model Predictions**. Versão arXiv:1705.07874. [*S. l.*]: arXiv, 2017. Disponível em: http://arxiv.org/abs/1705.07874. Acesso em: 20 jul. 2022.
MA, X. Epidemiology of myelodysplastic syndromes. **The American journal of medicine**, [*s. l.*], v. 125, n. 7 Suppl, p. S2-5, 2012.
MULLAINATHAN, S.; OBERMEYER, Z. Does Machine Learning Automate Moral Hazard and Error?. **The American economic review**, [*s. l.*], v. 107, n. 5, p. 476–480, 2017.
MURATI, A.; BRECQUEVILLE, M.; DEVILLIER, R.; MOZZICONACCI, M.-J.; GELSI-BOYER, V.; BIRNBAUM, D. Myeloid malignancies: mutations, models and management. **BMC Cancer**, [*s. l.*], v. 12, n. 1, p. 304, 2012.
NAZHA, A.; SEKERES, M. A.; BEJAR, R.; RAUH, M. J.; OTHUS, M.; KOMROKJI, R. S.; BARNARD, J.; HILTON, C. B.; KERR, C. M.; STEENSMA, D. P.; DEZERN, A.; ROBOZ, G.; GARCIA-MANERO, G.; ERBA, H.; EBERT, B. L.; MACIEJEWSKI, J. P. Genomic Biomarkers to Predict Resistance to Hypomethylating Agents in Patients With Myelodysplastic Syndromes Using Artificial Intelligence. **JCO precision oncology**, [*s. l.*], v. 3, 2019. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/31663066.
NIH. **Impact of NIH Research | National Institutes of Health (NIH)**. , 2019. Disponível em: https://www.nih.gov/about-nih/what-we-do/impact-nih-research. Acesso em: 29 ago. 2019.
NTZIACHRISTOS, P.; MULLENDERS, J.; TRIMARCHI, T.; AIFANTIS, I. Mechanisms of Epigenetic Regulation of Leukemia Onset and Progression. *Em*: ADVANCES IN IMMUNOLOGY. [*S. l.: s. n.*], 2013. v. 117, p. 1–38. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/23611284 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3830954 https://linkinghub.elsevier.com/retrieve/pii/B9780124105249000013.
OUBOUNYT, M.; ADLUNG, L.; PATRONI, F.; WENKE, N. K.; MAIER, A.; HARTUNG, M.; BAUMBACH, J.; ELKJAER, M. L. Inference of differential key regulatory networks and mechanistic drug repurposing candidates from scRNA-seq data with SCANet. **Bioinformatics**, [*s. l.*], v. 39, n. 11, p. btad644, 2023.
PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; MÜLLER, A.; NOTHMAN, J.; LOUPPE, G.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, É. Scikit-learn: Machine Learning in Python. [*s. l.*], 2012. Disponível em: http://arxiv.org/abs/1201.0490.
PROKHORENKOVA, L.; GUSEV, G.; VOROBEV, A.; DOROGUSH, A. V.; GULIN, A. **CatBoost: unbiased boosting with categorical features** Versão arXiv:1706.09516. [*S. l.*]: arXiv, 2019. Disponível em: http://arxiv.org/abs/1706.09516. Acesso em: 28 jun. 2023.
REILLY, B.; TANAKA, T. N.; DIEP, D.; YEERNA, H.; TAMAYO, P.; ZHANG, K.; BEJAR, R. DNA methylation identifies genetically and prognostically distinct subtypes of myelodysplastic syndromes. **Blood Advances**, [*s. l.*], v. 3, n. 19, p. 2845–2858, 2019.
RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. **"Why Should I Trust You?": Explaining the Predictions of Any Classifier** *Em*: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2016, New York, NY, USA. **Anais** [...]. New York, NY, USA:

Association for Computing Machinery, 2016. p. 1135–1144. Disponível em: https://doi.org/10.1145/2939672.2939778. Acesso em: 20 jul. 2022.
RICHARD A. DUSCHL; HEIDI A. SCHWEINGRUBER; ANDREW W. SHOUSE. **Taking Science to School: Learning and Teaching Science in Grades K-8**. [*S. l.*]: National Academies of Sciences, Engineering, and Medicine, 2007. 2007. Disponível em: https://www.nap.edu/read/11625/chapter/8. Acesso em: 15 ago. 2023.
SHAPLEY, L. S. 17. A Value for n-Person Games. *Em*: 17. A VALUE FOR N-PERSON GAMES. [*S. l.*]: Princeton University Press, 1953. p. 307–318. Disponível em: https://www.degruyter.com/document/doi/10.1515/9781400881970-018/html. Acesso em: 20 jul. 2022.
THE MYTHOS OF MODEL INTERPRETABILITY - ACM QUEUE. [*S. l.: s. n.*], [*s. d.*]. Disponível em: https://queue.acm.org/detail.cfm?id=3241340. Acesso em: 20 jul. 2022.
TOMEK, I. Two Modifications of CNN. **IEEE Transactions on Systems, Man, and Cybernetics**, [*s. l.*], v. SMC-6, n. 11, p. 769–772, 1976.
WALTER, R. B.; OTHUS, M.; BURNETT, A. K.; LÖWENBERG, B.; KANTARJIAN, H. M.; OSSENKOPPELE, G. J.; HILLS, R. K.; VAN MONTFORT, K. G. M.; RAVANDI, F.; EVANS, A.; PIERCE, S. R.; APPELBAUM, F. R.; ESTEY, E. H. Significance of FAB subclassification of “acute myeloid leukemia, NOS” in the 2008 WHO classification: analysis of 5848 newly diagnosed patients. **Blood**, [*s. l.*], v. 121, n. 13, p. 2424, 2013.
WEST, A. H.; GODLEY, L. A.; CHURPEK, J. E. Familial myelodysplastic syndrome/acute leukemia syndromes: a review and utility for translational investigations. **Annals of the New York Academy of Sciences**, [*s. l.*], v. 1310, n. 1, p. 111–118, 2014.
YAUNEY, G.; SHAH, P. **Reinforcement Learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regimen Selection**. , 2018. Disponível em: http://proceedings.mlr.press/v85/yauney18a.html.
ZITNIK, M.; NGUYEN, F.; WANG, B.; LESKOVEC, J.; GOLDENBERG, A.; HOFFMAN, M. M. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. arXiv e-prints, [*s. l.*], 2018. Disponível em: http://arxiv.org/abs/1807.00123 http://dx.doi.org/10.1016/j.inffus.2018.09.012.

1. D. E. Johnson, B. Burtness, C. R. Leemans, V. W. Y. Lui, J. E. Bauman, J. R. Grandis, Head and neck squamous cell carcinoma. Nat. Rev. Dis. Prim. 2020 61. 6, 1–22 (2020).

2. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA. Cancer J. Clin. 71, 209–249 (2021).

3. A. S. Ho, S. Kim, M. Tighiouart, C. Gudino, A. Mita, K. S. Scher, A. Laury, R. Prasad, S. L. Shiao, J. E. Van Eyk, Z. S. Zumsteg, Metastatic lymph node burden and survival in oral cavity cancer. J. Clin. Oncol. 35, 3601–3609 (2017).

4. A. Almangush, A. A. Mäkitie, A. Triantafyllou, R. de Bree, P. Strojan, A. Rinaldo, J. C. Hernandez-Prera, C. Suárez, L. P. Kowalski, A. Ferlito, I. Leivo, Staging and grading of oral squamous cell carcinoma: An update. Oral Oncol. 107, 104799 (2020).

5. M. Sharma, P. Sah, S. Sharma, R. Radhakrishnan, Molecular changes in invasive front of oral cancer. J. Oral Maxillofac. Pathol. 17, 240 (2013).

6. A. Almangush, R. D. Coletta, I. O. Bello, C. Bitu, H. Keski-Säntti, L. K. Mäkinen, J. H. Kauppila, M. Pukkila, J. Hagström, J. Laranne, S. Tommola, Y. Soini, V.-M. Kosma, P. Koivunen, L. P. Kowalski, P. Nieminen, R. Grénman, I. Leivo, T. Salo, A simple novel prognostic model for early stage oral tongue cancer. Int. J. Oral Maxillofac. Surg. 44, 143–150 (2015).

7. F. Chasma, R. Pedr King, S. Y. Ker, Are there diagnostic alternatives to histopathology in detecting oral cancer? Evidence-Based Dent. 2022 231. 23, 24–25 (2022).

8. P. Ostasiewicz, D. F. Zielinska, M. Mann, J. R. Wisniewski, Proteome, phosphoproteome, and N-glycoproteome are quantitatively preserved in formalin-fixed paraffin-embedded tissue and analyzable by high-resolution mass spectrometry. J. Proteome Res. 9, 3688–3700 (2010).

9. A. Mangé, P. Chaurand, H. Perrochia, P. Roger, R. M. Caprioli, J. Solassol, Liquid chromatography-tandem and MALDI imaging mass spectrometry analyses of RCL2/CS100-fixed, paraffin-embedded tissues: Proteomics evaluation of an alternate fixative for biomarker discovery. J. Proteome Res. 8, 5619–5628 (2009).

10. C. M. Carnielli, C. C. S. Macedo, T. De Rossi, D. C. Granato, C. Rivera, R. R. Domingues, B. A. Pauletti, S. Yokoo, H. Heberle, A. F. Busso-Lopes, N. K. Cervigne, I. Sawazaki-Calone, G. V. Meirelles, F. A. Marchi, G. P. Telles, R. Minghim, A. C. P. Ribeiro, T. B. Brandão, G. de Castro, W. A. González-Arriagada, A. Gomes, F. Penteado, A. R. Santos-Silva, M. A. Lopes, P. C. Rodrigues, E. Sundquist, T. Salo, S. D. da Silva, M. A. Alaoui-Jamali, E. Graner, J. W. Fox, R. Della Coletta, A. F. Paes Leme, Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer. Nat. Commun. 9, 3598 (2018).

11. Y. T. Chen, Y. M. Chong, C. W. Cheng, C. L. Ho, H. W. Tsai, F. H. Kasten, Y. L. Chen, C. F. Chang, Identification of novel tumor markers for oral squamous cell carcinoma using glycoproteomic analysis. Clin. Chim. Acta. 420, 45–53 (2013).

12. B. N. Vajaria, K. A. Patel, P. S. Patel, Role of aberrant glycosylation enzymes in oral cancer progression. J. Carcinog. 17 (2018), doi:10.4103/JCAR.JCAR_7_18.

13. G. Liu, P. K. Sengupta, B. Jamal, H. Y. Yang, M. P. Bouchie, V. Lindner, X. Varelas, M. A. Kukuruzinska, N-glycosylation induces the CTHRC1 protein and drives oral cancer cell migration. J. Biol. Chem. (2013), doi:10.1074/jbc.M113.473785.

14. C. Reily, T. J. Stewart, M. B. Renfrow, J. Novak, Glycosylation in health and disease. Nat. Rev. Nephrol. (2019), , doi:10.1038/s41581-019-0129-4.

15. M. Iwatsuki, K. Mimori, T. Yokobori, H. Ishi, T. Beppu, S. Nakamori, H. Baba, M. Mori, Epithelial–mesenchymal transition in cancer development and its clinical significance. Cancer Sci. 101, 293–299 (2010).

16. M. Nita-Lazar, V. Noonan, I. Rebustini, J. Walker, A. S. Menko, M. A. Kukuruzinska, Overexpression of DPAGT1 leads to aberrant N-glycosylation of E-cadherin and cellular discohesion in oral cancer. Cancer Res. (2009), doi:10.1158/0008-5472.CAN-08-4512.

17. B. Jamal, P. K. Sengupta, Z. N. Gao, M. Nita-Lazar, B. Amin, S. Jalisi, M. P. Bouchie, M. A. Kukuruzinska, Aberrant amplification of the crosstalk between canonical Wnt signaling and N-glycosylation gene DPAGT1 promotes oral cancer. Oral Oncol. (2012), doi:10.1016/j.oraloncology.2012.01.010.

18. T. Zhan, N. Rindtorff, M. Boutros, Wnt signaling in cancer. Oncogene (2017), , doi:10.1038/onc.2016.304.

19. J.-T. Chen, C.-H. Chen, K.-L. Ku, M. Hsiao, C.-P. Chiang, T.-L. Hsu, M.-H. Chen, C.-H. Wong, Glycoprotein B7-H3 overexpression and aberrant glycosylation in oral cancer and immune response. Proc. Natl. Acad. Sci. (2015), doi:10.1073/pnas.1516991112.

20. B. N. Vajaria, K. R. Patel, R. Begum, J. B. Patel, F. D. Shah, G. M. Joshi, P. S. Patel, Salivary Glyco-sialylation changes monitors oral carcinogenesis. Glycoconj. J. (2014), doi:10.1007/s10719-014-9561-7.

21. T. Zhao, C. Zhang, W. Ma, Y. Xiong, J. Yao, G. Yan, G. Chen, H. Lu, A practical approach to enrich intact tryptic N-glycopeptides through size exclusion chromatography and hydrophilicity (SELIC) using an acrylamide-agarose composite gel system. Anal. Chim. Acta (2019), doi:10.1016/j.aca.2019.01.044.

22. P. H. Jensen, N. G. Karlsson, D. Kolarich, N. H. Packer, Structural analysis of N- and O-glycans released from glycoproteins. Nat. Protoc. (2012), doi:10.1038/nprot.2012.063.

23. H. Hinneburg, S. Chatterjee, F. Schirmeister, T. Nguyen-Khuong, N. H. Packer, E. Rapp, M. Thaysen-Andersen, Post-Column Make-Up Flow (PCMF) Enhances the Performance of Capillary-Flow PGC-LC-MS/MS-Based Glycomics. Anal. Chem. (2019), doi:10.1021/acs.analchem.8b05720.

24. A. L. Oberg, O. Vitek, Statistical design of quantitative mass spectrometry-based proteomic experiments. J. Proteome Res. 8 (2009), pp. 2144–2156.

25. W. Revelle, Package “psych” - Procedures for Psychological, Psychometric and Personality Research. R Packag. (2015).

26. C. Ashwood, C. H. Lin, M. Thaysen-Andersen, N. H. Packer, Discrimination of Isomers of Released N- and O-Glycans Using Diagnostic Product Ions in Negative Ion PGC-LC-ESI-MS/MS. J. Am. Soc. Mass Spectrom. (2018), doi:10.1007/s13361-018-1932-z.

27. R. Kawahara, S. Recuero, M. Srougi, K. R. M. Leite, M. Thaysen-Andersen, G. Palmisano, The complexity and dynamics of the tissue glycoproteome associated with prostate cancer progression. Mol. Cell. Proteomics (2020), doi:10.1074/mcp.ra120.002320.

28. C. A. Cooper, E. Gasteiger, N. H. Packer, GlycoMod - A software tool for determining glycosylation compositions from mass spectrometric data. Proteomics (2001), doi:10.1002/1615-9861(200102)1:2<340::AID-PROT340>3.0.CO;2-B.

29. J. L. Abrahams, M. P. Campbell, N. H. Packer, Building a PGC-LC-MS N-glycan retention library and elution mapping resource. Glycoconj. J. 35, 15–29 (2018).

30. A. V. Everest-Dass, D. Kolarich, M. P. Campbell, N. H. Packer, Tandem mass spectra of glycan substructures enable the multistage mass spectrometric identification of determinants on oligosaccharides. Rapid Commun. Mass Spectrom. 27, 931–939 (2013).

31. K. J. Adams, B. Pratt, N. Bose, L. G. Dubois, L. St. John-Williams, K. M. Perrott, K. Ky, P. Kapahi, V. Sharma, M. J. Maccoss, M. A. Moseley, C. A. Colton, B. X. Maclean, B. Schilling, J. W. Thompson, Skyline for Small Molecules: A Unifying Software Package for Quantitative Metabolomics. J. Proteome Res. (2020), doi:10.1021/acs.jproteome.9b00640.

32. A. Ceroni, K. Maass, H. Geyer, R. Geyer, A. Dell, S. M. Haslam, GlycoWorkbench: A tool for the computer-assisted annotation of mass spectra of glycans. J. Proteome Res. (2008), doi:10.1021/pr7008252.

33. J. Villén, S. P. Gygi, The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. Nat. Protoc. 3, 1630–1638 (2008).

34. S. Mysling, G. Palmisano, P. Hojrup, M. Thaysen-Andersen, Utilizing ion-pairing hydrophilic interaction chromatography solid phase extraction for efficient glycopeptide enrichment in glycoproteomics. Anal. Chem. (2010), doi:10.1021/ac100530w.

35. M. Bern, Y. J. Kil, C. Becker, Byonic: Advanced peptide and protein identification software. Curr. Protoc. Bioinforma. (2012), doi:10.1002/0471250953.bi1320s40.

36. M. W. Bern, Y. J. Kil, Two-dimensional target decoy strategy for shotgun proteomics. J. Proteome Res. (2011), doi:10.1021/pr200780j.

37. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372 (2008).

38. J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, M. Mann, Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. J. Proteome Res. 10, 1794–1805 (2011).

39. M. Vaidya, C. Dmello, S. Mogre, Utility of Keratins as Biomarkers for Human Oral Precancer and Cancer. Life 2022, Vol. 12, Page 343. 12, 343 (2022).

40. M. Brandwein-Gensler, M. S. Teixeira, C. M. Lewis, B. Lee, L. Rolnitzky, J. J. Hille, E. Genden, M. L. Urken, B. Y. Wang, Oral squamous cell carcinoma: histologic risk assessment, but not margin status, is strongly predictive of local disease-free and overall survival. Am. J. Surg. Pathol. 29, 167–78 (2005).

41. L. X. Neves, D. C. Granato, A. F. Busso-Lopes, C. M. Carnielli, F. M. de Sá Patroni, T. de Rossi, A. K. Oliveira, A. C. P. Ribeiro, T. B. Brandão, A. N. Rodrigues, P. A. Lacerda, M. Uno, N. K. Cervigne, A. R. Santos-Silva, L. P. Kowalski, M. A. Lopes, A. F. Paes Leme, Peptidomics-driven strategy reveals peptides and predicted proteases associated with oral cancer prognosis. Mol. Cell. Proteomics. 20 (2021), doi:10.1074/MCP.RA120.002227.

42. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez, M. Müller, pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics (2011), doi:10.1186/1471-2105-12-77.

43. D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57 (2009).

44. E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. Meirelles, N. R. Clark, A. Ma'ayan, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 14, 128 (2013).

45. T. H. Chau, A. Chernykh, J. Ugonotti, B. L. Parker, R. Kawahara, M. Thaysen-Andersen, Glycomics-Assisted Glycoproteomics Enables Deep and Unbiased N-Glycoproteome Profiling of Complex Biological Specimens. Methods Mol. Biol. 2628, 235–263 (2023).

46. Z. Sumer-Bayraktar, O. C. Grant, V. Venkatakrishnan, R. J. Woods, N. H. Packer, M. Thaysen-Andersen, Asn347 Glycosylation of Corticosteroid-binding Globulin Fine-tunes the Host Immune Response by Modulating Proteolysis by Pseudomonas aeruginosa and Neutrophil Elastase. J. Biol. Chem. 291, 17727–17742 (2016).

47. M. Thaysen-Andersen, E. Chertova, C. Bergamaschi, E. S. X. Moh, O. Chertov, J. Roser, R. Sowder, J. Bear, J. Lifson, N. H. Packer, B. K. Felber, G. N. Pavlakis, Recombinant human heterodimeric IL-15 complex displays extensive and reproducible N- and O-linked glycosylation. Glycoconj. J. 33, 417–433 (2016).

48. Z. Sumer-Bayraktar, T. Nguyen-Khuong, R. Jayo, D. D. Y. Chen, S. Ali, N. H. Packer, M. Thaysen-Andersen, Micro- and macroheterogeneity of N-glycosylation yields size and charge isoforms of human sex hormone binding globulin circulating in serum. Proteomics. 12, 3315–3327 (2012).

49. Y. Watanabe, J. D. Allen, D. Wrapp, J. S. McLellan, M. Crispin, Site-specific glycan analysis of the SARS-CoV-2 spike. Science. 369, 330–333 (2020).

50. D. A. Mitchell, A. Kanatas, C. Murphy, P. Chengot, A. B. Smith, T. K. Ong, Margins and survival in oral cancer. Br. J. Oral Maxillofac. Surg. 56, 820–829 (2018).

51. M. H. Shah, S. D. Telang, P. M. Shah, P. S. Patel, Tissue and serum α2-3- and α2-6-linkage specific sialylation changes in oral carcinogenesis. Glycoconj. J. 25, 279–290 (2008).

52. N. Sinevici, S. Mittermayr, G. P. Davey, J. Bones, J. O'Sullivan, Salivary N-glycosylation as a biomarker of oral cancer: A pilot study. Glycobiology (2019), doi:10.1093/glycob/cwz046.

53. S. C. Chang, W. L. Lin, Y. F. Chang, C. T. Lee, J. S. Wu, P. H. Hsu, C. F. Chang, Glycoproteomic identification of novel plasma biomarkers for oral cancer. J. Food Drug Anal. 27, 483–493 (2019).

54. Y.-L. Wong, R. Anand, K. M. Yuen, W. M. W. Mustafa, M. T. Abraham, K. K. Tay, Z. A. A. Rahman, Y. Chen, Identification of potential glycoprotein biomarkers in oral squamous cell carcinoma using sweet strategies. Glycoconj. J. (2021), doi:10.1007/s10719-021-09973-z.

55. S.-Y. Guu, T.-H. Lin, S.-C. Chang, R.-J. Wang, L.-Y. Hung, P.-J. Fang, W.-C. Tang, P. Yu, C.-F. Chang, Serum N-glycome characterization and anti-carbohydrate antibody profiling in oral squamous cell carcinoma patients. PLoS One. 12, e0178927 (2017).

56. M. Thaysen-Andersen, D. Kolarich, N. H. Packer, Glycomics & Glycoproteomics: From Analytics to Function. Mol. Omi. 17, 8–10 (2021).

57. C. Grandi, M. Alloisio, D. Moglia, S. Podrecca, L. Sala, P. Salvatori, R. Molinari, Prognostic significance of lymphatic spread in head and neck carcinomas: therapeutic implications. Head Neck Surg. 8, 67–73 (1985).

58. L. P. Kowalski, A. Sanabria, Elective neck dissection in oral carcinoma: a critical review of the evidence. Acta Otorhinolaryngol. Ital. 27, 113 (2007).

59. R. Kawahara, A. Chernykh, K. Alagesan, M. Bern, W. Cao, R. J. Chalkley, K. Cheng, M. S. Choo, N. Edwards, R. Goldman, M. Hoffmann, Y. Hu, Y. Huang, J. Y. Kim, D. Kletter, B. Liquet, M. Liu, Y. Mechref, B. Meng, S. Neelamegham, T. Nguyen-Khuong, J. Nilsson, A. Pap, G. W. Park, B. L. Parker, C. L. Pegg, J. M. Penninger, T. K. Phung, M. Pioch, E. Rapp, E. Sakalli, M. Sanda, B. L. Schulz, N. E. Scott, G. Sofronov, J. Stadlmann, S. Y. Vakhrushev, C. M. Woo, H. Y. Wu, P. Yang, W. Ying, H. Zhang, Y. Zhang, J. Zhao, J. Zaia, S. M. Haslam, G. Palmisano, J. S. Yoo, G. Larson, K. H. Khoo, K. F. Medzihradszky, D. Kolarich, N. H. Packer, M. Thaysen-Andersen, Community evaluation of glycoproteomics informatics solutions reveals high-performance search strategies for serum glycopeptide analysis. Nat. Methods 2021 1811. 18, 1304–1316 (2021).

60. T. H. Chau, A. Chernykh, R. Kawahara, M. Thaysen-Andersen, Critical considerations in N-glycoproteomics. Curr. Opin. Chem. Biol. 73 (2023), doi:10.1016/J.CBPA.2023.102272.

61. H. E. Miwa, Y. Song, R. Alvarez, R. D. Cummings, P. Stanley, The bisecting GlcNAc in cell growth control and tumor progression. Glycoconj. J. (2012), doi:10.1007/s10719-012-9373-6.

62. M. K. Sethi, M. Thaysen-Andersen, J. T. Smith, M. S. Baker, N. H. Packer, W. S. Hancock, S. Fanayan, Comparative N-glycan profiling of colorectal cancer cell lines reveals unique bisecting GlcNAc and α-2,3-linked sialic acid determinants are associated with membrane proteins of the more metastatic/aggressive cell lines. J. Proteome Res. 13, 277–288 (2014).

63. L. Cheng, L. Cao, Y. Wu, W. Xie, J. Li, F. Guan, Z. Tan, Bisecting N-Acetylglucosamine on EGFR Inhibits Malignant Phenotype of Breast Cancer via Down-Regulation of EGFR/Erk Signaling. Front. Oncol. 10 (2020), doi:10.3389/FONC.2020.00929.

64. Y. Wang, Q. Li, L. Niu, L. Xu, Y. Guo, L. Wang, C. Guo, Suppression of G6PD induces the expression and bisecting GlcNAc-branched N-glycosylation of E-Cadherin to block epithelial-mesenchymal transition and lymphatic metastasis. Br. J. Cancer (2020), doi:10.1038/s41416-020-1007-3.

65. A. Magalhães, H. O. Duarte, C. A. Reis, Aberrant Glycosylation in Cancer: A Novel Molecular Mechanism Controlling Metastasis. Cancer Cell. 31, 733–735 (2017).

66. O. M. T. Pearce, H. Läubli, Sialic acids in cancer biology and immunity. Glycobiology. 26, 111–128 (2016).

67. S. Chatterjee, J. Ugonotti, L. Y. Lee, A. Everest-Dass, R. Kawahara, M. Thaysen-Andersen, Trends in oligomannosylation and α1,2-mannosidase expression in human cancers. Oncotarget. 12, 2188–2205 (2021).

68. S. Chatterjee, L. Y. Lee, R. Kawahara, J. L. Abrahams, B. Adamczyk, M. Anugraham, C. Ashwood, Z. Sumer-Bayraktar, M. T. Briggs, J. H. L. Chik, A. Everest-Dass, S. Förster, H. Hinneburg, K. R. M. Leite, I. Loke, U. Möginger, E. S. X. Moh, M. Nakano, S. Recuero, M. K. Sethi, M. Srougi, K. Stavenhagen, V. Venkatakrishnan, K. Wongtrakul-Kish, S. Diestel, P. Hoffmann, N. G. Karlsson, D. Kolarich, M. P. Molloy, M. H. Muders, M. K. Oehler, N. H. Packer, G. Palmisano, M.

Thaysen-Andersen, Protein Paucimannosylation Is an Enriched N-Glycosylation Signature of Human Cancers. Proteomics. 19, 1900010 (2019).

69. Y. Xiang, K. Karaveg, K. W. Moremen, Substrate recognition and catalysis by GH47 α-mannosidases involved in Asn-linked glycan maturation in the mammalian secretory pathway. Proc. Natl. Acad. Sci. U. S. A. 113, E7890–E7899 (2016).

70. J. Ugonotti, R. Kawahara, I. Loke, Y. Zhu, S. Chatterjee, H. C. Tjondro, Z. Sumer-Bayraktar, S. Neelamegham, M. Thaysen-Andersen, N-acetyl-β-D-hexosaminidases mediate the generation of paucimannosidic proteins via a putative noncanonical truncation pathway in human neutrophils. Glycobiology. 32, 218–229 (2022).

71. K. T. Schjoldager, Y. Narimatsu, H. J. Joshi, H. Clausen, Global view of human protein glycosylation pathways and functions. Nat. Rev. Mol. Cell Biol. 2020 2112. 21, 729–749 (2020).

72. C. Bonnans, J. Chou, Z. Werb, Remodelling the extracellular matrix in development and disease. Nat. Rev. Mol. Cell Biol. 15, 786–801 (2014).

73. G. Marsico, L. Russo, F. Quondamatteo, A. Pandit, Glycosylation and Integrin Regulation in Cancer. Trends in Cancer. 4, 537–552 (2018).

74. J. Winkler, A. Abisoye-Ogunniyan, K. J. Metcalf, Z. Werb, Concepts of extracellular matrix remodelling in tumour progression and metastasis. Nat. Commun. 2020 111. 11, 1–19 (2020).

75. A. Braun, H. J. Anders, T. Gudermann, E. Mammadova-Bach, Platelet-Cancer Interplay: Molecular Mechanisms and New Therapeutic Avenues. Front. Oncol. 11, 2576 (2021).

76. M. Schlesinger, Role of platelets and platelet receptors in cancer metastasis. J. Hematol. Oncol. 2018 111. 11, 1–15 (2018).

77. F. Mollinedo, Neutrophil Degranulation, Plasticity, and Cancer Metastasis. Trends Immunol. 40, 228–242 (2019).

78. H. C. Tjondro, J. Ugonotti, R. Kawahara, S. Chatterjee, S. Chen, F. Soltermann, H. Hinneburg, B. L. Parker, V. Venkatakrishnan, I. Loke, R. Dieckmann, O. C. Grant, J. Bylund, A. Rodger, R. J. Woods, A. Karlsson-Bengtsson, W. B. Struwe, M. Thaysen-Andersen, Hyper-truncated Asn355- and Asn391-glycans modulate the activity of neutrophil granule myeloperoxidase. J. Biol. Chem. 296 (2021), doi:10.1074/JBC.RA120.016342.

79. V. Venkatakrishnan, R. Dieckmann, I. Loke, H. C. Tjondro, S. Chatterjee, J. Bylund, M. Thaysen-Andersen, N. G. Karlsson, A. Karlsson-Bengtsson, Glycan analysis of human neutrophil granules implicates a maturation-dependent glycosylation machinery. J. Biol. Chem. 295, 12648–12660 (2020).

80. I. Loke, N. H. Packer, M. Thaysen-Andersen, Complementary LC-MS/MS-Based N-Glycan, N-Glycopeptide, and Intact N-Glycoprotein Profiling Reveals Unconventional Asn71-Glycosylation of Human Neutrophil Cathepsin G. Biomolecules. 5, 1832–1854 (2015).

81. I. Loke, O. Østergaard, N. H. H. Heegaard, N. H. Packer, M. Thaysen-Andersen, Paucimannose-Rich N -glycosylation of Spatiotemporally Regulated Human Neutrophil Elastase Modulates Its Immune Functions . Mol. Cell. Proteomics (2017), doi:10.1074/mcp.m116.066746.

82. M. E. Shaul, Z. G. Fridlender, Tumour-associated neutrophils in patients with cancer. Nat. Rev. Clin. Oncol. 2019 1610. 16, 601–620 (2019).

83. M. Domnich, J. Riedesel, E. Pylaeva, C. H. L. Kürten, J. Buer, S. Lang, J. Jablonska, Oral Neutrophils: Underestimated Players in Oral Cancer. Front. Immunol. 0, 2529 (2020).

84. N. Wang, Y. Feng, Q. Wang, S. Liu, L. Xiang, M. Sun, X. Zhang, G. Liu, X. Qu, F. Wei, Neutrophils infiltration in the tongue squamous cell carcinoma and its correlation with CEACAM1 expression on tumor cells. PLoS One. 9 (2014), doi:10.1371/JOURNAL.PONE.0089991.

85. I. Chatzistefanou, J. Lubek, K. Markou, R. A. Ord, The role of perineural invasion in treatment decisions for oral cancer patients: A review of the literature. J. Craniomaxillofac. Surg. 45, 821–825 (2017).

86. S. K. Rajappa, U. Maheshwari, D. Ram, V. P. B. Koyyala, G. Mandal, R. Kumar, A. K. Dewan, G. Vishwakarma, Extracapsular extension in oral cavity cancers-predictive factors and impact on recurrence pattern and survival. Int. J. Oral Maxillofac. Surg. 48, 989–994 (2019).

87. F. Liang, C. Zhang, H. Guo, S. H. Gao, F. Y. Yang, G. B. Zhou, G. Z. Wang, Comprehensive analysis of BTN3A1 in cancers: mining of omics data and validation in patient samples and cellular models. FEBS Open Bio. 11, 2586–2599 (2021).

88. S. W. de Vroome, S. Holst, M. R. Girondo, Y. E. M. van der Burgt, W. E. Mesker, R. A. E. M. Tollenaar, M. Wuhrer, Serum N-glycome alterations in colorectal cancer associate with survival. Oncotarget. 9, 30610 (2018).

89. K. Kodar, J. Stadlmann, K. Klaamas, B. Sergeyev, O. Kurtenkov, Immunoglobulin G Fc N-glycan profiling in patients with gastric cancer by LC-ESI-MS: relation to tumor progression and survival. Glycoconjugate J. 2011 291. 29, 57–66 (2011).

90. J. A. Vizcaíno, E. W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Ríos, J. A. Dianes, Z. Sun, T. Farrah, N. Bandeira, P.-A. Binz, I. Xenarios, M. Eisenacher, G. Mayer, L. Gatto, A. Campos, R. J. Chalkley, H.-J. Kraus, J. P. Albar, S. Martinez-Bartolomé, R. Apweiler, G. S. Omenn, L. Martens, A. R. Jones, H. Hermjakob, ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol. 32, 223–226 (2014).

91. P. R. Baker, R. J. Chalkley, MS-Viewer: A Web-based Spectral Viewer for Proteomics Results. Mol. Cell. Proteomics. 13, 1392–1396 (2014).

92. Y. Watanabe, K. F. Aoki-Kinoshita, Y. Ishihama, S. Okuda, GlycoPOST realizes FAIR principles for glycomics mass spectrometry data. Nucleic Acids Res. 49, D1523–D1528 (2021).

5. Apêndices

5.1. Apêndice A

Tabela S1. Ranqueamento dos modelos treinados com dados unificados de LMA sem otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
Gradient Boosting	0.76	0.46	0.44	0.45	0.53	0.31	0.53
LightGBM	0.74	0.43	0.41	0.42	0.57	0.28	0.57
Decision Tree	0.66	0.43	0.43	0.43	0.57	0.28	0.57
XGBClassifier	0.75	0.42	0.41	0.41	0.57	0.28	0.57
CatBoost	0.75	0.42	0.4	0.41	0.57	0.28	0.57
Random Forest	0.73	0.42	0.4	0.4	0.57	0.27	0.57
k-nearest neighbors	0.69	0.37	0.37	0.37	0.63	0.23	0.63
Multi-layer Perceptron2	0.7	0.28	0.23	0.24	0.71	0.14	0.71
Gaussian Process	0.66	0.28	0.24	0.24	0.71	0.15	0.71
Multi-layer Perceptron	0.7	0.27	0.22	0.22	0.73	0.13	0.73
Support Vector	0.65	0.27	0.22	0.22	0.72	0.13	0.72
Logistic Regression	0.69	0.26	0.2	0.2	0.74	0.12	0.74
Gaussian Naive Bayes	0.67	0.25	0.21	0.21	0.74	0.13	0.74
Quadratic Discriminant	0.7	0.24	0.19	0.2	0.76	0.12	0.76
Stochastic Gradient Descent	0.55	0.24	0.18	0.18	0.76	0.1	0.76
Multinomial Naive Bayes	0.63	0.19	0.1	0.1	0.8	0.06	0.8
AdaBoost	0.57	0.18	0.16	0.16	0.82	0.09	0.82

Tabela S2. Ranqueamento dos modelos treinados com dados unificados de LMA após otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
k-nearest neighbors_opt	0.74	0.44	0.45	0.45	0.56	0.3	0.56
XGBClassifier_opt	0.75	0.4	0.38	0.39	0.59	0.26	0.59
Multi-layer Perceptron_opt	0.7	0.39	0.37	0.38	0.6	0.25	0.6
CatBoost_opt	0.73	0.37	0.35	0.36	0.62	0.23	0.62
LightGBM_opt	0.74	0.36	0.33	0.34	0.63	0.21	0.63
Random Forest_opt	0.66	0.35	0.34	0.33	0.66	0.2	0.66
Multi-layer Perceptron2_opt	0.68	0.32	0.29	0.29	0.68	0.18	0.68
Stochastic Gradient Descent_opt	0.67	0.31	0.29	0.29	0.69	0.17	0.69
Gradient Boosting_opt	0.67	0.29	0.24	0.24	0.71	0.14	0.71
Gaussian Process_opt	0.66	0.29	0.25	0.25	0.71	0.15	0.71
Quadratic Discriminant_opt	0.68	0.28	0.23	0.24	0.72	0.15	0.72
Decision Tree_opt	0.64	0.28	0.22	0.22	0.72	0.13	0.72
Support Vector_opt	0.68	0.27	0.21	0.22	0.72	0.13	0.72
Logistic Regression_opt	0.69	0.25	0.2	0.2	0.74	0.12	0.74

Gaussian Naive Bayes_opt	0.67	0.25	0.21	0.21	0.74	0.13	0.74
AdaBoost_opt	0.63	0.24	0.18	0.19	0.76	0.11	0.76
Multinomial Naive Bayes_opt	0.63	0.19	0.1	0.1	0.8	0.06	0.8

Tabela S3. Ranqueamento dos modelos treinados com dados de LMA do TCGA sem otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
Multi-layer Perceptron	0.81	0.55	0.54	0.54	0.46	0.39	0.46
Logistic Regression	0.78	0.5	0.51	0.51	0.5	0.39	0.5
XGBClassifier	0.78	0.49	0.49	0.49	0.52	0.35	0.52
Gradient Boosting	0.8	0.48	0.45	0.45	0.52	0.32	0.52
LightGBM	0.81	0.47	0.46	0.46	0.54	0.32	0.54
CatBoost	0.77	0.47	0.48	0.48	0.54	0.34	0.54
Random Forest	0.77	0.45	0.45	0.45	0.56	0.31	0.56
Multi-layer Perceptron2	0.8	0.42	0.44	0.43	0.58	0.3	0.58
k-nearest neighbors	0.76	0.42	0.44	0.44	0.58	0.31	0.58
Stochastic Gradient Descent	0.6	0.42	0.44	0.44	0.58	0.32	0.58
Decision Tree	0.63	0.38	0.38	0.38	0.62	0.24	0.62
Support Vector	0.74	0.36	0.39	0.39	0.64	0.25	0.64
Gaussian Process	0.74	0.36	0.37	0.37	0.64	0.24	0.64
Multinomial Naive Bayes	0.74	0.26	0.23	0.23	0.74	0.14	0.74
Gaussian Naive Bayes	0.73	0.24	0.19	0.19	0.76	0.11	0.76
AdaBoost	0.63	0.16	0.11	0.11	0.84	0.06	0.84
Quadratic Discriminant	0.51	0.16	0.12	0.12	0.84	0.08	0.84

Tabela S4. Ranqueamento dos modelos treinados com dados de LMA do TCGA após otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
Support Vector_opt	0.78	0.57	0.57	0.56	0.43	0.43	0.43
Stochastic Gradient Descent_opt	0.75	0.53	0.52	0.52	0.47	0.37	0.47
CatBoost_opt	0.73	0.52	0.52	0.52	0.48	0.36	0.48
XGBClassifier_opt	0.79	0.5	0.53	0.53	0.5	0.38	0.5
LightGBM_opt	0.77	0.47	0.47	0.47	0.54	0.31	0.54
k-nearest neighbors_opt	0.78	0.44	0.47	0.47	0.56	0.32	0.56
AdaBoost_opt	0.76	0.44	0.43	0.43	0.56	0.31	0.56
Quadratic Discriminant_opt	0.69	0.43	0.44	0.44	0.57	0.35	0.57
Logistic Regression_opt	0.78	0.42	0.44	0.44	0.58	0.29	0.58
Gaussian Process_opt	0.73	0.42	0.42	0.42	0.58	0.28	0.58
Multi-layer Perceptron2_opt	0.79	0.4	0.42	0.41	0.6	0.27	0.6
Random Forest_opt	0.75	0.39	0.37	0.37	0.61	0.25	0.61
Gradient Boosting_opt	0.77	0.36	0.36	0.36	0.64	0.24	0.64

Gaussian Naive Bayes_opt	0.73	0.34	0.33	0.33	0.66	0.22	0.66
Multi-layer Perceptron_opt	0.77	0.31	0.3	0.3	0.69	0.19	0.69
Multinomial Naive Bayes_opt	0.77	0.26	0.23	0.23	0.74	0.14	0.74
Decision Tree_opt	0.53	0.14	0.09	0.1	0.86	0.05	0.86

Tabela S5. Ranqueamento dos modelos treinados com dados de LMA do Hemocentro sem otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
AdaBoost	0.76	0.67	0.58	0.53	0.38	0.47	0.38
Random Forest	0.8	0.44	0.34	0.34	0.59	0.22	0.59
Gaussian Process	0.76	0.42	0.36	0.36	0.59	0.24	0.59
Stochastic Gradient Descent	0.64	0.42	0.37	0.4	0.57	0.28	0.57
Decision Tree	0.62	0.42	0.37	0.39	0.59	0.27	0.59
CatBoost	0.76	0.4	0.32	0.33	0.62	0.21	0.62
Gaussian Naive Bayes	0.57	0.38	0.3	0.33	0.59	0.24	0.59
Logistic Regression	0.57	0.37	0.33	0.36	0.61	0.24	0.61
k-nearest neighbors	0.62	0.33	0.3	0.31	0.65	0.19	0.65
Quadratic Discriminant	0.58	0.33	0.22	0.2	0.67	0.13	0.67
XGBClassifier	0.78	0.31	0.25	0.26	0.7	0.15	0.7
LightGBM	0.71	0.31	0.26	0.27	0.68	0.17	0.68
Multi-layer Perceptron2	0.53	0.31	0.26	0.27	0.68	0.17	0.68
Multi-layer Perceptron	0.57	0.27	0.23	0.24	0.73	0.15	0.73
Multinomial Naive Bayes	0.76	0.25	0.13	0.13	0.74	0.09	0.74
Support Vector	0.24	0.25	0.1	0.11	0.74	0.07	0.74
Gradient Boosting	0.72	0.23	0.2	0.21	0.77	0.12	0.77

Tabela S6. Ranqueamento dos modelos treinados com dados de LMA do Hemocentro após otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
AdaBoost_opt	0.73	0.52	0.45	0.46	0.47	0.36	0.47
LightGBM_opt	0.72	0.35	0.29	0.31	0.63	0.25	0.63
Support Vector_opt	0.54	0.35	0.32	0.32	0.66	0.21	0.66
Stochastic Gradient Descent_opt	0.49	0.35	0.32	0.31	0.66	0.2	0.66
Multinomial Naive Bayes_opt	0.64	0.32	0.27	0.26	0.66	0.18	0.66
Multi-layer Perceptron2_opt	0.65	0.3	0.27	0.26	0.71	0.16	0.71
Decision Tree_opt	0.64	0.3	0.25	0.25	0.69	0.2	0.69
XGBClassifier_opt	0.58	0.3	0.27	0.26	0.71	0.16	0.71
Logistic Regression_opt	0.54	0.3	0.3	0.27	0.71	0.18	0.71
CatBoost_opt	0.6	0.28	0.21	0.23	0.7	0.15	0.7
k-nearest neighbors_opt	0.66	0.25	0.14	0.17	0.7	0.12	0.7
Random Forest_opt	0.61	0.22	0.18	0.18	0.78	0.11	0.78

Gradient Boosting_opt	0.73	0.2	0.11	0.13	0.76	0.08	0.76
Multi-layer Perceptron_opt	0.56	0.18	0.19	0.17	0.84	0.11	0.84
Quadratic Discriminant_opt	0.34	0.12	0.17	0.17	0.87	0.13	0.87
Gaussian Naive Bayes_opt	0.5	0.1	0.11	0.14	0.88	0.09	0.88
Gaussian Process_opt	0.48	0.1	0.11	0.13	0.88	0.08	0.88

Tabela S7. Ranqueamento dos modelos treinados com dados de LMA do TARGET sem otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
Random Forest	0.82	0.48	0.43	0.44	0.52	0.3	0.52
Quadratic Discriminant	0.79	0.47	0.46	0.46	0.53	0.32	0.53
CatBoost	0.82	0.46	0.43	0.44	0.53	0.3	0.53
XGBClassifier	0.81	0.45	0.42	0.42	0.54	0.29	0.54
Gaussian Naive Bayes	0.78	0.45	0.42	0.42	0.54	0.3	0.54
LightGBM	0.8	0.44	0.41	0.42	0.55	0.28	0.55
k-nearest neighbors	0.76	0.44	0.45	0.46	0.55	0.31	0.55
Gradient Boosting	0.8	0.43	0.4	0.41	0.56	0.27	0.56
Decision Tree	0.66	0.41	0.38	0.39	0.58	0.25	0.58
Multi-layer Perceptron	0.81	0.32	0.26	0.27	0.68	0.17	0.68
Support Vector	0.8	0.31	0.26	0.27	0.68	0.17	0.68
Gaussian Process	0.78	0.31	0.27	0.28	0.68	0.17	0.68
AdaBoost	0.64	0.29	0.23	0.24	0.7	0.15	0.7
Logistic Regression	0.8	0.28	0.22	0.23	0.72	0.14	0.72
Multi-layer Perceptron2	0.79	0.28	0.22	0.23	0.71	0.14	0.71
Stochastic Gradient Descent	0.57	0.24	0.19	0.19	0.76	0.12	0.76
Multinomial Naive Bayes	0.67	0.16	0.08	0.08	0.84	0.05	0.84

Tabela S8. Ranqueamento dos modelos treinados com dados de LMA do TARGET após otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
CatBoost_opt	0.84	0.48	0.45	0.46	0.51	0.32	0.51
k-nearest neighbors_opt	0.84	0.45	0.43	0.44	0.54	0.31	0.54
XGBClassifier_opt	0.8	0.45	0.41	0.42	0.54	0.28	0.54
Gaussian Naive Bayes_opt	0.78	0.45	0.42	0.42	0.54	0.3	0.54
LightGBM_opt	0.79	0.43	0.4	0.41	0.56	0.27	0.56
Multi-layer Perceptron_opt	0.78	0.38	0.35	0.36	0.61	0.23	0.61
Gaussian Process_opt	0.77	0.34	0.31	0.31	0.65	0.21	0.65
Quadratic Discriminant_opt	0.81	0.31	0.25	0.26	0.68	0.16	0.68
Support Vector_opt	0.8	0.3	0.24	0.25	0.69	0.15	0.69
Gradient Boosting_opt	0.79	0.3	0.22	0.22	0.7	0.14	0.7
Stochastic Gradient Descent_opt	0.78	0.3	0.25	0.25	0.7	0.16	0.7

Multi-layer Perceptron2_opt	0.77	0.3	0.24	0.25	0.69	0.15	0.69
AdaBoost_opt	0.7	0.29	0.22	0.22	0.71	0.14	0.71
Logistic Regression_opt	0.77	0.27	0.22	0.22	0.72	0.13	0.72
Decision Tree_opt	0.7	0.25	0.18	0.18	0.75	0.11	0.75
Random Forest_opt	0.73	0.23	0.14	0.15	0.76	0.09	0.76
Multinomial Naive Bayes_opt	0.69	0.16	0.08	0.08	0.84	0.05	0.84

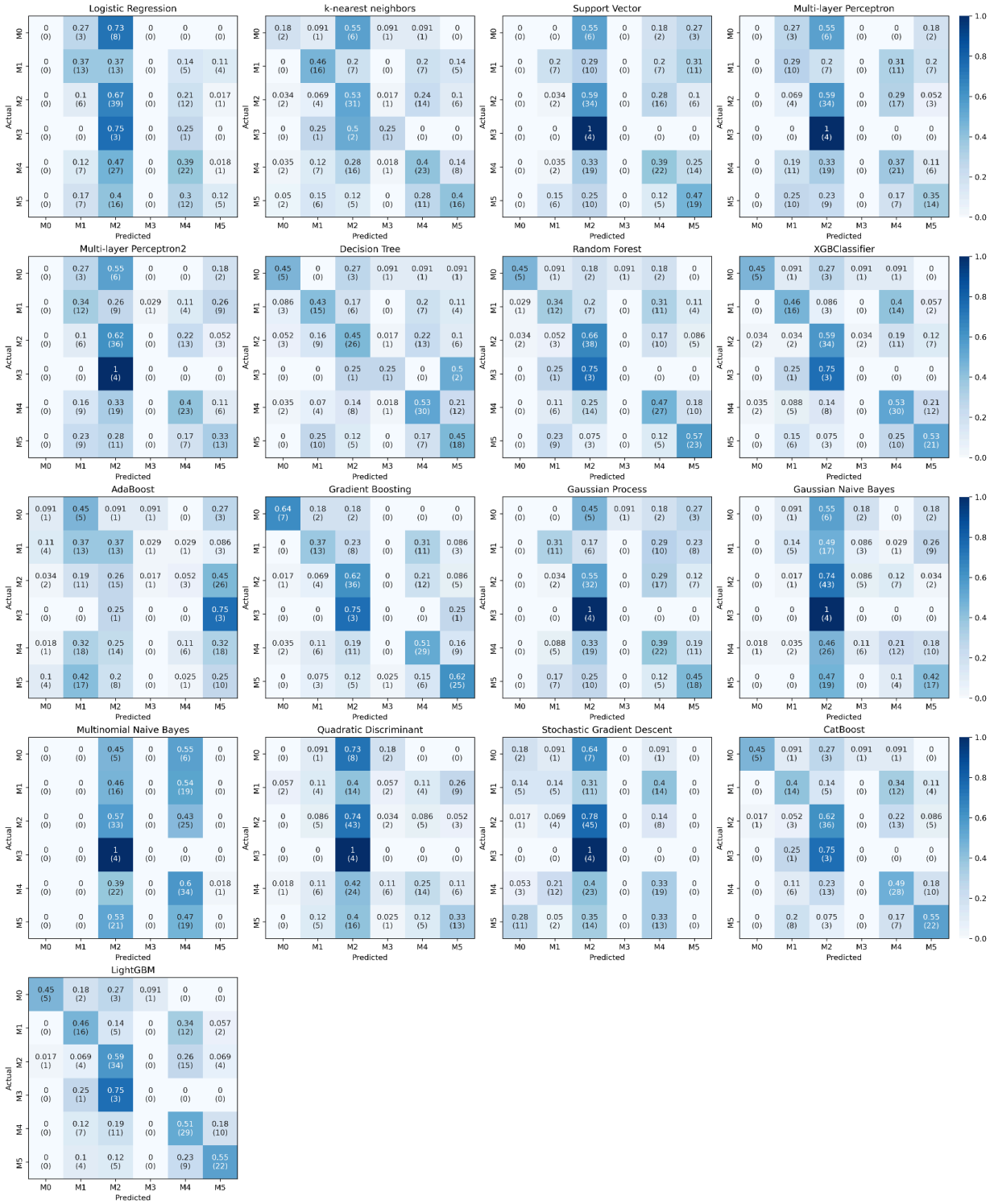

Figura S1. ***Matriz de confusão dos modelos não-otimizados treinados com o conjunto LMA unificado.***

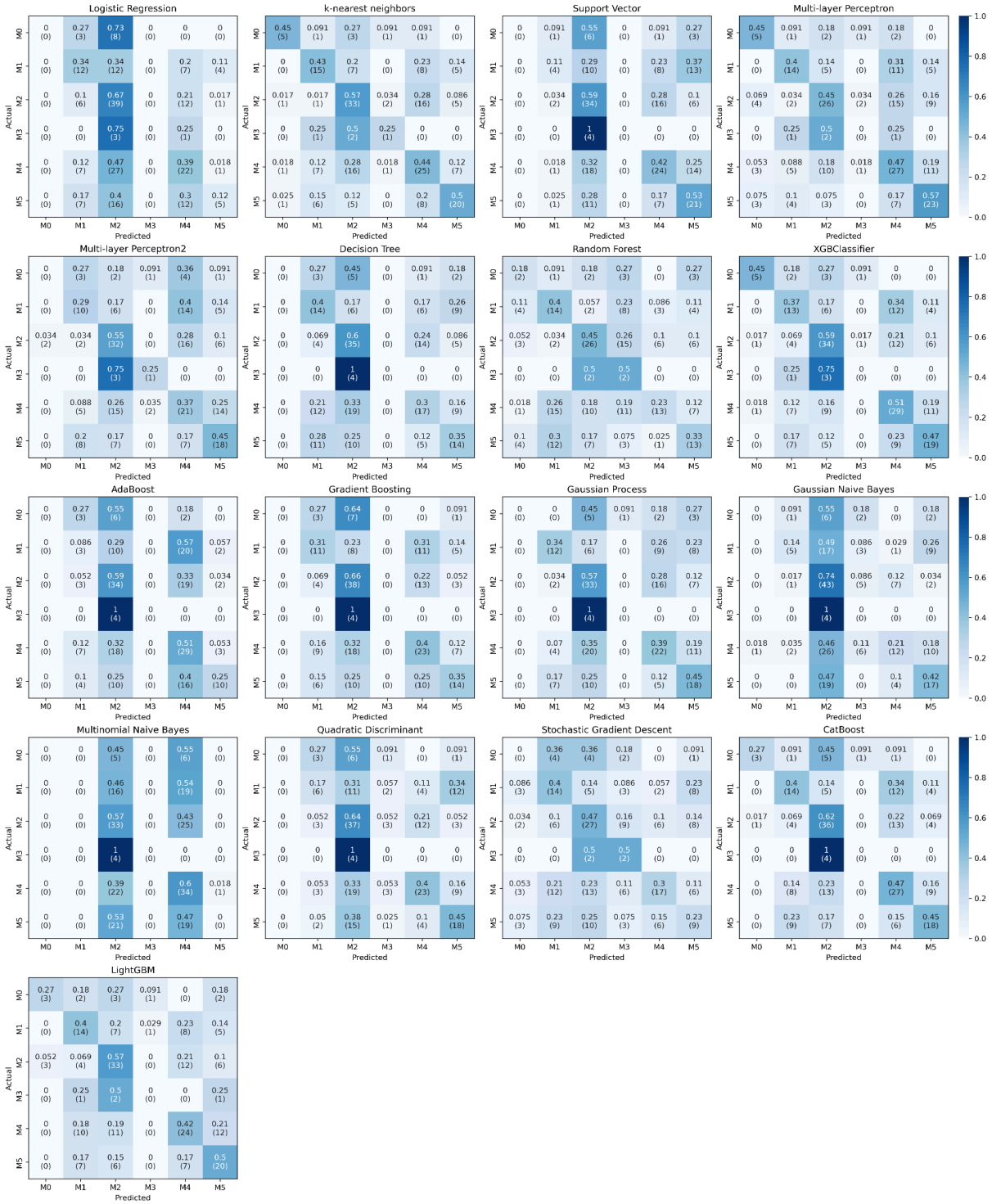

Figura S2. ***Matriz de confusão dos modelos otimizados treinados com o conjunto LMA unificado.***

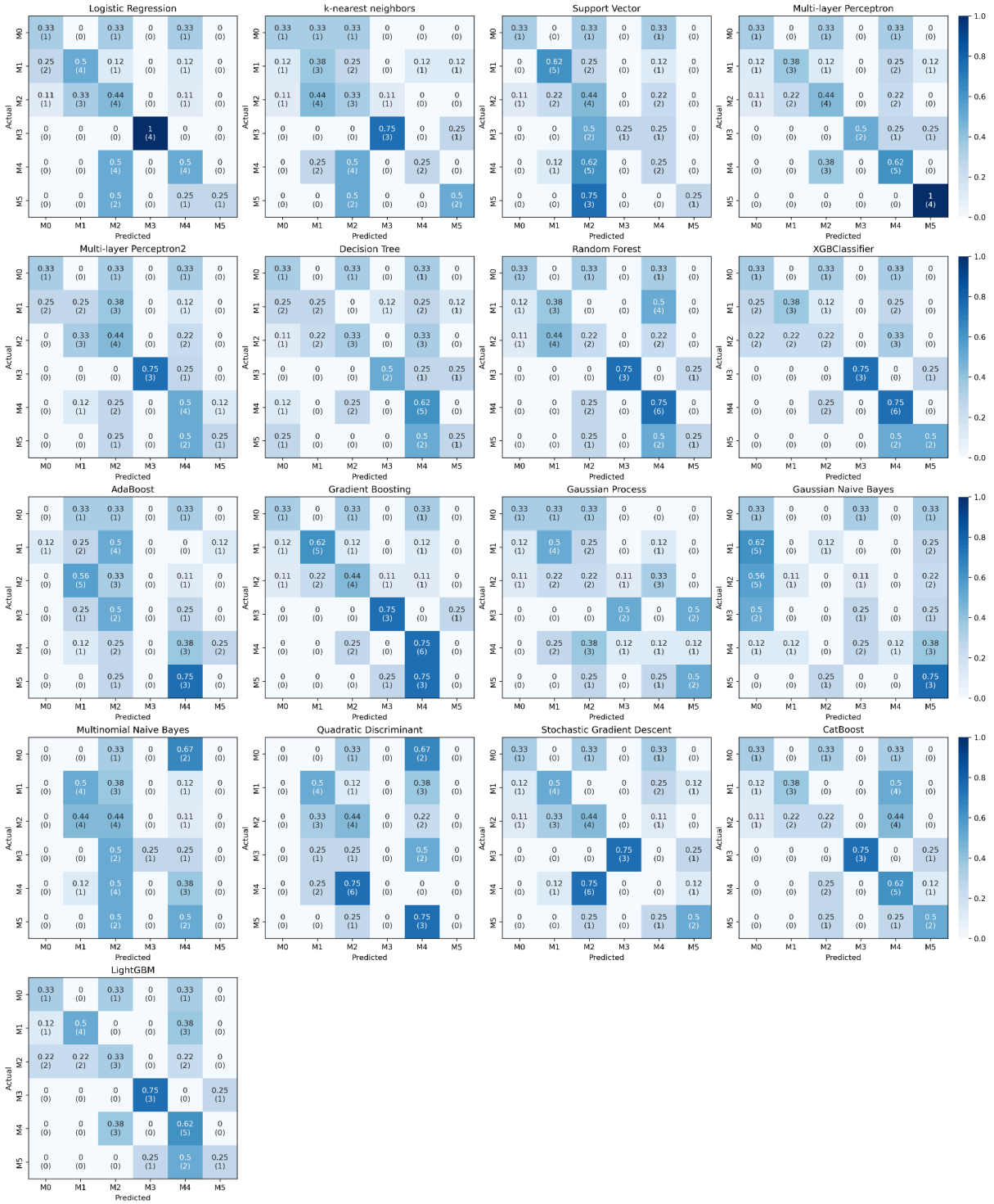

Figura S3. ***Matriz de confusão dos modelos não-otimizados treinados com o conjunto LMA do TCGA.***

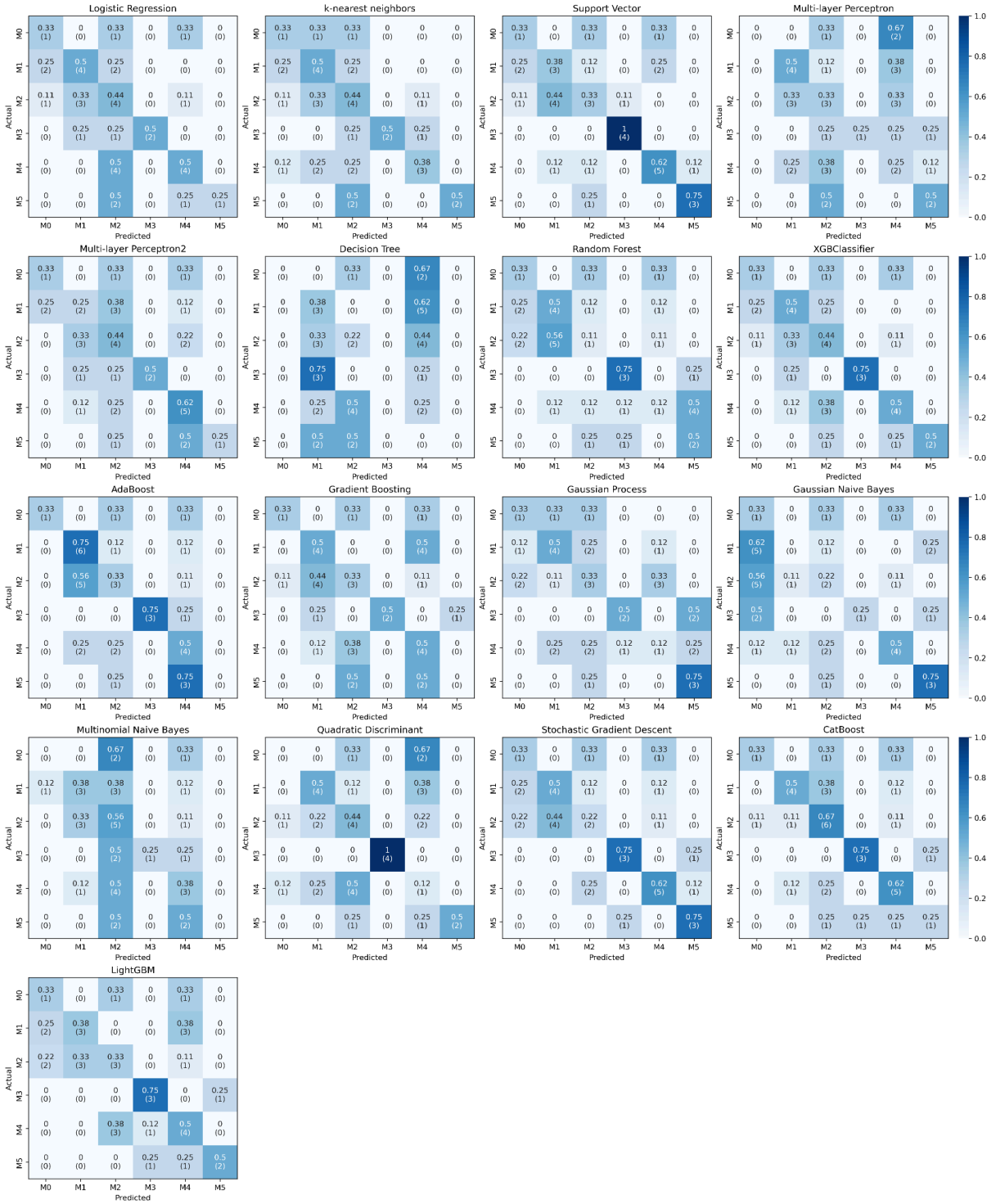

Figura S4. ***Matriz de confusão dos modelos otimizados treinados com o conjunto LMA do TCGA.***

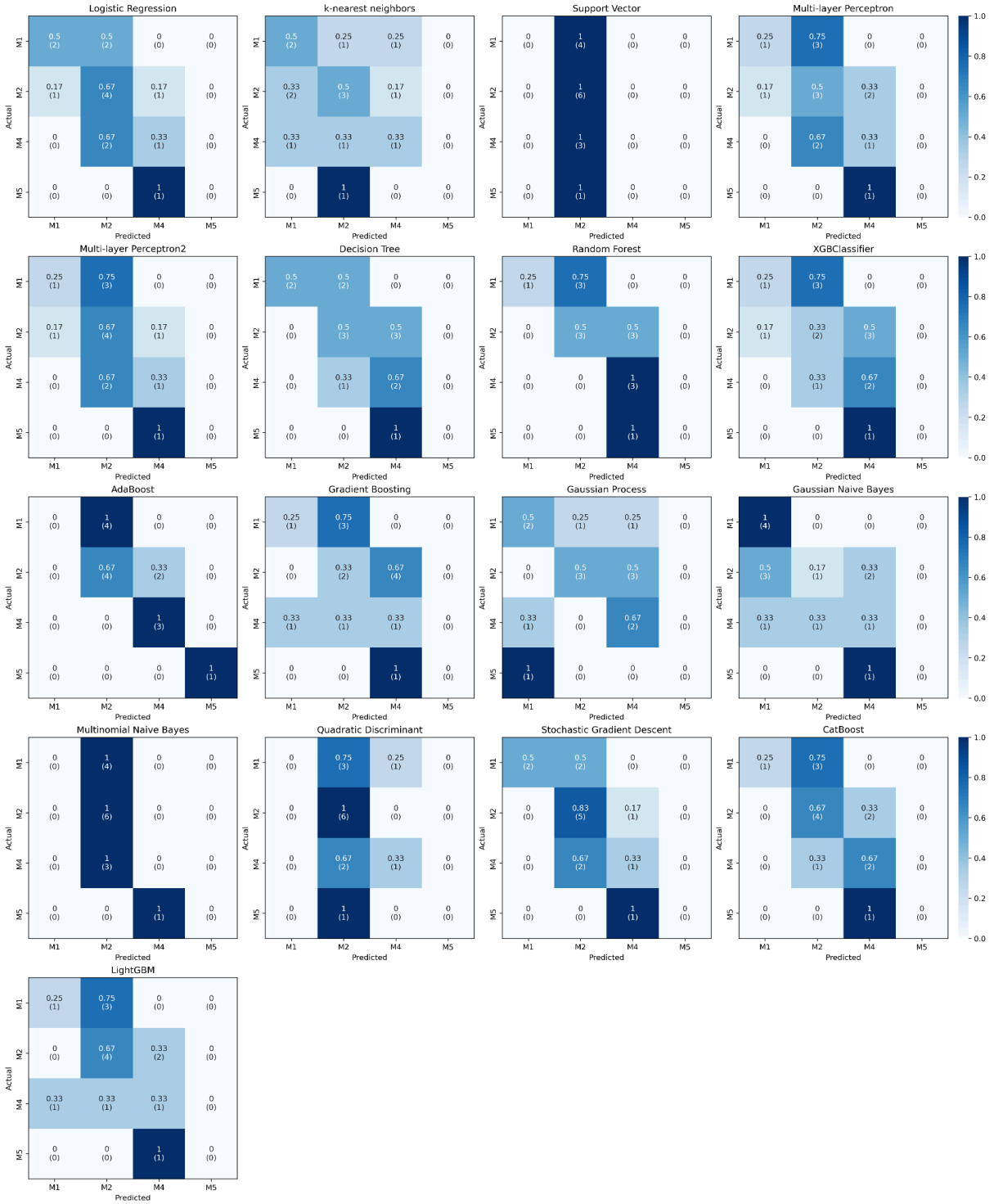

Figura S5. ***Matriz de confusão dos modelos não-otimizados treinados com o conjunto LMA do Hemocentro.***

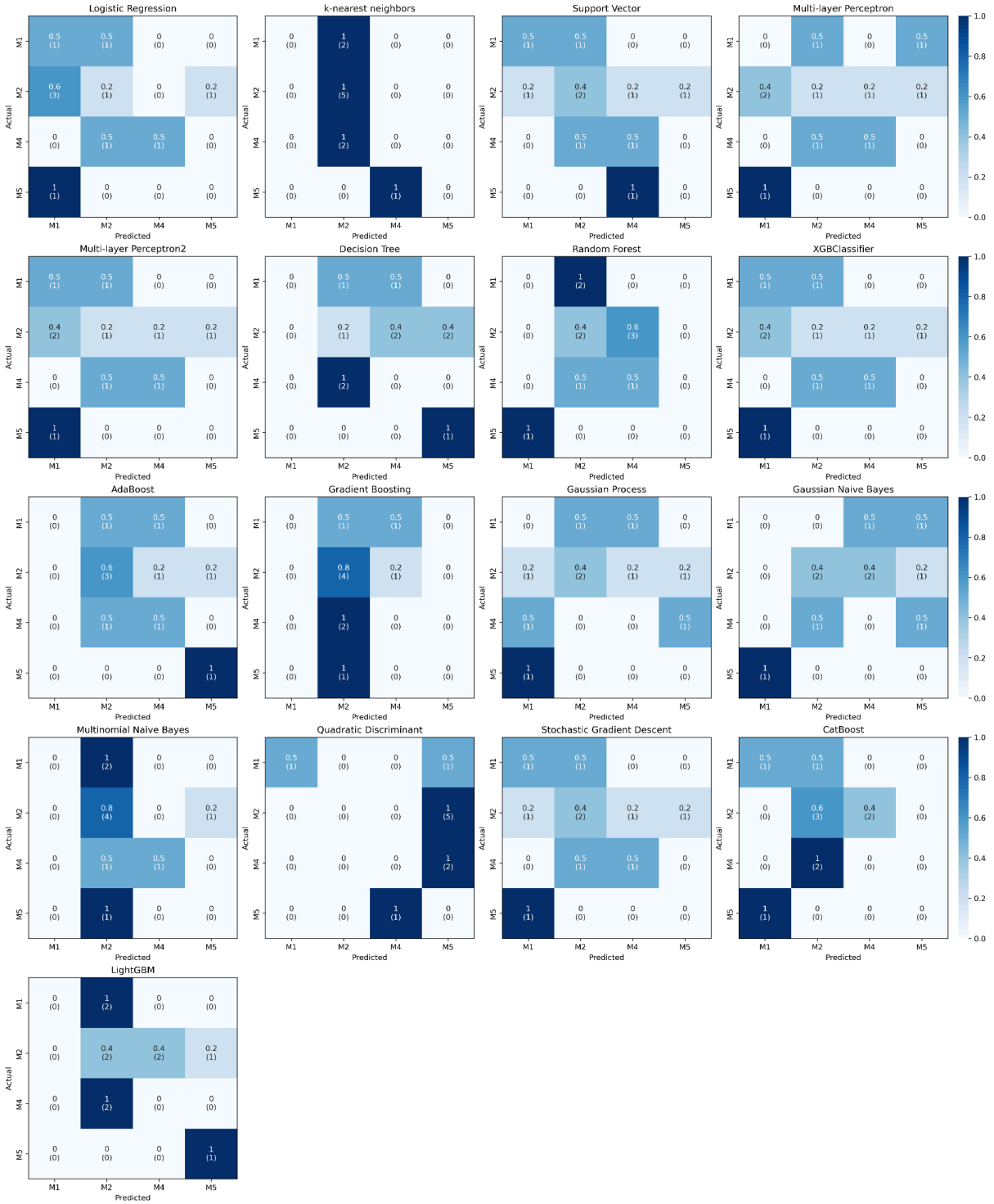

Figura S6. ***Matriz de confusão dos modelos otimizados treinados com o conjunto LMA do Hemocentro.***

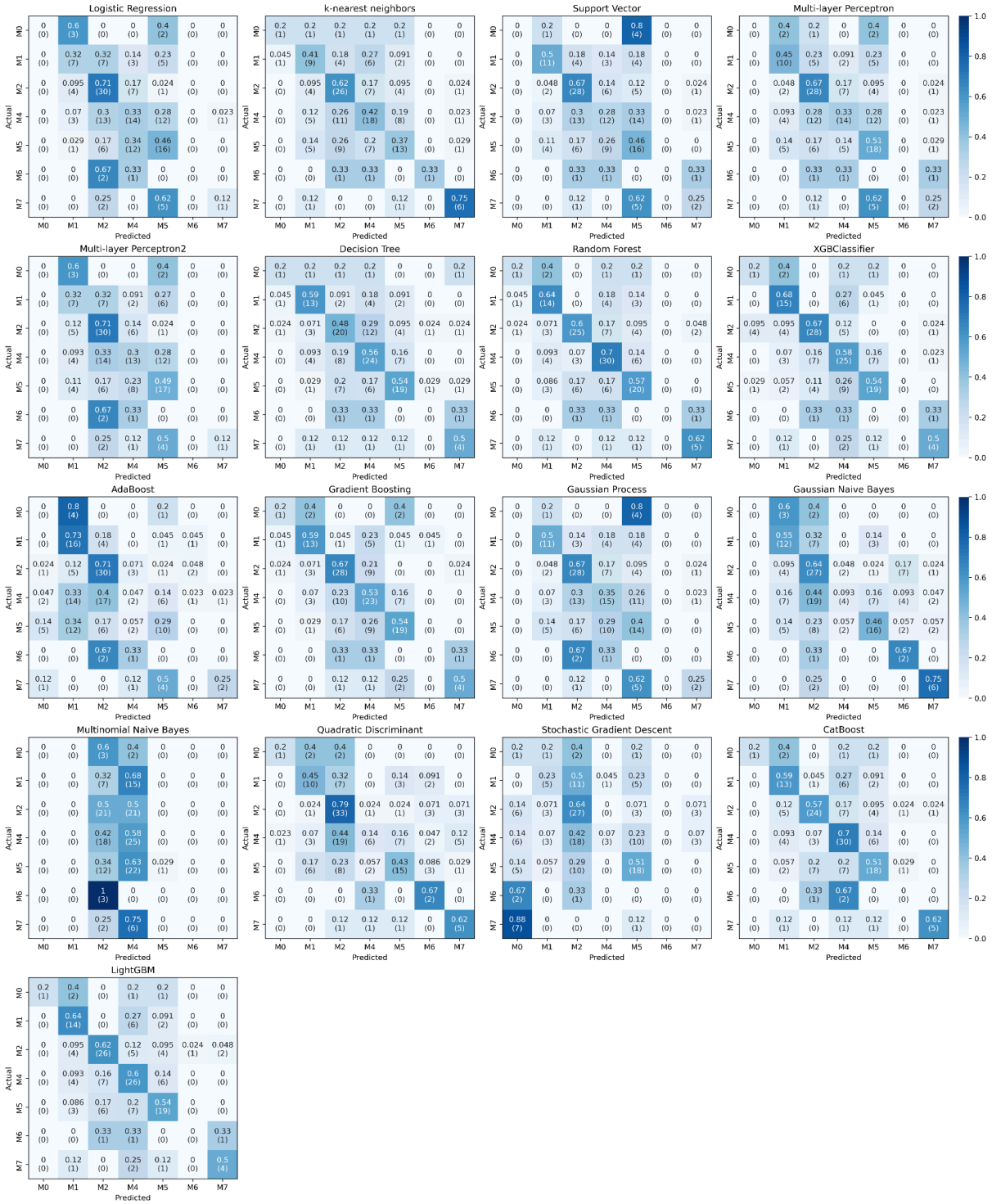

Figura S7. ***Matriz de confusão dos modelos não-otimizados treinados com o conjunto LMA do TARGET.***

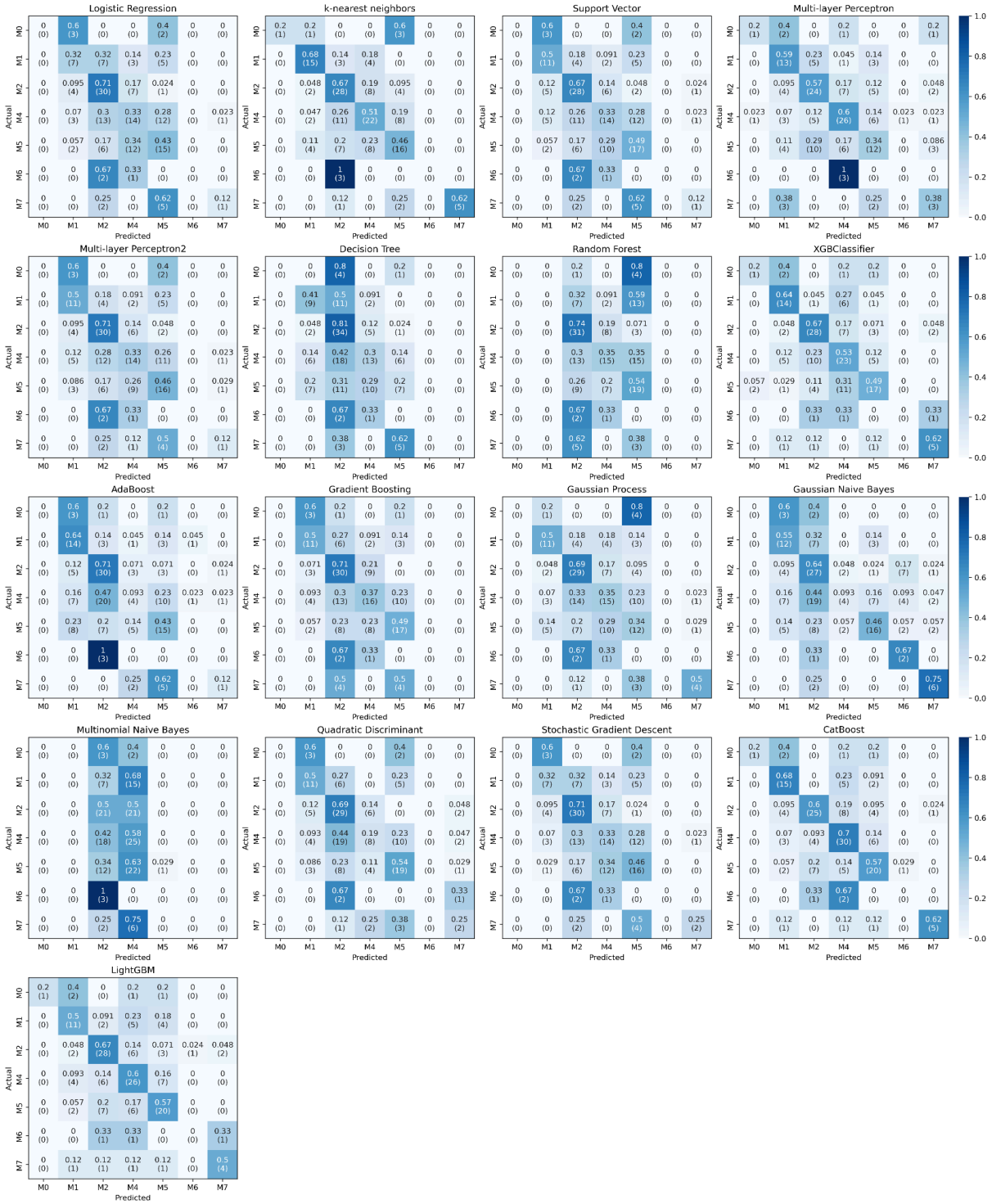

Figura S8. ***Matriz de confusão dos modelos otimizados treinados com o conjunto LMA do TARGET.***

5.2. Apêndice B

Tabela S9. Ranqueamento dos modelos treinados com dados unificados de SMD sem otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
LightGBM	0.84	0.53	0.53	0.53	0.47	0.42	0.47
Random Forest	0.84	0.53	0.51	0.51	0.46	0.41	0.46
Gaussian Naive Bayes	0.84	0.52	0.47	0.47	0.48	0.39	0.48
Multi-layer Perceptron	0.85	0.51	0.47	0.47	0.49	0.38	0.49
CatBoost	0.84	0.51	0.49	0.49	0.49	0.4	0.49
Gaussian Process	0.83	0.51	0.46	0.46	0.49	0.36	0.49
Quadratic Discriminant	0.83	0.51	0.45	0.46	0.49	0.37	0.49
Multi-layer Perceptron2	0.85	0.5	0.45	0.45	0.5	0.37	0.5
Gradient Boosting	0.84	0.5	0.48	0.48	0.5	0.39	0.5
XGBClassifier	0.83	0.5	0.49	0.49	0.5	0.39	0.5
Support Vector	0.84	0.49	0.45	0.45	0.51	0.36	0.51
Decision Tree	0.7	0.49	0.48	0.48	0.51	0.38	0.51
k-nearest neighbors	0.77	0.46	0.43	0.43	0.54	0.33	0.54
Logistic Regression	0.83	0.43	0.36	0.37	0.57	0.28	0.57
Stochastic Gradient Descent	0.67	0.39	0.32	0.32	0.61	0.25	0.61
AdaBoost	0.75	0.34	0.29	0.29	0.66	0.18	0.66
Multinomial Naive Bayes	0.79	0.25	0.16	0.16	0.75	0.13	0.75

Tabela S10. Ranqueamento dos modelos treinados com dados unificados de SMD após otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
Multi-layer Perceptron2_opt	0.85	0.53	0.48	0.48	0.47	0.39	0.47
Support Vector_opt	0.84	0.53	0.5	0.5	0.47	0.4	0.47
CatBoost_opt	0.85	0.52	0.49	0.49	0.48	0.4	0.48
XGBClassifier_opt	0.84	0.52	0.51	0.51	0.48	0.41	0.48
Gaussian Naive Bayes_opt	0.84	0.52	0.47	0.47	0.48	0.39	0.48
Multi-layer Perceptron_opt	0.84	0.51	0.49	0.49	0.49	0.39	0.49
LightGBM_opt	0.83	0.51	0.49	0.49	0.49	0.4	0.49
Gaussian Process_opt	0.83	0.51	0.46	0.46	0.49	0.36	0.49
Quadratic Discriminant_opt	0.83	0.51	0.45	0.45	0.49	0.37	0.49
k-nearest neighbors_opt	0.83	0.5	0.47	0.47	0.5	0.37	0.5
Logistic Regression_opt	0.83	0.43	0.36	0.36	0.57	0.27	0.57
Stochastic Gradient Descent_opt	0.81	0.41	0.35	0.35	0.59	0.26	0.59
AdaBoost_opt	0.66	0.41	0.32	0.32	0.59	0.29	0.59
Random Forest_opt	0.81	0.3	0.18	0.18	0.7	0.12	0.7
Decision Tree_opt	0.75	0.29	0.19	0.19	0.71	0.13	0.71

Multinomial Naive Bayes_opt	0.8	0.26	0.16	0.16	0.74	0.13	0.74
Gradient Boosting_opt	0.5	0.14	0.04	0.04	0.86	0.02	0.86

Tabela S11. Ranqueamento dos modelos treinados com dados de SMD do Hemocentro sem otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
Gaussian Naive Bayes	0.97	0.95	0.95	0.96	0.04	0.92	0.04
Support Vector	0.95	0.95	0.95	0.96	0.04	0.92	0.04
Random Forest	0.94	0.9	0.9	0.91	0.09	0.85	0.09
XGBClassifier	0.92	0.9	0.9	0.91	0.09	0.85	0.09
LightGBM	0.97	0.89	0.89	0.9	0.1	0.83	0.1
AdaBoost	0.96	0.89	0.89	0.9	0.1	0.83	0.1
Gradient Boosting	0.95	0.86	0.85	0.86	0.13	0.78	0.13
CatBoost	0.94	0.86	0.85	0.86	0.13	0.78	0.13
Multi-layer Perceptron	0.94	0.84	0.84	0.85	0.14	0.77	0.14
Logistic Regression	0.94	0.79	0.79	0.81	0.19	0.71	0.19
Multi-layer Perceptron2	0.94	0.79	0.76	0.75	0.25	0.61	0.25
Gaussian Process	0.92	0.79	0.76	0.75	0.25	0.61	0.25
Decision Tree	0.87	0.79	0.79	0.81	0.19	0.71	0.19
k-nearest neighbors	0.9	0.69	0.67	0.66	0.33	0.51	0.33
Stochastic Gradient Descent	0.87	0.68	0.67	0.71	0.29	0.61	0.29
Quadratic Discriminant	0.91	0.65	0.65	0.65	0.36	0.48	0.36
Multinomial Naive Bayes	0.97	0.33	0.16	0.14	0.7	0.09	0.7

Tabela S12. Ranqueamento dos modelos treinados com dados de SMD do Hemocentro após otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
Gaussian Naive Bayes_opt	0.97	0.95	0.95	0.96	0.04	0.92	0.04
Quadratic Discriminant_opt	0.96	0.95	0.95	0.96	0.04	0.92	0.04
Multi-layer Perceptron2_opt	0.95	0.95	0.95	0.96	0.04	0.92	0.04
Support Vector_opt	0.95	0.9	0.9	0.91	0.09	0.84	0.09
Random Forest_opt	0.94	0.9	0.9	0.91	0.09	0.85	0.09
k-nearest neighbors_opt	0.93	0.9	0.9	0.91	0.09	0.85	0.09
Decision Tree_opt	0.92	0.9	0.9	0.91	0.09	0.85	0.09
AdaBoost_opt	0.96	0.89	0.89	0.9	0.1	0.83	0.1
LightGBM_opt	0.96	0.89	0.89	0.9	0.1	0.83	0.1
CatBoost_opt	0.94	0.89	0.89	0.9	0.1	0.83	0.1
Stochastic Gradient Descent_opt	0.96	0.86	0.85	0.86	0.13	0.78	0.13
XGBClassifier_opt	0.93	0.86	0.85	0.86	0.13	0.78	0.13
Logistic Regression_opt	0.94	0.84	0.84	0.85	0.14	0.77	0.14
Multi-layer Perceptron_opt	0.95	0.79	0.79	0.81	0.19	0.71	0.19
Gaussian Process_opt	0.92	0.79	0.76	0.75	0.25	0.61	0.25

Gradient Boosting_opt	0.89	0.55	0.44	0.39	0.51	0.31	0.51
Multinomial Naive Bayes_opt	0.97	0.4	0.27	0.25	0.64	0.16	0.64

Tabela S13. Ranqueamento dos modelos treinados com dados de SMD da UFC sem otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
Decision Tree	0.74	0.57	0.57	0.57	0.42	0.46	0.42
Multi-layer Perceptron	0.87	0.56	0.48	0.45	0.47	0.38	0.47
Support Vector	0.77	0.52	0.44	0.4	0.51	0.32	0.51
XGBClassifier	0.81	0.51	0.48	0.45	0.53	0.4	0.53
Gradient Boosting	0.79	0.51	0.5	0.47	0.52	0.41	0.52
Quadratic Discriminant	0.71	0.51	0.48	0.45	0.52	0.39	0.52
Random Forest	0.77	0.43	0.41	0.39	0.59	0.29	0.59
Logistic Regression	0.76	0.43	0.36	0.32	0.61	0.25	0.61
Gaussian Naive Bayes	0.75	0.43	0.4	0.38	0.59	0.31	0.59
CatBoost	0.75	0.43	0.41	0.38	0.59	0.29	0.59
Gaussian Process	0.79	0.42	0.39	0.36	0.61	0.28	0.61
k-nearest neighbors	0.71	0.42	0.38	0.35	0.61	0.27	0.61
LightGBM	0.77	0.39	0.4	0.37	0.65	0.31	0.65
Multi-layer Perceptron2	0.75	0.39	0.31	0.28	0.65	0.2	0.65
Stochastic Gradient Descent	0.65	0.39	0.33	0.3	0.65	0.23	0.65
AdaBoost	0.71	0.34	0.24	0.25	0.66	0.17	0.66
Multinomial Naive Bayes	0.64	0.17	0.05	0.05	0.84	0.03	0.84

Tabela S14. Ranqueamento dos modelos treinados com dados de SMD da UFC após otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
Multi-layer Perceptron_opt	0.81	0.6	0.6	0.57	0.43	0.49	0.43
Multi-layer Perceptron2_opt	0.81	0.59	0.49	0.45	0.45	0.37	0.45
Random Forest_opt	0.79	0.57	0.49	0.45	0.46	0.4	0.46
Gradient Boosting_opt	0.81	0.55	0.53	0.5	0.48	0.43	0.48
Support Vector_opt	0.73	0.55	0.48	0.45	0.48	0.39	0.48
CatBoost_opt	0.81	0.53	0.49	0.46	0.5	0.41	0.5
LightGBM_opt	0.79	0.53	0.5	0.47	0.5	0.4	0.5
Quadratic Discriminant_opt	0.71	0.51	0.48	0.45	0.52	0.39	0.52
Stochastic Gradient Descent_opt	0.75	0.46	0.37	0.34	0.57	0.26	0.57
XGBClassifier_opt	0.81	0.45	0.42	0.39	0.59	0.35	0.59
Gaussian Naive Bayes_opt	0.75	0.43	0.4	0.38	0.59	0.31	0.59
AdaBoost_opt	0.75	0.4	0.27	0.27	0.59	0.2	0.59
Gaussian Process_opt	0.71	0.4	0.37	0.34	0.64	0.27	0.64

Logistic Regression_opt	0.76	0.34	0.26	0.23	0.7	0.16	0.7
Decision Tree_opt	0.7	0.34	0.27	0.28	0.65	0.19	0.65
k-nearest neighbors_opt	0.67	0.33	0.29	0.27	0.69	0.17	0.69
Multinomial Naive Bayes_opt	0.68	0.23	0.12	0.12	0.77	0.07	0.77

Tabela S15. Ranqueamento dos modelos treinados com dados de SMD do MDS-IWG sem otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
XGBClassifier	0.85	0.55	0.54	0.54	0.45	0.43	0.45
CatBoost	0.85	0.55	0.54	0.54	0.45	0.43	0.45
Random Forest	0.85	0.55	0.53	0.53	0.45	0.43	0.45
LightGBM	0.85	0.55	0.53	0.53	0.46	0.43	0.46
Multi-layer Perceptron	0.85	0.55	0.52	0.52	0.45	0.41	0.45
Gradient Boosting	0.86	0.54	0.52	0.52	0.46	0.42	0.46
Multi-layer Perceptron2	0.85	0.51	0.45	0.46	0.49	0.36	0.49
Gaussian Naive Bayes	0.84	0.51	0.45	0.45	0.49	0.36	0.49
Quadratic Discriminant	0.84	0.5	0.44	0.44	0.5	0.35	0.5
Gaussian Process	0.81	0.5	0.46	0.46	0.5	0.35	0.5
Support Vector	0.84	0.49	0.43	0.43	0.51	0.34	0.51
Decision Tree	0.69	0.47	0.45	0.45	0.53	0.35	0.53
k-nearest neighbors	0.78	0.46	0.44	0.44	0.54	0.32	0.54
Logistic Regression	0.82	0.45	0.38	0.38	0.55	0.29	0.55
Stochastic Gradient Descent	0.61	0.34	0.28	0.28	0.66	0.21	0.66
AdaBoost	0.78	0.33	0.28	0.28	0.67	0.18	0.67
Multinomial Naive Bayes	0.77	0.26	0.17	0.16	0.74	0.13	0.74

Tabela S16. Ranqueamento dos modelos treinados com dados de SMD do MDS-IWG após otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
CatBoost_opt	0.86	0.56	0.52	0.52	0.44	0.42	0.44
XGBClassifier_opt	0.85	0.55	0.52	0.52	0.45	0.42	0.45
LightGBM_opt	0.85	0.54	0.52	0.52	0.46	0.42	0.46
Multi-layer Perceptron2_opt	0.85	0.53	0.47	0.47	0.47	0.39	0.47
Support Vector_opt	0.84	0.52	0.51	0.51	0.48	0.39	0.48
Multi-layer Perceptron_opt	0.84	0.51	0.49	0.49	0.49	0.38	0.49
Gaussian Naive Bayes_opt	0.84	0.51	0.45	0.45	0.49	0.36	0.49
Quadratic Discriminant_opt	0.84	0.5	0.44	0.44	0.5	0.35	0.5
Gaussian Process_opt	0.81	0.5	0.46	0.46	0.5	0.35	0.5
k-nearest neighbors_opt	0.82	0.47	0.44	0.44	0.53	0.33	0.53
Logistic Regression_opt	0.82	0.46	0.4	0.4	0.54	0.31	0.54
Stochastic Gradient Descent_opt	0.81	0.43	0.37	0.37	0.57	0.28	0.57

Decision Tree_opt	0.77	0.42	0.33	0.32	0.58	0.3	0.58
AdaBoost_opt	0.62	0.42	0.33	0.32	0.58	0.3	0.58
Random Forest_opt	0.8	0.3	0.19	0.19	0.7	0.13	0.7
Multinomial Naive Bayes_opt	0.77	0.26	0.17	0.16	0.74	0.13	0.74
Gradient Boosting_opt	0.5	0.14	0.04	0.04	0.86	0.02	0.86

Tabela S17. Ranqueamento dos modelos treinados com dados de SMD para IPSS-R sem otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
Gradient Boosting	0.88	0.61	0.6	0.6	0.39	0.45	0.39
Random Forest	0.88	0.6	0.59	0.59	0.4	0.44	0.4
CatBoost	0.88	0.6	0.59	0.59	0.4	0.44	0.4
LightGBM	0.87	0.58	0.57	0.57	0.42	0.42	0.42
Quadratic Discriminant	0.87	0.58	0.56	0.56	0.42	0.41	0.42
XGBClassifier	0.87	0.57	0.56	0.56	0.43	0.41	0.43
AdaBoost	0.76	0.57	0.51	0.51	0.42	0.39	0.42
Multi-layer Perceptron	0.88	0.56	0.56	0.56	0.44	0.4	0.44
Support Vector	0.87	0.56	0.56	0.56	0.44	0.4	0.44
Gaussian Naive Bayes	0.85	0.56	0.55	0.55	0.44	0.39	0.44
Decision Tree	0.72	0.55	0.55	0.55	0.45	0.4	0.45
k-nearest neighbors	0.83	0.54	0.54	0.54	0.46	0.37	0.46
Multi-layer Perceptron2	0.87	0.53	0.52	0.52	0.47	0.36	0.47
Gaussian Process	0.86	0.53	0.52	0.52	0.47	0.36	0.47
Logistic Regression	0.83	0.45	0.43	0.43	0.55	0.28	0.55
Stochastic Gradient Descent	0.68	0.28	0.21	0.21	0.72	0.12	0.72
Multinomial Naive Bayes	0.77	0.21	0.08	0.08	0.8	0.05	0.8

Tabela S18. Ranqueamento dos modelos treinados com dados de SMD para IPSS-R após otimização de hiperparâmetros.

models	auc_roc	balanced_acc	f1_ma_avg	f1_weighted	hamming	jaccard	loss_01
CatBoost_opt	0.89	0.64	0.62	0.62	0.36	0.48	0.36
LightGBM_opt	0.88	0.62	0.6	0.6	0.38	0.46	0.38
XGBClassifier_opt	0.88	0.61	0.6	0.6	0.39	0.45	0.39
Multi-layer Perceptron_opt	0.87	0.59	0.59	0.59	0.41	0.44	0.41
Support Vector_opt	0.87	0.59	0.58	0.58	0.41	0.43	0.41
AdaBoost_opt	0.8	0.59	0.53	0.53	0.41	0.41	0.41
Quadratic Discriminant_opt	0.87	0.57	0.55	0.55	0.43	0.4	0.43
Multi-layer Perceptron2_opt	0.88	0.56	0.55	0.55	0.44	0.4	0.44
k-nearest neighbors_opt	0.86	0.56	0.55	0.55	0.44	0.39	0.44
Gaussian Naive Bayes_opt	0.85	0.56	0.55	0.55	0.44	0.39	0.44
Gaussian Process_opt	0.86	0.53	0.52	0.52	0.47	0.36	0.47

Logistic Regression_opt	0.83	0.46	0.45	0.45	0.54	0.29	0.54
Stochastic Gradient Descent_opt	0.82	0.46	0.44	0.44	0.54	0.29	0.54
Decision Tree_opt	0.79	0.39	0.29	0.29	0.61	0.19	0.61
Random Forest_opt	0.8	0.27	0.15	0.15	0.73	0.09	0.73
Multinomial Naive Bayes_opt	0.77	0.21	0.09	0.09	0.79	0.05	0.79
Gradient Boosting_opt	0.5	0.2	0.07	0.07	0.8	0.04	0.8

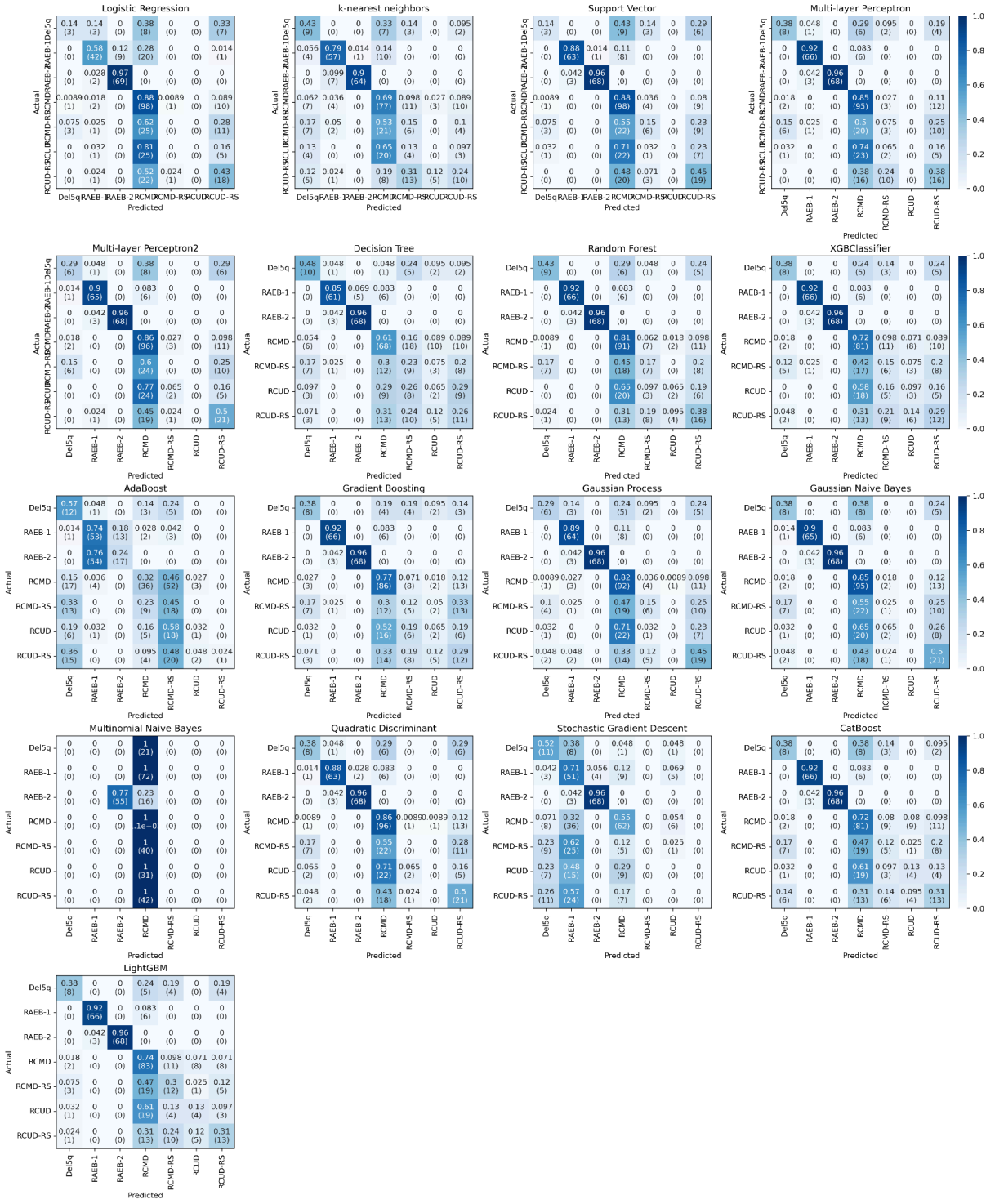

Figura S9. ***Matriz de confusão dos modelos não-otimizados treinados com o conjunto SMD unificado.***

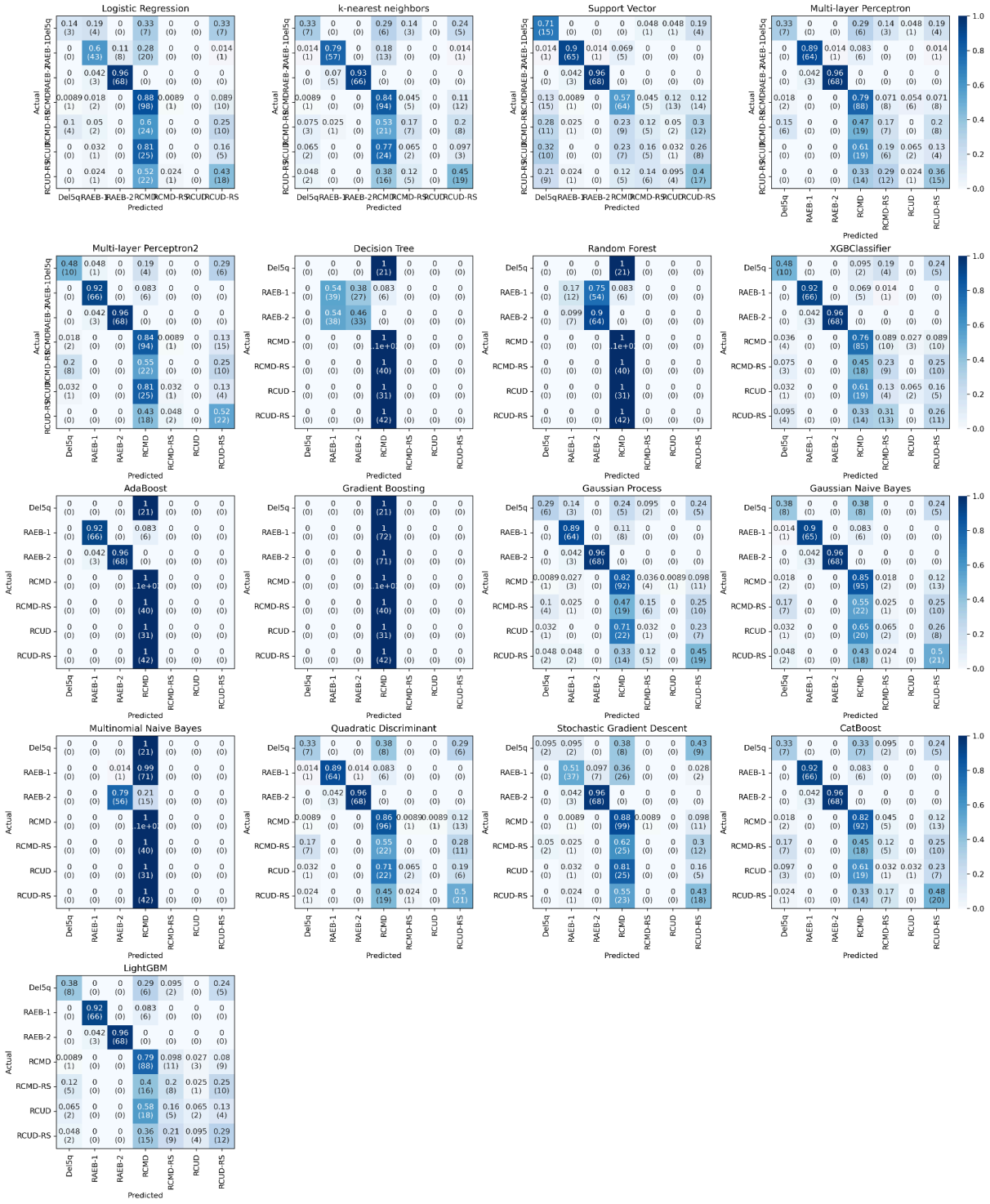

Figura S10. ***Matriz de confusão dos modelos otimizados treinados com o conjunto SMD unificado.***

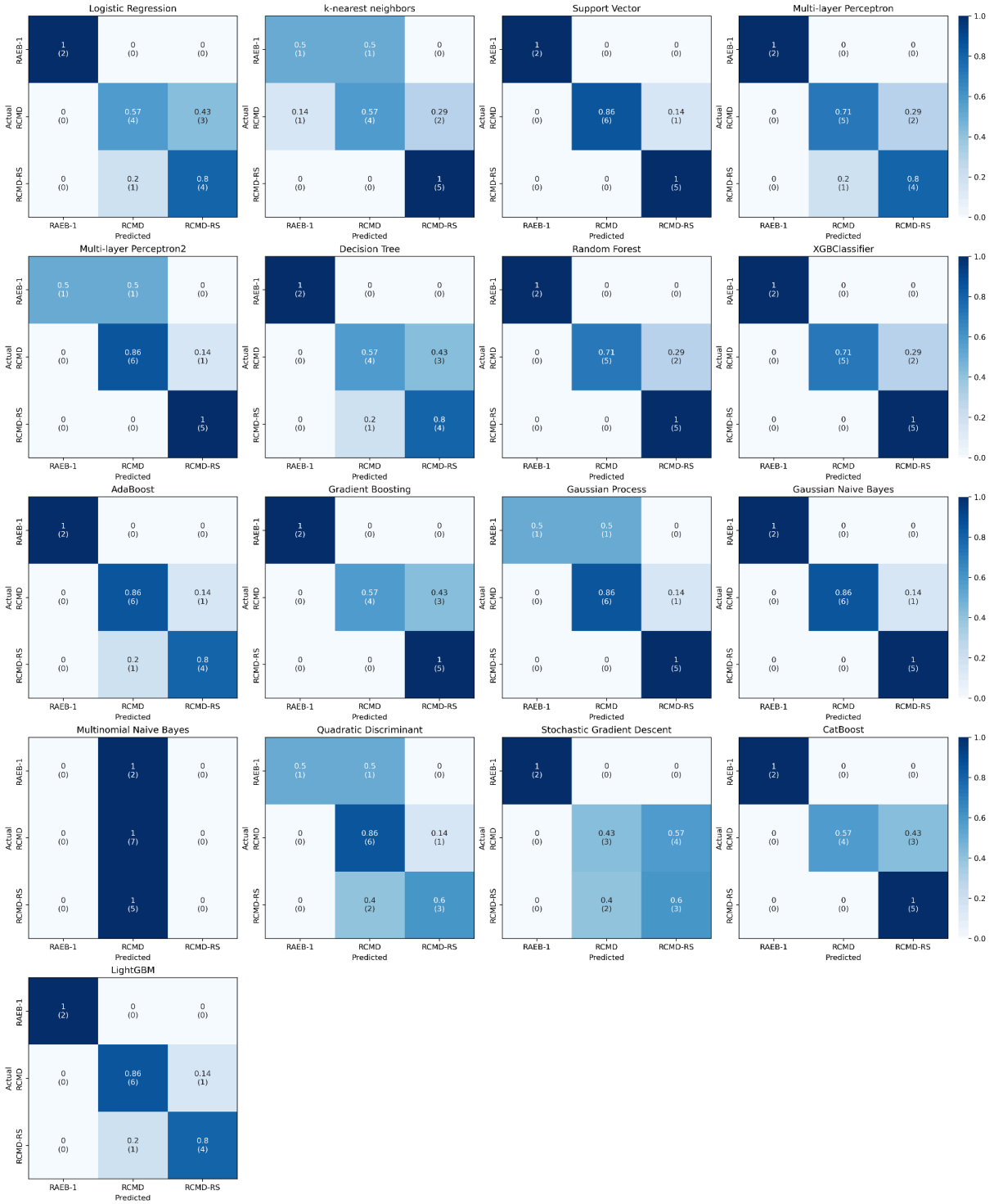

Figura S11. ***Matriz de confusão dos modelos não-otimizados treinados com o conjunto SMD do Hemocentro.***

Figura S12. ***Matriz de confusão dos modelos otimizados treinados com o conjunto SMD do Hemocentro.***

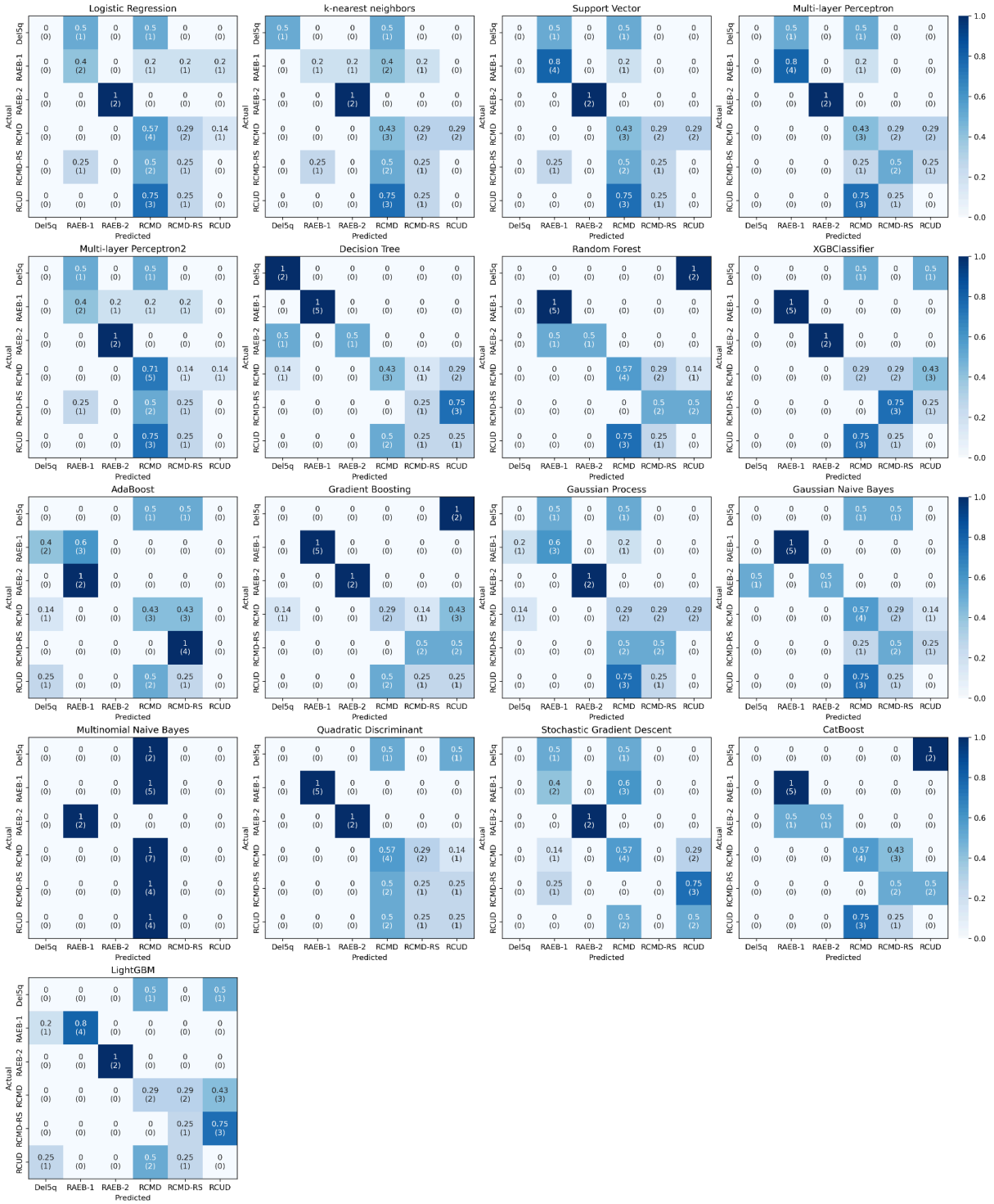

Figura S13. ***Matriz de confusão dos modelos não-otimizados treinados com o conjunto SMD da UFC.***

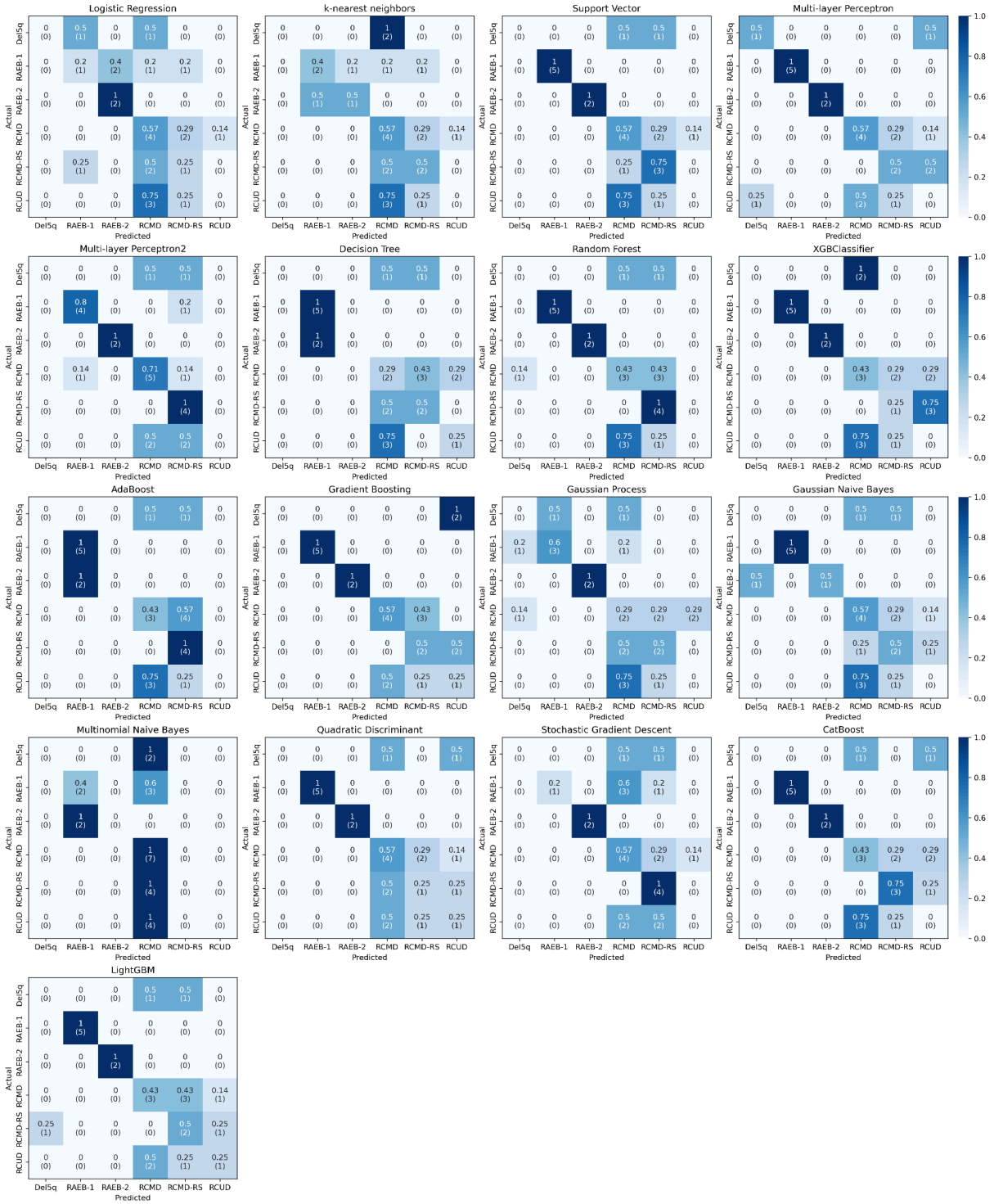

Figura S14. ***Matriz de confusão dos modelos otimizados treinados com o conjunto SMD da UFC.***

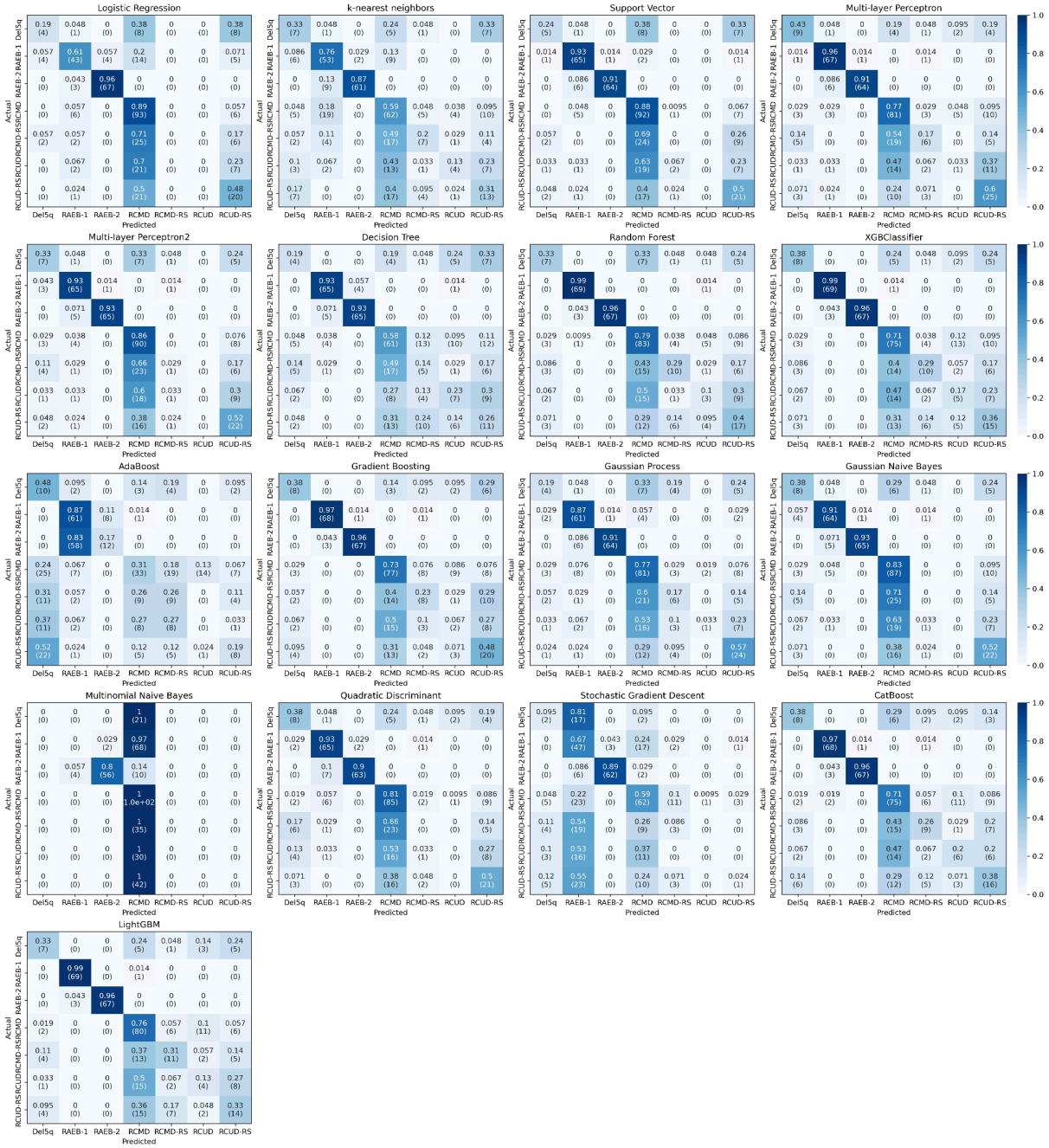

Figura S15. ***Matriz de confusão dos modelos não-otimizados treinados com o conjunto SMD do MDS-IWG.***

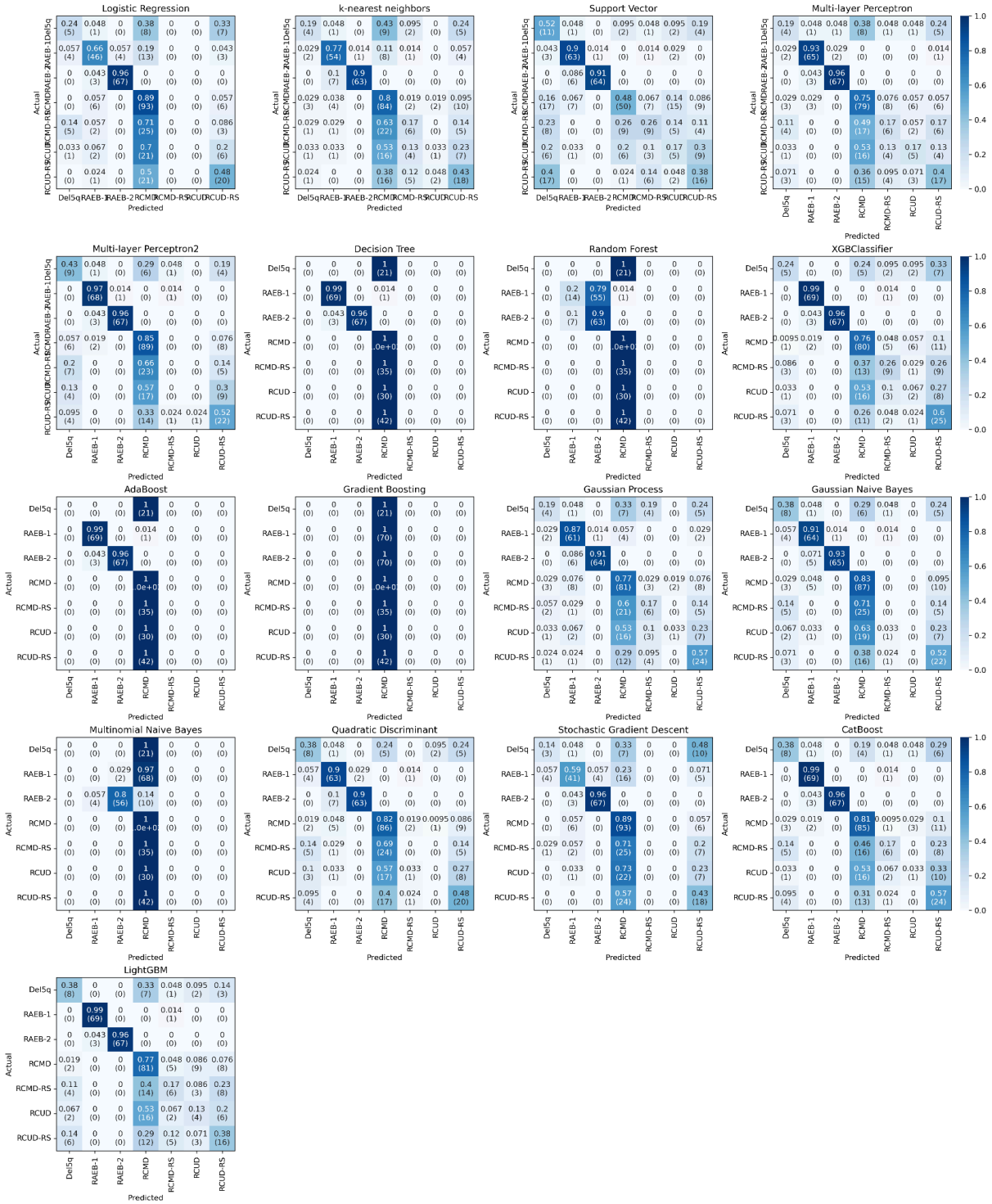

Figura S16. ***Matriz de confusão dos modelos otimizados treinados com o conjunto SMD do MDS-IWG.***

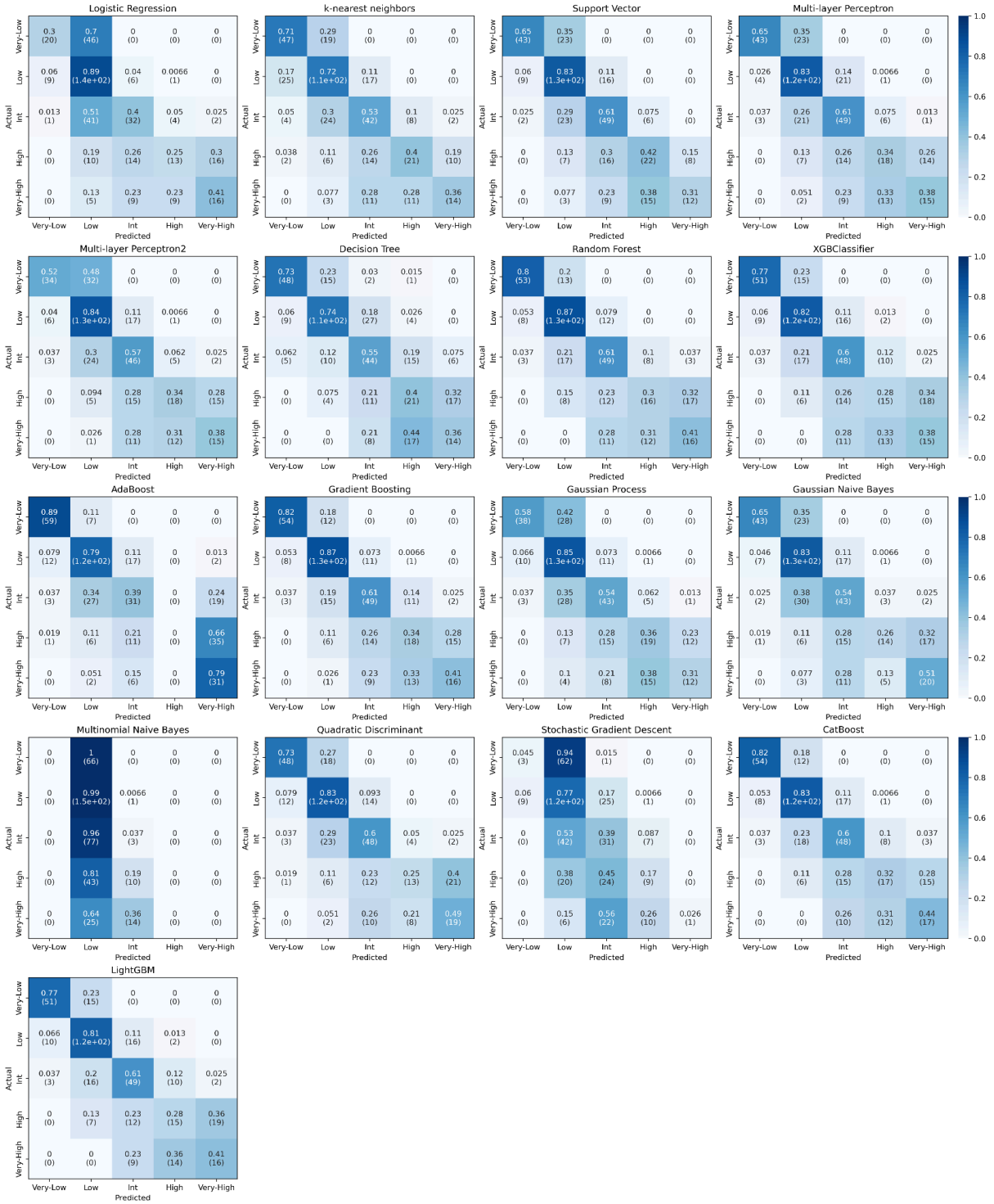

Figura S17. ***Matriz de confusão dos modelos não-otimizados treinados com o conjunto SMD para IPSS-R.***

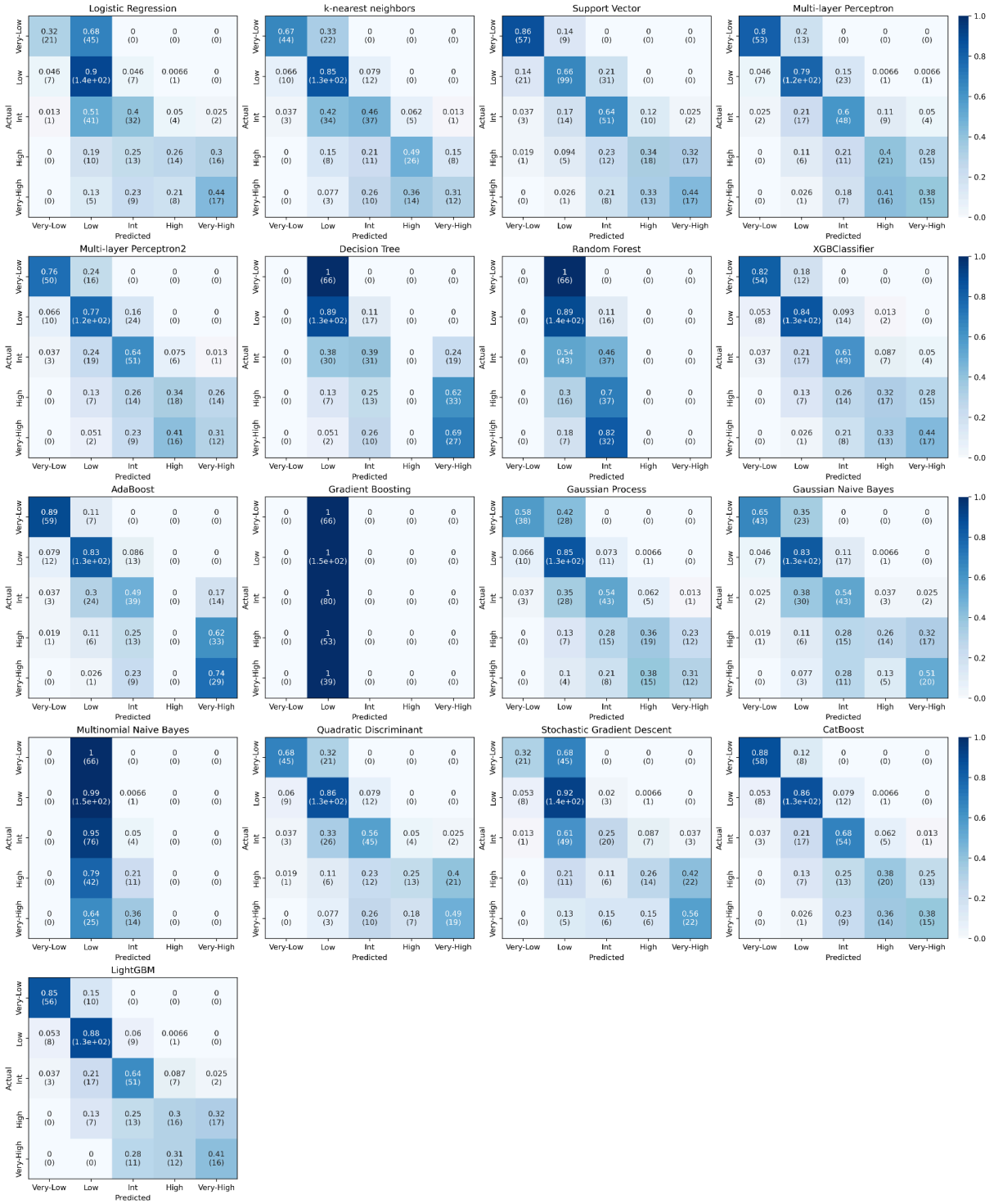

Figura S18. ***Matriz de confusão dos modelos otimizados treinados com o conjunto SMD para IPSS-R.***

6. Anexos

6.1. Anexo A

Foundation

/// Bayer Foundation Carl-Duisberg-Fellowship:
Welcome to Bayer Foundation

Leverkusen, 06.06.2022

Dear Fábio Malta de Sá Patroni,

We would like to congratulate you on behalf of Bayer Foundation for being selected as a recipient of the Bayer Foundation Carl-Duisberg-Fellowship for Medical Sciences.

At Bayer Foundation, we identify and nurture the next generation of exceptional talents with the ultimate goal of embedding a collaborative and interdisciplinary mindset in our future scientific leaders. Our scientific fellowships are a key component of supporting the future of cutting-edge research, and we are proud that we can support you on your journey.

In addition to the financial support, we would like to invite you to join the Bayer Foundation Science Network and will continue to reach out in this regard.

A fellowship agreement is attached, and we request that you complete and return this as soon as possible to ensure the timely commencement of the fellowship.

We look forward to our future interactions, and wish you all the best for your scientific endeavors.

Kind regards,

Dr. Karl Collins
Director, Science

Fabian Haß
Foundation Office

Bayer Foundation
Kaiser-Wilhelm-Allee 1
51368 Leverkusen

Germany

bayer.foundation@bayer.com
www.bayer-foundation.com

Bayer Science & Education Foundation

///Vorstand:
Serena Lin
Dr. Monika Lessl
///Vorsitzender des Kuratoriums:
Prof. Dr. Patrick Cramer

Bayer Cares Foundation

///Vorstand:
Serena Lin
Dr. Monika Lessl
///Vorsitzender des Kuratoriums:
Prof. Dr. André Habisch

/// Sitz der Stiftung: Leverkusen

6.2. Anexo B

DECLARAÇÃO

Em observância ao **§5º do Artigo 1º da Informação CCPG-UNICAMP/001/15**, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Tese de Doutorado, intitulada "**APRENDIZADO DE MÁQUINA APLICADO AO ESTUDO DE SUBTIPOS DAS MALIGNIDADES MIELOIDES**", desenvolvida no Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos à Biossegurança.

Assinatura : Fábio M. de Sá Patroni

Nome do(a) autor(a): Fábio Malta de Sá Patroni

RG n.° 49523302-X-SSP/SP

Assinatura : Marcelo Brandão

Nome do(a) autor(a): Marcelo Mendes Brandão

RG n.° 1406735-SSP/DF

6.3. Anexo C

DECLARAÇÃO

As cópias de artigos de minha autoria ou de minha coautoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada "**APRENDIZADO DE MÁQUINA APLICADO AO ESTUDO DE SUBTIPOS DAS MALIGNIDADES MIELOIDES**", não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 26 de Agosto de 2024

Assinatura : Fábio M. de Sá Patroni

Nome do(a) autor(a): Fábio Malta de Sá Patroni

RG n.° 49523302-X-SSP/SP

Assinatura : Marcelo Brandão

Nome do(a) autor(a): Marcelo Mendes Brandão

RG n.° 1406735-SSP/DF