



UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE
CURSO DE BIOTECNOLOGIA

DÉBORA VIRGÍNIA DA COSTA E LIMA

**O uso de redes neurais artificiais na análise de dados clínicos e
genéticos do câncer de pulmão**

**Mossoró/RN
Fevereiro 2020**

DÉBORA VIRGÍNIA DA COSTA E LIMA

O uso de redes neurais artificiais na análise de dados clínicos e genéticos do câncer de pulmão

Monografia apresentada à Universidade Federal Rural do Semi-Árido - UFERSA, Campus Mossoró, para a obtenção do título de Bacharel em Biotecnologia.

Orientador: Professor MsC Taffarel Melo Torres

Mossoró/RN
Fevereiro 2020

© Todos os direitos estão reservados a Universidade Federal Rural do Semi-Árido. O conteúdo desta obra é de inteira responsabilidade do (a) autor (a), sendo o mesmo, passível de sanções administrativas ou penais, caso sejam infringidas as leis que regulamentam a Propriedade Intelectual, respectivamente, Patentes: Lei nº 9.279/1996 e Direitos Autorais: Lei nº 9.610/1998. O conteúdo desta obra tomar-se-á de domínio público após a data de defesa e homologação da sua respectiva ata. A mesma poderá servir de base literária para novas pesquisas, desde que a obra e seu (a) respectivo (a) autor (a) sejam devidamente citados e mencionados os seus créditos bibliográficos.

L732 Lima, Débora Virgínia da Costa e.
u O uso de redes neurais artificiais na análise
 de dados clínicos e genéticos do câncer de pulmão /
 Débora Virgínia da Costa e Lima. - 2020.
 48 f. : il.

 Orientador: Taffarel Melo Torres.
 Monografia (graduação) - Universidade Federal
Rural do Semi-árido, Curso de Biotecnologia,
2020.

 1. Aprendizado de máquina. 2. Inteligência
artificial. 3. Dados Oncológicos. I. Melo Torres,
Taffarel, orient. II. Título.

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC's) foi desenvolvido pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP) e gentilmente cedido para o Sistema de Bibliotecas da Universidade Federal Rural do Semi-Árido (SISBI-UFERSA), sendo customizado pela Superintendência de Tecnologia da Informação e Comunicação (SUTIC) sob orientação dos bibliotecários da instituição para ser adaptado às necessidades dos alunos dos Cursos de Graduação e Programas de Pós-Graduação da Universidade.

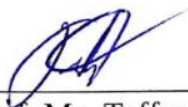
DÉBORA VIRGÍNIA DA COSTA E LIMA

**O uso de redes neurais artificiais na análise de dados
clínicos e genéticos do câncer de pulmão**

Monografia apresentada à Universidade
Federal Rural do Semi-Árido - UFERSA,
Campus Mossoró, para a obtenção do título
de Bacharel em Biotecnologia.

Defendida em: 07/02/2020

BANCA EXAMINADORA



Prof. Me. Taffarel Melo Torres
Orientador



Profa. Dra. Ana Carla Diógenes Suassuna Bezerra
Examinadora



Prof. Dr. Lívio Carvalho de Figueirêdo
Examinador

AGRADECIMENTOS

Agradeço, primeiramente a Deus por me permitir essa realização e por me iluminar sempre.

Agradeço aos meus pais, Francisco Chagas de Lima Júnior e Sandra Maria de Melo Costa e Lima, por me inspirarem e por me incentivarem sempre na busca de conhecimento. Aos meus irmãos, Euller Sadraque e Anna Clara que sempre me deram apoio.

Agradeço ao meu Orientador, Taffarel Melo Torres, que me guiou durante toda essa jornada, me aconselhando e capacitando em todos os momentos de dificuldade. Aos professores que me instruíram durante todo esse tempo.

Agradeço aos meus colegas de laboratório, Allen, Dyonatan, Maria Júlia e Stephany por sempre me auxiliarem. Aos meus amigos Maria da Conceição, Mônica, Vivian, Júlia Letycia, e Gregory que foram meus companheiros nessa caminhada.

Agradeço a todos os meus amigos, em especial, Júlia Alves, Matheus, Dara, Allan, Beatriz, Lara, Vinicius, Lauro, Samuel e João Paulo que sempre me ouviram, me motivaram e estiveram comigo nos momentos de descontração. Agradeço a Alan, Sinádia, Wladmir e Bruna que me escutaram e aconselharam.

Por fim, agradeço a todas as pessoas que não foram mencionadas, mas que de alguma forma ajudaram em minha formação.

“It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.”

Sir Arthur Conan Doyle, Sherlock Holmes

RESUMO

A área da saúde produz um grande volume de dados, que quando analisados, auxiliam na prevenção, diagnóstico e implementação de políticas públicas, mas a análise de tantos dados é complicada, logo as técnicas de big data surgem como uma possível solução para esse problema. O presente estudo teve como proposta a análise de dados clínicos e genéticos, relativos a neoplasias pulmonares, através do uso de redes neurais artificiais como método de predição de dados. As etapas deste estudo são: obtenção de dados clínicos e genéticos presentes em bancos de dados internacionais (GDC Data Portal) e locais (Registro Hospitalar de Câncer da Liga Mossoroense de Estudo e Combate ao Câncer), a seleção, organização e pré-processamento dos dados para inserção na rede neural, a implementação, treinamento e teste do algoritmo de rede neural artificial, e os resultados da predição e da validação das redes neurais. A etapa inicial realizada foi a obtenção dos dados clínicos e genéticos provenientes dos bancos de dados tanto internacional quanto local, e foram filtrados de acordo com as variáveis a serem utilizadas nas redes neurais. Após a etapa de filtragem e organização, iniciou-se o pré-processamento dos dados para permitir a inserção no algoritmo. Cinco redes foram construídas três delas implementadas com dados clínicos internacionais, uma rede implementada com dados clínicos internacionais e locais, e uma utilizando dados genéticos. A rede que apresentou melhor performance foi uma das redes com dados clínicos internacionais a qual apresentou em sua validação uma acurácia = 86,06%. Demonstrando que a implementação de redes neurais podem ser uma boa alternativa para a análise em pacientes com câncer.

Palavras-chave: Aprendizado de máquina, Inteligência artificial, Dados Oncológicos.

ABSTRACT

The health area produces a large volume of data, which when analyzed help in the prevention, diagnosis and implementation of public policies, but the analysis of so much data is complicated, so big data techniques appear as a possible solution to this problem.

The present study aimed to analyze clinical and genetic data of lung neoplasms, using artificial neural networks as a method for data prediction. The steps of this study are: using clinical and genetic data present in international databases (GDC Data Portal) and local (Hospital Cancer Registry of the Mossoroense League for the Study and Fight against Cancer), a selection, organization and pre-processing of the data for insertion in the neural network, an implementation, training and testing of the artificial neural network algorithm and the results of the prediction and validation of the neural networks. The initial step was to recover the clinical and genetic data from databases, both international and local, and were filtered according to the variables to be used in neural networks. After a filtering and organizing step, was started the pre-processing of the data to allow insertion into the algorithm. Five networks were built, three implemented with international clinical data, a network implemented with international and local clinical data and a network with genetic data. The network that performed best was one of the networks with international clinical data that presented its validation accuracy = 86.06%. Demonstrating that the implementation of neural networks can be a good alternative for analysis in cancer patients.

Keywords: Machine learning, Artificial intelligence, Oncological data.

LISTA DE FIGURAS

Figura 1 - Passos na descoberta de conhecimento em banco de dados -----	16
Figura 2 - Uma ilustração da estrutura da rede neural artificial. As setas conectam a saída de um nó à entrada de outro -----	17
Figura 3 - Ilustração da curva de características operacionais do receptor -----	19
Figura 4 - Esquema da construção do banco de dados integrado -----	23

LISTA DE GRÁFICOS

Gráfico 1 - Rede neural internacional contendo todas as variáveis clínicas -----	28
Gráfico 2 – Representação das variáveis por nível de importância na predição da primeira Rede neural, obtida pelo algoritmo de Garson -----	30
Gráfico 3 - Rede neural internacional contendo oito variáveis selecionadas -----	31
Gráfico 4 - Rede neural internacional contendo treze variáveis selecionadas -----	32
Gráfico 5 - Rede neural mista -----	33
Gráfico 6 - Rede neural genética -----	34
Gráfico 7 - Taxa real contra taxa predita pela rede clínica com todas as variáveis -----	35
Gráfico 8 - Curva de características operacionais do receptor da rede clínica com todas as variáveis -----	36
Gráfico 9 - Curva de características operacionais do receptor da rede clínica com oito variáveis -----	38
Gráfico 10 - Taxa real contra taxa predita pela rede clínica com oito variáveis -----	38
Gráfico 11 - Taxa real contra taxa predita pela rede clínica com treze variáveis -----	39
Gráfico 12 - Curva de características operacionais do receptor da rede clínica com treze variáveis -----	39
Gráfico 13 - Taxa real contra taxa predita pela rede mista -----	41
Gráfico 14 - Taxa real contra taxa predita pela rede genética -----	42
Gráfico 15 - Curva de características operacionais do receptor da rede genética -----	42

LISTA DE QUADROS

Quadro 1 - Performance das redes neurais artificiais -----	35
Quadro 2 - Matriz de Confusão das redes -----	37

LISTA DE ABREVIATURAS E SIGLAS

ALK	<i>Anaplastic Lymphoma Kinase</i>
ANN	<i>Artificial Neural Network</i>
AUC	Área sob a curva
BRAF	<i>Proto-oncogene B-Raf</i>
DM	<i>Data Mining</i>
EGFR	<i>Epidermal growth factor receptor</i>
GDC	<i>Genomic Data Commons Data Portal</i>
KDD	<i>Knowledge Discovery in Databases</i>
KRAS	<i>Kirsten Rat sarcoma viral oncogene homolog</i>
LKB1	<i>Liver kinase B1</i>
LMECC	Liga Mossoroense de Estudo e Combate ao Câncer
MEK1	<i>Mitogen-activated protein kinase 1</i>
MET	<i>Hepatocyte growth factor receptor</i>
ML	<i>Machine Learning</i>
MSE	Erro médio quadrático
NRAS	<i>NRAS proto-oncogene</i>
NSCLC	<i>Non-small cell lung cancer</i>
PIK3CA	<i>Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha</i>
RET	<i>RET proto-oncogene</i>
RHC	Registro Hospitalar de Câncer
ROC	<i>Receiver Operating Characteristic</i>
ROS1	<i>Proto-oncogene tyrosine-protein kinase ROS</i>
SCLC	<i>Small cell lung cancer</i>
TNM	Classificação de tumores malignos
WHO	<i>World Health Organization</i>

SUMÁRIO

1 INTRODUÇÃO	12
2 FUNDAMENTAÇÃO TEÓRICA	14
2.1 CÂNCER DE PULMÃO	14
2.2 BANCOS DE DADOS E DATA MINING	15
2.3 REDES NEURAIS ARTIFICIAIS	16
2.4 ACURÁCIA, ERRO MÉDIO QUADRÁTICO E CURVA DE CARACTERÍSTICAS OPERACIONAIS DO RECEPTOR	18
3. OBJETIVOS DA PESQUISA	20
3.1 OBJETIVO GERAL	20
3.2 OBJETIVOS ESPECÍFICOS	20
4. MATERIAL E MÉTODOS	21
4.1 ASPECTOS ÉTICOS	21
4.2 RECUPERAÇÃO DOS DADOS	21
4.2.1 Obtenção dos dados clínicos	21
4.2.2 Obtenção dos dados genéticos	22
4.3 ORGANIZAÇÃO DOS DADOS	22
4.3.1 Dados das redes clínicas	23
4.3.2 Dados da rede genética	24
4.4 IMPLEMENTAÇÃO DAS REDES NEURAIS	24
4.4.1 Implementação das redes Clínicas	24
4.4.2 Implementação da rede Genética	25
5. RESULTADOS E DISCUSSÃO	27
5.1 BANCO DE DADOS	27
5.2 REDES CLÍNICAS	27
5.3 REDE GENÉTICA	33
5.4 AVALIAÇÃO DA PREDIÇÃO DAS REDES	34
6. CONSIDERAÇÕES FINAIS	43
REFERÊNCIAS	44

1. INTRODUÇÃO

O câncer, neoplasia ou tumor, é o termo utilizado para nomear um conjunto de mais de 100 doenças caracterizadas pelo crescimento desordenado de células causando a invasão de tecidos e órgãos, ocasionando a formação de tumores, que podem se espalhar para outras regiões do corpo, esse grupo heterogêneo de doenças classifica-se em diversos subtipos anatômicos e moleculares, de maneira que casos diferentes exigem estratégias específicas de diagnóstico e manejo (INCA, 2019). O controle do câncer consiste em componentes como a prevenção, diagnóstico precoce e triagem, tratamentos, cuidados paliativos e cuidados de sobrevivência, sendo que tal controle deve ser avaliado por meio de monitoramento, que incluem os registros de câncer, e possibilita cuidados integrados centrados no paciente (WHO, 2017).

Muitas neoplasias podem ser prevenidas evitando a exposição a fatores de risco comuns, como o fumo do tabaco, e muitas podem ser curadas por métodos como cirurgia, radioterapia e quimioterapia, especialmente em casos de diagnóstico precoce (MINISTÉRIO DA SAÚDE, 2014). A prevenção é a estratégia de saúde pública mais eficaz em termos de custo no controle de doenças não transmissíveis (WHO, 2017). No entanto, apenas a prevenção não é suficiente para combater o câncer, visto que suas causas não se detêm apenas a fatores ambientais, possuindo também associações com os fatores genéticos (RENAULT, 2017). Portanto são necessárias estratégias que priorizem a realização do diagnóstico, tratamento e identificação precoce, para que seja possível a detecção de câncer em estágios iniciais e um tratamento mais eficaz e menos dispendioso (WHO, 2017).

Segundo a WHO (*World Health Organization*) o câncer é uma das principais causas de morte no mundo, estima-se que foi responsável por 9,6 milhões de mortes em 2018, além disso, através das estimativas, é possível estabelecer prioridades para o controle do câncer. Em 2008, com base nos dados internacionais, o ônus do câncer quando considerada a incidência e a mortalidade foram estimados em 12,4 milhões de novos casos e 7,6 milhões de mortes e uma das neoplasias mais comuns em termos de incidência foi de pulmão com cerca de 1,52 milhões de casos, e ainda, devido ao mau prognóstico o câncer de pulmão, também foi a causa mais comum de morte, estimada em 1,31 milhões, logo os números de câncer de pulmão, tanto em incidência quanto em mortalidade estão entre os mais altos no panorama mundial (IARC, 2008).

O câncer de pulmão é uma das principais causas de morte por câncer em mulheres e

homens, em geral apenas 14% das pessoas diagnosticadas com câncer de pulmão sobrevivem cinco anos após o diagnóstico (KRISHNAIAH *et al*, 2013). Em razão da relevância desta neoplasia, o estabelecimento de evidências robustas de predisposição genética para o risco de câncer de pulmão tem uma utilidade clínica potencial não apenas para estratificação de risco populacional, mas também para prevenção primária (WANG *et al*, 2017).

O setor de saúde gera continuamente grande volume de dados que ainda são subutilizados. A mineração de dados é capaz de procurar informações novas e valiosas e está sendo usada na área da saúde principalmente com o propósito de prever o comportamento de várias doenças, bem como na assistência de diagnóstico para os médicos e na tomada de decisão clínica (JOTHI *et al*, 2015).

O perfil molecular do câncer de pulmão tem sido amplamente alimentado pelo avanço científico e tecnológico na pesquisa do genoma do câncer. O uso de dados genéticos presentes em bancos de dados de diferentes populações, associados ao desenvolvimento de algoritmos e escolhas terapêuticas são importantes ferramentas para uma medicina personalizada para câncer de pulmão guiada por genômica (EL-TELBANY, 2012).

Um método utilizado para análise e previsão de dados são as redes neurais artificiais, que consistem em um método de aprendizado de máquina, e pode ser utilizado para melhorar a precisão da previsão de resultados do câncer (BERTOLACCINI *et al*, 2017). As redes neurais artificiais são modelos estatísticos, nos quais a estrutura matemática reproduz a organização da rede neural biológica (CHEN *et al*, 2015).

Neste contexto, a presente proposta objetivou utilizar dados genéticos e clínicos de pacientes com câncer de pulmão, presentes em bancos de dados públicos internacionais de câncer, em modelos preditivos de redes neurais artificiais. Além disso, comparou-se os resultados com os dados de pacientes locais presentes no Registro Hospitalar de Câncer (RHC) da Liga Mossoroense de Estudo e Combate ao Câncer (LMECC).

2. FUNDAMENTAÇÃO TEÓRICA

2.1 CÂNCER DE PULMÃO

O câncer de pulmão é a neoplasia mais comum no mundo e a causa mais comum de morte por câncer, apresentando uma alta letalidade e padrões de mortalidade próximos aos de incidência (FERLAY *et al*, 2014). Pode ser classificado em duas categorias principais: o câncer de pulmão de células não pequenas (NSCLC - *Non-small cell lung cancer*) e o câncer de pulmão de células pequenas (SCLC - *small cell lung cancer*). O tratamento, prognóstico e sobrevivência são dependentes de fatores como o tipo histológico, estágio e saúde geral do paciente (MINISTÉRIO DA SAÚDE, 2014). Dentre os possíveis tratamentos para esta neoplasia estão a cirurgia, quimioterapia e radioterapia (KRISHNAIAH *et al*, 2013).

Em relação aos fatores de risco para o câncer de pulmão, o mais relevante é o tabagismo, onde a maioria dos cânceres de pulmão poderia ser evitada com a eliminação do início do tabagismo e com a cessação do tabagismo entre fumantes (MINISTÉRIO DA SAÚDE, 2014). No entanto, além do tabaco outros fatores de risco relacionados a esses tumores são exposição a carcinógenos ambientais e ocupacionais e a poluição do ar (TORRE *et al*, 2015). Além da identificação dos fatores de riscos ambientais a identificação de fatores genéticos que influenciam o risco de câncer é fundamental para obtenção de novas estratégias relacionadas à prevenção, diagnóstico e tratamento de tumores (INCA, 2017).

Uma grande quantidade de informação genômica pode transformar a compreensão atual do câncer de pulmão, abrindo portas para uma personalização das terapias, a partir do sequenciamento clínico e montagem do perfil molecular de pacientes, com a finalidade de fortalecer decisões (RENAULT, 2017). Desta forma, por meio de dados como esses, é possível também compreender diferenças genéticas dos tumores de pulmão em populações diferentes (YOUOLDEN *et al*, 2008). As modificações genéticas como variações no número de cópias e mutações em genes associados ao câncer de pulmão, tais como EGFR, BRAF, KRAS, MET, LKB1, PIK3CA, ALK, ROS1, RET, NRAS e MEK1, estão fortemente ligadas a forma como a doença se desenvolve, portanto, a análise de genes revela informações vitais sobre a estrutura molecular da neoplasia (EL-TELBANY, 2012; VAVALÀ *et al*, 2014 ; SHOLL *et al*, 2015).

2.2 BANCOS DE DADOS E DATA MINING

Diversos campos científicos já se tornaram orientados por dados, entre eles estão a bioinformática e a biologia computacional, essa característica desses campos da ciência podem modificar paradigmas da biologia tradicional, por meio de análises que integram bancos de dados e mineração de dados (CHEN *et al*, 2014). A análise dos dados genômicos de câncer sugere que a descoberta de todas as alterações genéticas que promovem fenótipos malignos está longe de acontecer, mas há uma grande necessidade de descobrir fatores genéticos raros, pois a sua detecção pode ser um fator chave no sucesso da terapia (GROSSMAN *et al*, 2016)

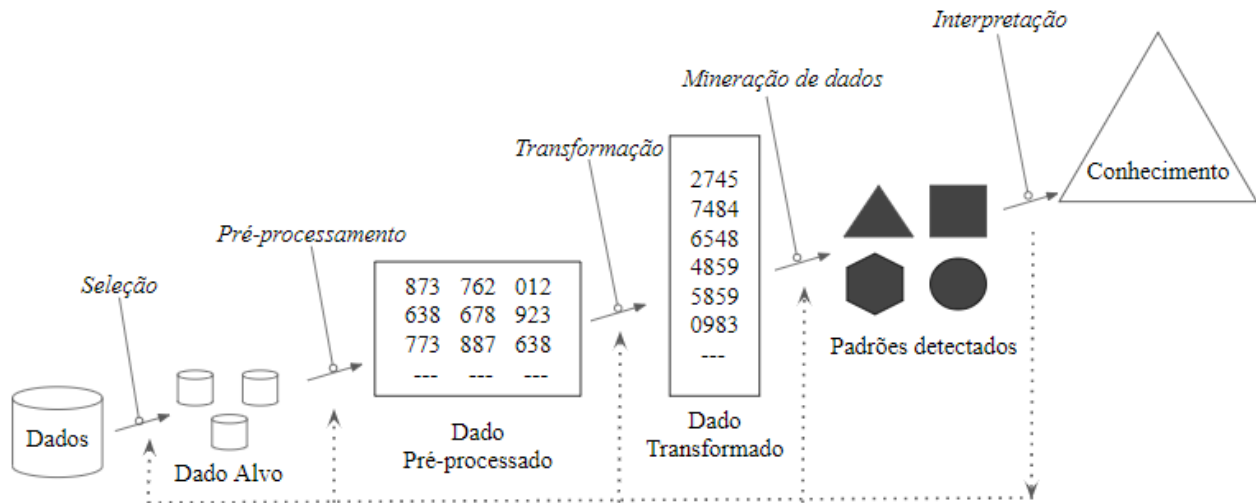
Para integrar dados de estudos diferentes a fim de obter uma conclusão com maior poder estatístico é possível utilizar metanálise, sendo que algumas já foram realizadas para câncer de pulmão (LEE *et al*, 2015; WANG *et al*, 2015; LEE *et al*, 2017). No entanto, a maioria limita-se a investigar uma variante genética, várias variantes de um gene ou várias variantes ao longo de uma via. A utilização de metanálises é importante para aumentar o conhecimento acerca das associações genéticas de doenças e as evidências de predisposição genética para o câncer de pulmão, tal conhecimento pode ter grande utilidade na área clínica e nas estratégias de prevenção (WANG *et al*, 2017).

O processo de descoberta de conhecimento em bancos de dados (KDD - *Knowledge Discovery in Databases*¹) e as técnicas de mineração de dados ou *Data Mining*² (DM) tornam possível verificar padrões entre variáveis, para prever resultados de uma doença, por meio de análise de casos armazenados em bancos de dados (KRISHNAIAH *et al*, 2013). O KDD é uma área ampla que visa extrair conhecimento útil a partir de um grande volume de dados através de mineração de dados, isso é possível devido aos avanços em áreas como: reconhecimento de padrões, bancos de dados, inteligência artificial e aprendizado de máquina (CHAURASIA, 2017). Desta forma, para que o processo de descoberta de conhecimento ocorra é preciso seguir passos (Figura 1) que vão desde a obtenção dos dados brutos até o novo conhecimento obtido (KRISHNAIAH *et al*, 2013). O *data mining* é um dos processos que permitem a obtenção de informações a partir dos dados, esse método pode ser usado em campos variados de tecnologia, negócios e ciência, usando áreas de conhecimento como aprendizado de máquina, inteligência artificial, probabilidade e estatística (JOTHI *et al*, 2015).

¹ Em tradução livre: DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS.

² Em tradução livre: MINERAÇÃO DE DADOS.

Figura 1: Passos na descoberta de conhecimento em banco de dados.



Fonte: Adaptada de KRISHNAIAH *et al*, 2013

2.3 REDES NEURAIS ARTIFICIAIS

A classificação de pacientes com câncer em grupos de alto ou baixo risco motiva pesquisadores de áreas como biomedicina e bioinformática a estudar métodos de aprendizado de máquina (*Machine Learning*³ - ML), um dos métodos que têm sido amplamente utilizados para desenvolvimento de modelos preditivos na pesquisa do câncer incluem Redes Neurais Artificiais (*Artificial Neural Network*⁴ - ANN). Os métodos de ML têm potencial para melhorar a compreensão da progressão do câncer, mas necessitam de validação para que sejam considerados na prática clínica (KOUROU *et al*, 2015).

Uma ANN é um modelo matemático inspirado em sistemas nervosos biológicos, uma rede neural apresenta-se em um conjunto de neurônios simulados interconectados, utilizando essas conexões como abordagem para o processamento de informações (AGRAWAL e AGRAWAL, 2015). As ANNs têm como base a organização hierárquica dos neurônios no cérebro, de forma que um neurônio recebe múltiplas entradas atribuindo a cada entrada um peso e realiza tomada de decisão dependendo da soma dos dados ponderados (Equação 1). De maneira

³ Em tradução livre: APRENDIZADO DE MÁQUINA.

⁴ Em tradução livre: REDES NEURAIS ARTIFICIAIS.

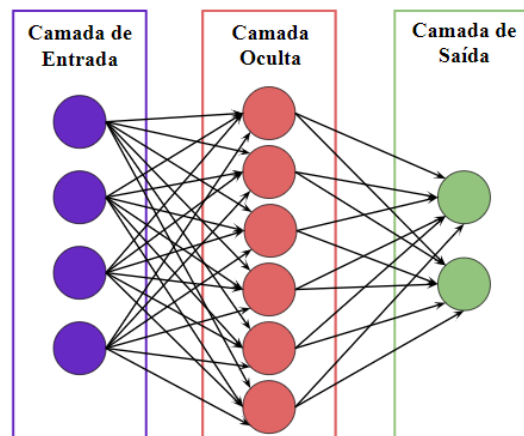
geral uma rede neural consiste em, pelo menos, três camadas compostas de unidades lógicas (Figura 2), nas quais a camada de entrada recebe um conjunto de dados, uma ou mais camadas ocultas sintetizam os dados e a camada de saída recebe a(s) camada(s) oculta(s) para gerar a resposta de interesse para pesquisa (BERTOLACCINI *et al*, 2017). Além disso, outros elementos que são essenciais em uma rede neural são os pesos e biases, os pesos são o fator mais importante na conversão da entrada para a formação da saída, pois tratam-se de parâmetros numéricos os quais determinam a intensidade da influência de um neurônio em outro, já um bias é um parâmetro adicional utilizado para ajustar a saída juntamente com a soma ponderada das entradas para o neurônio (CIABURRO,2017).

Equação 1: Processo realizado por um neurônio.

$$Saída = Soma (Pesos * Entradas) + bias$$

Fonte: Adaptada de CIABURRO,2017.

Figura 2: Uma ilustração da estrutura da rede neural artificial. As setas conectam a saída de um nó à entrada de outro.



Fonte: Adaptada de KOUROU *et al*, 2015.

A estrutura da ANN está intimamente ligada ao algoritmo de aprendizagem utilizado para treinar a rede, em geral essas estruturas podem ser classificadas em três classes de arquiteturas de rede fundamentalmente diferentes, são elas: as redes com camada única, as redes com múltiplas camadas e as redes recorrentes. As redes com múltiplas camadas se distinguem das outras pela presença de uma ou mais camadas ocultas, que tornam a rede capaz de extrair estatísticas de ordem elevada, tal habilidade é muito valiosa quando o tamanho da camada de entrada é elevada,

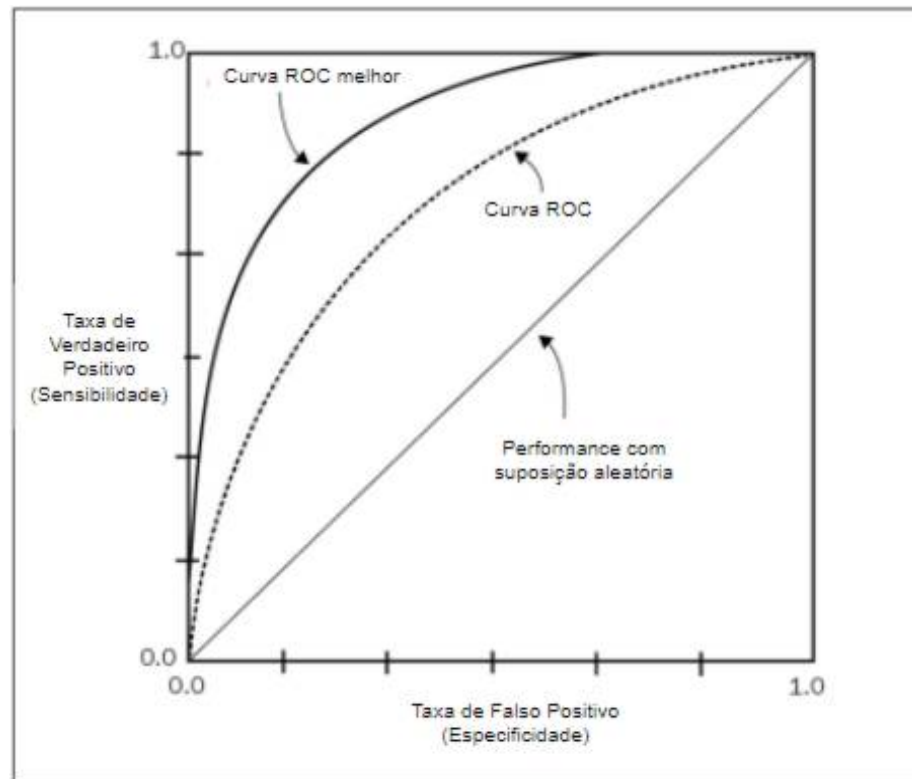
os sinais de saída de cada camada oculta funciona como sinais de entrada para a camada subsequente (HAYKIN, 2001). Em redes neurais artificiais, as camadas de neurônios são configuradas, e cada neurônio possui um peso que determina sua importância, um método para conseguir isso é atribuir pesos aleatórios e ajustá-lo iterativamente para aprimorar a correlação, ANNs que apresentam uma ou duas camadas ocultas são consideradas redes de aprendizado de máquina supervisionado, esse tipo de rede de aprendizado é um algoritmo que analisa um conjunto de dados de treinamento para criar uma função adequada e essa função será generalizada pela máquina para prever valores de saída desconhecidos (BIBAULT *et al*, 2016).

A rede neural pode modelar relacionamentos simples e complexos, e funciona de maneira adaptativa sofrendo mudanças na fase de aprendizado. As vantagens da utilização de ANN são: a aprendizagem adaptativa, refere-se capacidade de aprender a executar tarefas a partir dos dados fornecidos para treinamento, a auto-organização, após o recebimento de treinamento a rede neural pode criar sua própria organização, a operação em tempo real, cálculos de redes neurais podem ser realizados em paralelo, e a tolerância a falha por meio de codificação redundante de informações, pois apesar de danos parciais nas ANNs levarem a degradação do desempenho, as habilidades das redes podem ser recuperadas mesmo após grandes danos (AGRAWAL e AGRAWAL, 2015).

2.4 ACURÁCIA, ERRO MÉDIO QUADRÁTICO E CURVA DE CARACTERÍSTICAS OPERACIONAIS DO RECEPTOR

Uma ANN é um modelo capaz de aprender com os dados e prever a saída a partir desse aprendizado, para verificar se a rede é capaz de acertar a predição é necessário avaliar o resultado da rede neural. Alguns parâmetros que podem auxiliar na análise da predição da rede são: a acurácia e o erro médio quadrático (MSE), além disso a curva de características operacionais do receptor (curva ROC) também é uma forma de ilustrar a capacidade preditiva da rede. A acurácia é a medida de desempenho do modelo, ou seja, trata-se da proporção de previsões corretas para as previsões totais, o MSE mede quão longe as previsões estão dos dados reais, já a curva ROC (Figura 3) é um gráfico no qual possibilita a visualização da capacidade preditiva de um sistema classificador binário, sendo plotado através da sensibilidade (taxa de verdadeiro positivo) contra a especificidade (taxa de falso positivo) (CIABURRO, 2017).

Figura 3: Ilustração da curva de características operacionais do receptor.



Fonte: Adaptada de CIABURRO,2017.

3. OBJETIVOS DA PESQUISA

3.1 OBJETIVO GERAL

Descobrir associações entre dados genéticos e clínicos de câncer de pulmão e o status vital de pacientes utilizando redes neurais artificiais.

3.2 OBJETIVOS ESPECÍFICOS

- Obter e analisar dados do câncer de pulmão de bancos de dados públicos e do banco de dados do Registro Hospitalar de Câncer da Liga Mossoroense de Estudos e Combate ao Câncer;
- Produzir um banco de dados com os dados internacionais integrados;
- Gerar redes neurais artificiais com variáveis genéticas e clínicas;
- Utilizar a rede neural artificial para descobrir associações genéticas do câncer de pulmão;
- Comparar associações descobertas com dados do Registro Hospitalar de Câncer da Liga Mossoroense de Estudos e Combate ao Câncer.

4. MATERIAL E MÉTODOS

4.1 ASPECTOS ÉTICOS

Com o objetivo de proteger a identidade e privacidade dos pacientes, os dados clínicos e genéticos obtidos nos bancos de dados internacionais, não apresentam uma identificação dos pacientes, e os dados obtidos no Registro Hospitalar de Câncer da Liga Mossoroense de Estudos e Combate ao Câncer já estão disponíveis no IntegradorRHC⁵ e utilizam valores semi-identificadores. Além disso, qualquer inferência dos resultados da pesquisa relacionados a diagnóstico, tratamento e probabilidade de sobrevida estão apresentadas de forma desidentificada.

4.2 RECUPERAÇÃO DOS DADOS

A análise dos dados de câncer de pulmão foi planejada de modo que o conjunto de dados foi dividido em dois grupos, um contendo apenas dados clínicos e outro contendo dados genéticos, sendo aplicadas redes neurais diferentes para cada grupo.

Dentro de cada rede neural há ainda outra divisão referente ao conjunto de dados de treinamento, utilizado para treinar a rede por meio do aprendizado de máquina, além disso ainda há o conjunto de dados teste, utilizado para validação da rede como método de predição.

4.2.1 Obtenção dos dados clínicos

Os dados clínicos avaliados são dados de câncer de pulmão provenientes do banco de dados público internacional: *Genomic Data Commons Data Portal (GDC Data Portal)*⁶, além disso foram também recuperados dados de pacientes locais obtidos no Registro Hospitalar de Câncer da Liga Mossoroense de Estudos e Combate ao Câncer. Atendendo aos critérios éticos envolvidos no processo.

Os dados internacionais foram obtidos no GDC devido a maior organização e disponibilidade, selecionando entre eles, apenas os dados referentes as neoplasias pulmonares que

⁵ Disponível em: < <https://irhc.inca.gov.br/RHCNet/> > Acesso em: 20/01/2020.

⁶ Disponível em: < <https://portal.gdc.cancer.gov/> > Acesso em: 15/02/2019.

continham a variável de status vital do paciente, a qual será a variável utilizada como saída na rede neural, após esta seleção foram obtidos dados de 1008 pacientes.

Os dados de origem local foram obtidos de um conjunto de dados do RHC da LMECC, referentes ao câncer de pulmão, contendo 172 registros de pacientes.

4.2.2 Obtenção dos dados genéticos

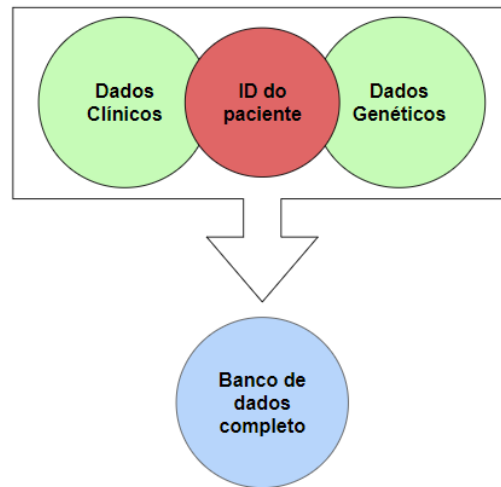
Os dados genéticos avaliados são dados de câncer de pulmão do *GDC Data Portal*. Foram obtidos os dados genéticos dos mesmos 1008 pacientes, para extrair o todo o conjunto de dados genéticos foram utilizados códigos na linguagem de programação *Python*, a extração ocorreu por meio do código identificador dos pacientes, que foi recuperado anteriormente no conjunto de dados clínicos, para resgatar o conjunto de genes de cada indivíduo.

A partir dessa metodologia foi possível recuperar apenas os genes que sofreram mutações na região codificante da proteína. Após a obtenção do conjunto de genes de cada paciente separadamente, foi utilizado um código em *Shell script* para condensar todos esses dados em um mesmo documento.

4.3 ORGANIZAÇÃO DOS DADOS

Os dados recuperados foram preparados e organizados em planilhas para posterior aplicação na rede neural. Já que os dados internacionais clínicos e genéticos mesmo pertencendo aos mesmos pacientes foram obtidos de forma diferente foi utilizado o programa *SQLite* para organizar um banco de dados completo com todos os dados internacionais, utilizando como variável comum o código identificador dos pacientes (Figura 4).

Figura 4: Esquema da construção do banco de dados integrado



4.3.1 Dados das redes clínicas

O banco de dados internacionais obtido continha variáveis vazias ou duplicadas, essas variáveis foram descartadas, e as variáveis restantes foram selecionadas.

Foram produzidas quatro redes neurais clínicas, as três primeiras foram implementadas utilizando apenas dados internacionais, de maneira que a primeira utilizou um conjunto de 20 variáveis de entrada, tais variáveis foram: idade, etnia, gênero, raça, ano de nascimento, classificação de tumores malignos (TNM), localização do tumor primário, morfologia, primeiro diagnóstico, malignidade prévia, tratamento prévio, malignidade sincronizada, local da biópsia, tecido ou órgão de origem, estágio do tumor, ano do diagnóstico, número de cigarros por dia, maços de cigarro por ano. A segunda e a terceira rede foram geradas utilizando o mesmo conjunto de dados da primeira rede, no entanto para estas, ocorreu a seleção de algumas variáveis da primeira rede, logo essas duas redes possuíram respectivamente 8 e 13 variáveis de entrada. A variável TNM foi considerada como três variáveis da extensão do câncer referentes ao tumor, linfonodos e metástases respectivamente.

Para a implementação da última rede neural clínica, os dados utilizados para treinamento são os dados obtidos no banco de dados internacional, e os dados teste são os obtidos no banco de dados local. Após a recuperação dos dados e a seleção dos pacientes, foram identificadas as

variáveis clínicas para utilização na rede neural, baseando-se nas variáveis comuns entre os dois conjuntos de dados obtidos, tais variáveis foram: idade, gênero, raça, TNM, localização do tumor primário, morfologia e status vital. Após identificadas as variáveis comuns foi construída uma tabela contendo todos os pacientes internacionais e locais, de forma que os dados foram organizados e padronizados para poderem se encaixar na rede neural.

4.3.2 Dados da rede genética

Na implementação da rede genética o conjunto de dados para o treinamento e teste foram obtidos apenas a partir do banco de dados internacional, as variáveis utilizadas foram o gene como variável de entrada e o status vital como variável de saída, como as duas variáveis foram obtidas separadamente respectivamente com o conjunto de dados genéticos e com o conjunto de dados clínico, obteve-se uma tabela a partir do banco de dados construído que contém todo o conjunto de dados internacionais, apenas com as duas variáveis da rede genética, posteriormente essa tabela foi padronizada para se encaixar na rede neural.

4.4 IMPLEMENTAÇÃO DAS REDES NEURAIIS

Para realizar a implementação da rede neural, inicialmente os dados separados para cada rede foram divididos em dois conjuntos, um conjunto de dados treinamento e um conjunto de dados teste, o conjunto de dados de treinamento é usado para ajustar os pesos que foram utilizados posteriormente para tratar os dados teste, esses dados testes foram usados para a validação da rede. Após a divisão dos conjuntos de dados, foi utilizado o programa *R Studio* e por meio de códigos implementados na linguagem *R* são aplicados os modelos de arquitetura de redes e o algoritmo de aprendizado. Para que se inicie o processo de treinamento, os dados devem ser adaptados e organizados para se adequar ao programa utilizado, e após o treinamento a rede deve ser testada para verificar sua capacidade de classificar os dados corretamente (TURBAN, 2009).

4.4.1 Implementação das redes Clínicas

Para implementação das redes clínica inicialmente realizou-se testes do algoritmo no

programa *R Studio* utilizando a linguagem de programação *R*, com auxílio do pacote *neuralnet*, utilizando dados de alguns pacientes e as variáveis gênero, raça, ano de diagnóstico e status vital, na construção de uma rede neural pouco robusta como teste para aprimorar o código. Em seguida, por meio desse teste foi possível observar que os dados necessitam de um pré-processamento, com a finalidade de codificar os dados qualitativos numericamente. Após realizar essa codificação o algoritmo foi aprimorado para receber o conjunto de dados completo. Além disso, utilizou-se os pacotes *readr*, *nnet*, *NeuralNetTools* e *PRROC* para possibilitar uma melhor importação dos dados e formação de tabelas e gráficos.

Foram produzidas quatro redes neurais clínicas para melhor avaliar a relação entre as variáveis presentes nos bancos de dados e o status vital dos pacientes com a finalidade de obter uma predição mais precisa.

Para a primeira rede clínica todas as variáveis selecionadas foram entradas, com exceção do status vital que foi a variável de saída e foi utilizada apenas uma camada oculta para possibilitar a visualização de quais variáveis foram mais importantes na predição. A segunda rede clínica foi construída com as oito variáveis mais importantes na predição da primeira rede para verificar se a seleção dessas variáveis é capaz de melhorar a predição, além disso foram utilizadas três camadas ocultas contendo 20, 10 e 5 nós respectivamente e duas repetições. A terceira rede clínica assim como a segunda foi composta por variáveis selecionadas, no entanto a seleção foi feita por meio de um refinamento na escolha das variáveis com finalidade de obter um melhor resultado, e foram utilizadas cinco camadas ocultas contendo 40, 20, 15, 10 e 5 nós respectivamente e duas repetições. Além disso, para essa rede, foram descartados todos os pacientes que possuíam alguma variável vazia.

A última rede foi a rede mista composta por dados internacionais e locais, contendo oito variáveis comuns entre os dois bancos de dados, e para essa rede, foram utilizadas três camadas ocultas contendo 20, 10 e 5 nós respectivamente e 2 repetições. Para todas as redes as linhas de interação entre as camadas representam pesos positivos na cor preta e pesos negativos na cor cinza.

Foi gerada a partir dessas redes tabelas e gráficos comparando o previsto com o real, um gráfico demonstrando quais foram as variáveis com mais peso para a rede sem camadas ocultas e curvas ROC para analisar a qualidade da predição.

4.4.2 Implementação da rede Genética

Na implementação da rede genética foi utilizada apenas uma variável de entrada que foi referente ao gene modificado e a variável de saída que foi o status vital do paciente, nessa rede apenas continha no conjunto de dados os genes que sofreram uma mutação na região codificante da proteína, e outros dados foram descartados. A fim de permitir uma melhor predição foram utilizadas três camadas ocultas, possuindo cada uma 20, 10 e 5 nós respectivamente e duas repetições. Assim como na rede clínica utilizou-se os pacotes *readr*, *nnet*, *NeuralNetTools* e *PRROC* para possibilitar uma melhor importação dos dados e formação de tabelas e gráficos. Foi gerada a partir dessa rede uma tabela e gráfico comparando o previsto com o real e uma curva ROC.

5. RESULTADO E DISCUSSÃO

5.1 BANCO DE DADOS

O GDC Data Portal é uma plataforma robusta baseada em dados que permite que pesquisadores sejam capazes de pesquisar e baixar dados de câncer para análise (GDC, 2020). No entanto na obtenção de dados a partir do GDC Data Portal a metodologia de recuperação direta aplicada aos dados clínicos e de exposição não é aplicável para os dados genéticos, ainda que eles também estejam disponíveis na plataforma, os dados genéticos apresentam uma recuperação mais dispendiosa. Logo a partir dos dados clínicos obtidos diretamente, o identificador referente a cada paciente presente nos dados resgatados e a utilização de algoritmos na linguagem de programação *Python* mostraram-se como uma possibilidade de metodologia para resolução do problema associado a seleção dos dados genéticos dos pacientes bem como sua obtenção.

No câncer, as interpretações clínicas e das mutações somáticas dependem de variáveis clínicas associadas ao tumor, tais como o tipo e o local do tumor, portanto um importante fator na análise de dados genômicos são as informações geradas a partir deles e o que essas informações significam em nível clínico (HUANG *et al*, 2017).

Desse modo devido à importância da associação das variáveis clínicas e genéticas e da condensação do conjunto de dados completo a construção de um banco de dados foi realizada, obtendo de tal forma um banco composto por todos os dados clínicos e genéticos dos 1008 pacientes com câncer de pulmão selecionados.

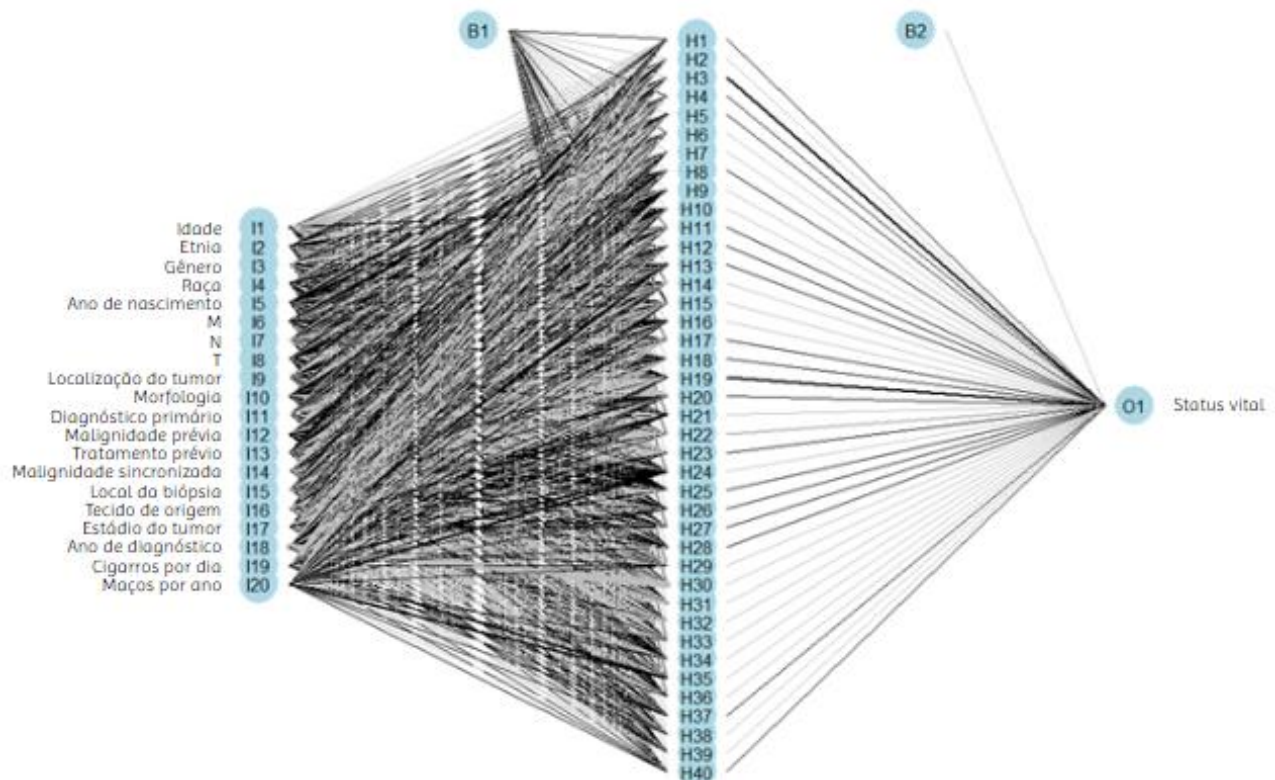
5.2 REDES CLÍNICAS

Foi possível visualizar diversas variáveis clínicas presentes no banco de dados, no entanto algumas dessas variáveis se apresentavam vazias, com a organização e padronização dos dados para se encaixarem na rede neural as variáveis vazias foram retiradas do conjunto de dados selecionados para serem utilizados nas redes neurais artificiais clínicas.

A primeira rede implementada foi uma rede contendo dados internacionais, de maneira que foi utilizado aproximadamente 70% dos dados para treinamento e 30% dos dados para teste. Esta rede contém as vinte variáveis de entrada selecionadas, após o processo de organização e padronização dos dados, e a variável de saída referente ao status vital, além disso nesta rede foi

implementada apenas uma camada oculta para permitir o uso do algoritmo de Garson. O resultado da estrutura da primeira rede construída pode ser observado no Gráfico 1.

Gráfico 1: Rede neural internacional contendo todas as variáveis clínicas. Círculos com letras I, B, H e O representam respectivamente variáveis de entrada, biases, camadas ocultas, e saída da rede e os números após as letras diferenciam as camadas. Linhas representam as interações entre as camadas. As Linhas pretas representam pesos positivos e as cinzas representam pesos negativos.



Por ser uma rede com uma camada oculta foi possível utilizar o algoritmo de Garson, pois ainda que o método seja limitado, devido a direção da resposta não poder ser determinada e apenas redes neurais com uma camada oculta e um nó de saída poderem ser avaliadas (BECK, 2018), tal limitação não influênciava tanto nessa rede, considerando que a mesma está dentro do padrão para a implementação deste algoritmo.

O algoritmo de Garson permite visualizar a importância relativa das variáveis de entrada em redes neurais para o resultado da saída (CIABURRO, 2017). No Gráfico 2 pode-se então observar dentre as vinte variáveis aplicadas nesta rede quais foram mais influentes no resultado final da primeira rede.

Após a verificação de quais eram as variáveis mais influentes na primeira rede construída foram selecionadas as oito variáveis mais importantes a fim de construir uma segunda rede neural

que obtivesse uma predição mais precisa baseada em um menor conjunto de variáveis.

A ANN com oito variáveis foi construída com 3 camadas ocultas como pode ser observado no Gráfico 3 e foi realizada uma repetição, além disso o grupo de dados de treinamento utilizado foi de aproximadamente 80%, essas adaptações foram feitas com a finalidade de melhorar capacidade de predição da rede.

Gráfico 2: Representação das variáveis por nível de importância na predição da primeira rede neural, obtida pelo algoritmo de Garson.

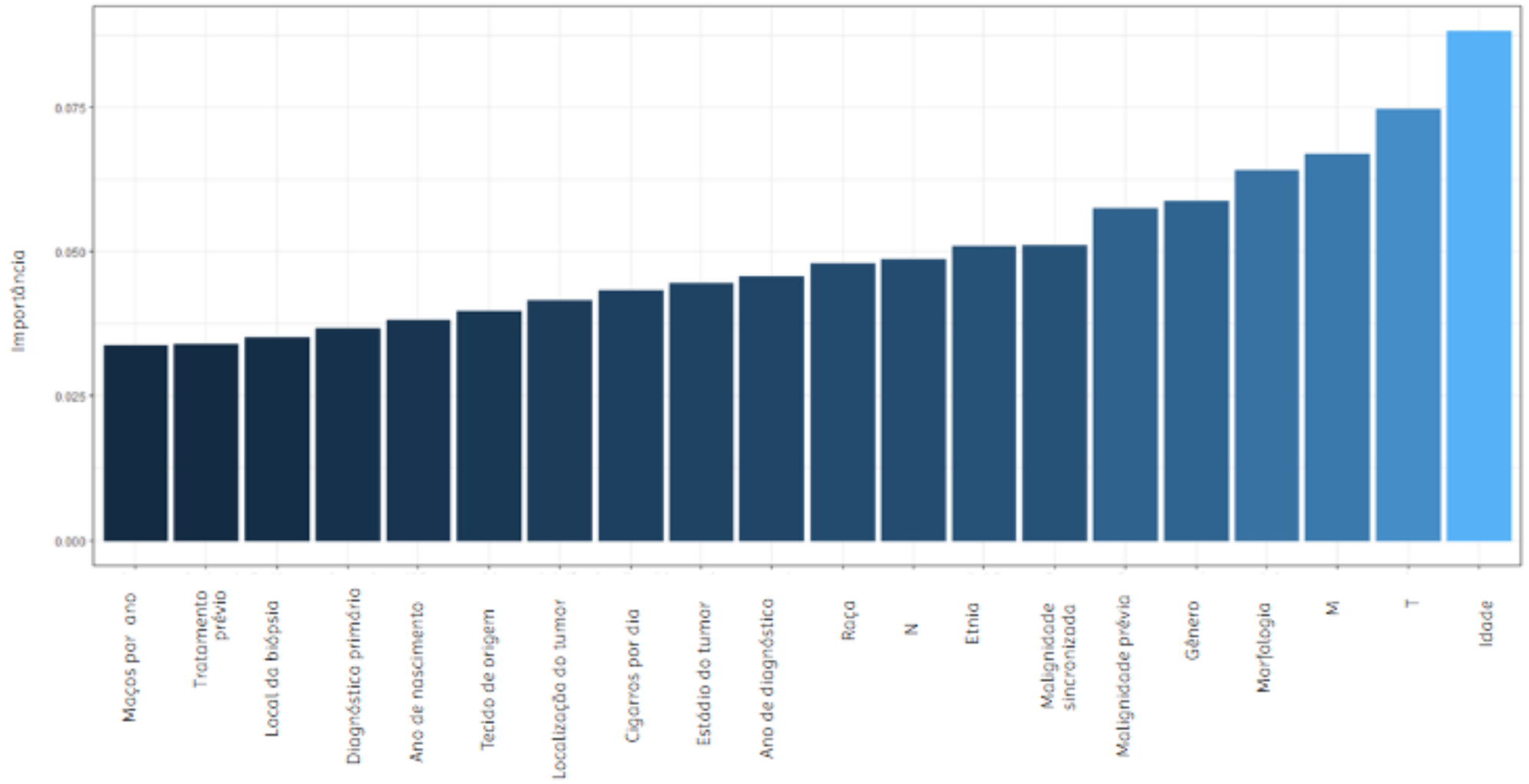
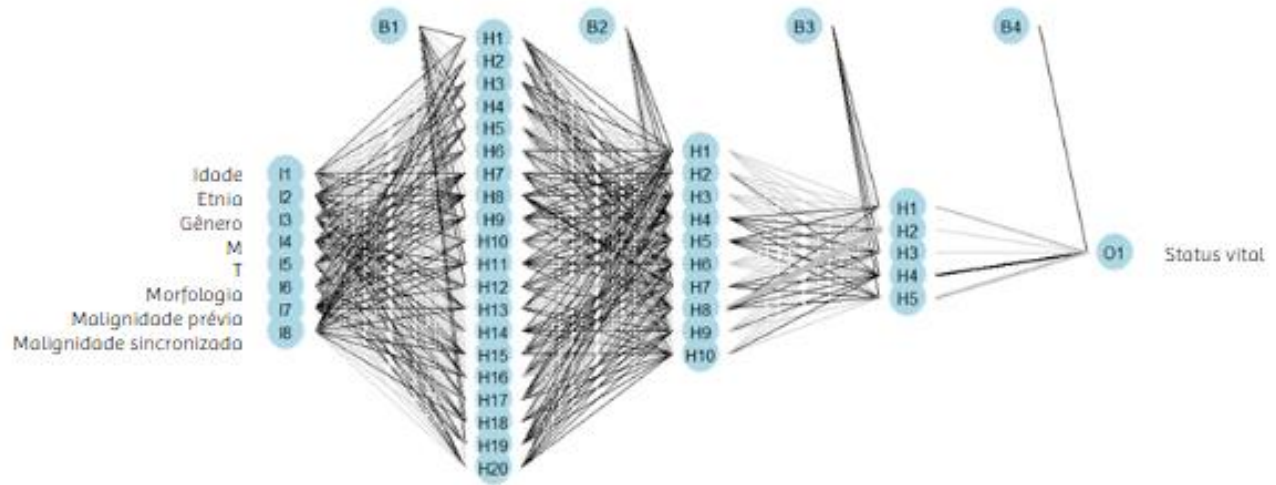


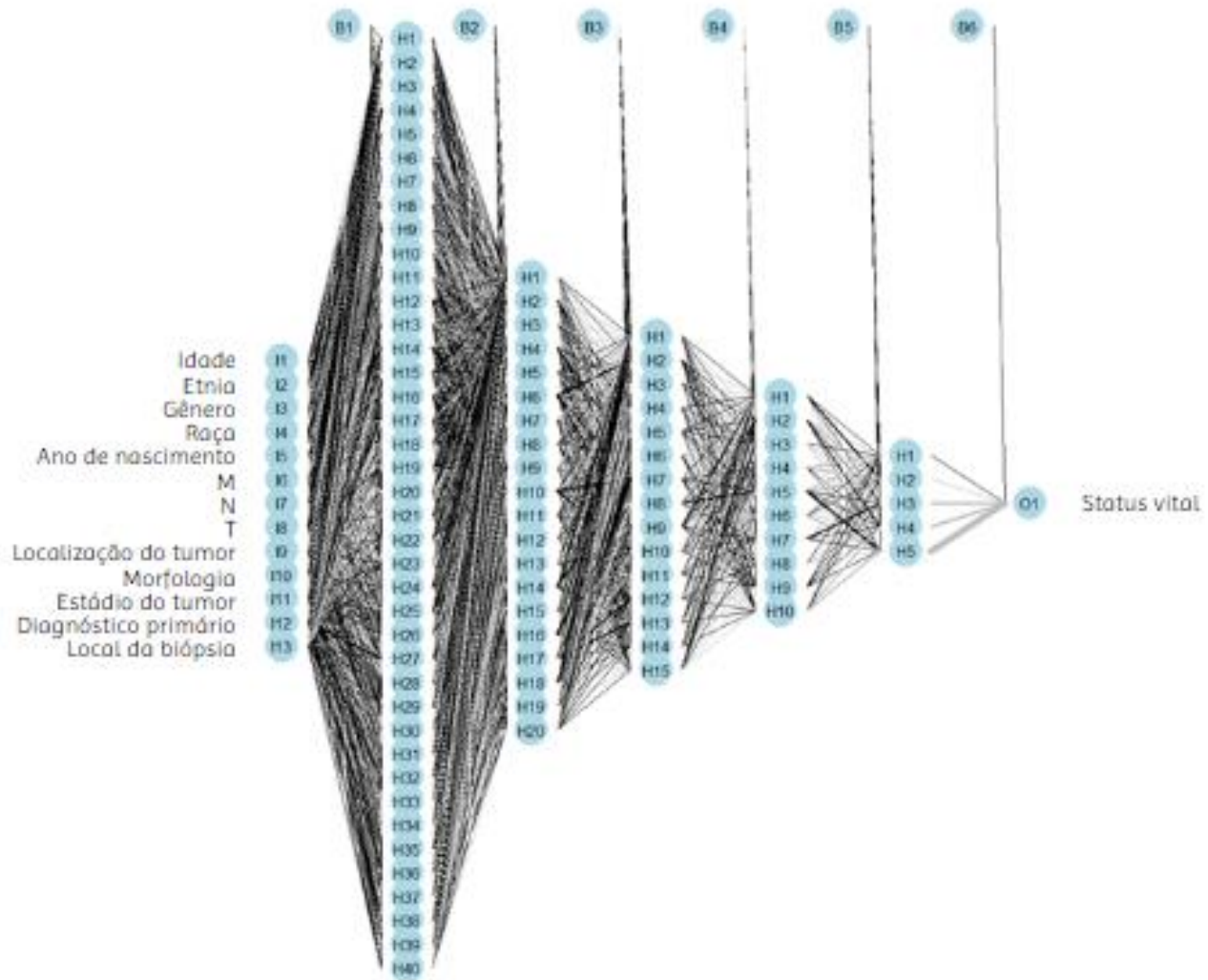
Gráfico 3: Rede neural internacional contendo oito variáveis selecionadas. Círculos com letras I, B, H e O representam respectivamente variáveis de entrada, biases, camadas ocultas, e saída da rede e os números após as letras diferenciam as camadas. Linhas representam as interações entre as camadas. As Linhas pretas representam pesos positivos e as cinzas representam pesos negativos.



Mediante os resultados de performance das duas primeiras redes, se fez necessário realizar a construção de uma terceira rede neural com variáveis clínicas internacionais. Esta rede foi construída por meio do refinamento das variáveis testando a performance da rede com a modificação aleatória das variáveis de maneira que o melhor resultado de predição foi selecionado. Além disso foram acrescentadas mais camadas ocultas, pois o aumento de camadas tem a capacidade melhorar a capacidade de predição, logo foram acrescentadas camadas até que esse aumento parou de influenciar no resultado.

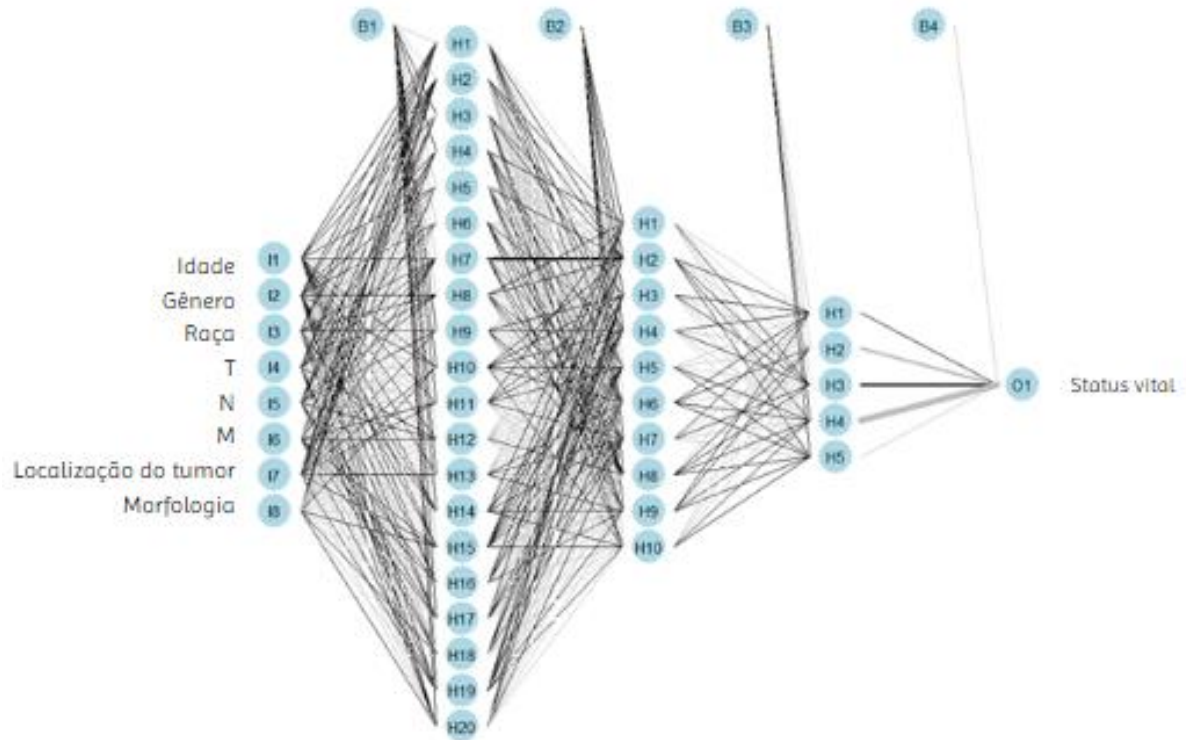
O resultado mais preciso para dados clínicos internacionais está apresentado nesta rede, na qual foram selecionadas treze variáveis, e foram utilizadas 5 camadas ocultas, duas repetições e com a finalidade de validação foi usado todo o conjunto de dados como treinamento, utilizando posteriormente 30% desse mesmo conjunto para validação da rede. A estrutura da rede com treze variáveis pode ser observada no Gráfico 4.

Gráfico 4: Rede neural internacional contendo treze variáveis selecionadas. Círculos com letras I, B, H e O representam respectivamente variáveis de entrada, biases, camadas ocultas, e saída da rede e os números após as letras diferenciam as camadas. Linhas representam as interações entre as camadas. As Linhas pretas representam pesos positivos e as cinzas representam pesos negativos.



A quarta rede implementada foi uma rede construída a partir de um conjunto de dados mistos, ou seja, constituído por dados internacionais e por dados locais, sendo os mesmos correspondentes ao conjunto de dados treinamento e teste respectivamente. Este procedimento foi realizado para avaliar se seria possível a partir desta rede prever o status vital dos pacientes locais por meio da utilização de dados internacionais. Esta rede foi feita com oito variáveis de entrada e 3 camadas ocultas, e duas repetições (Gráfico 5).

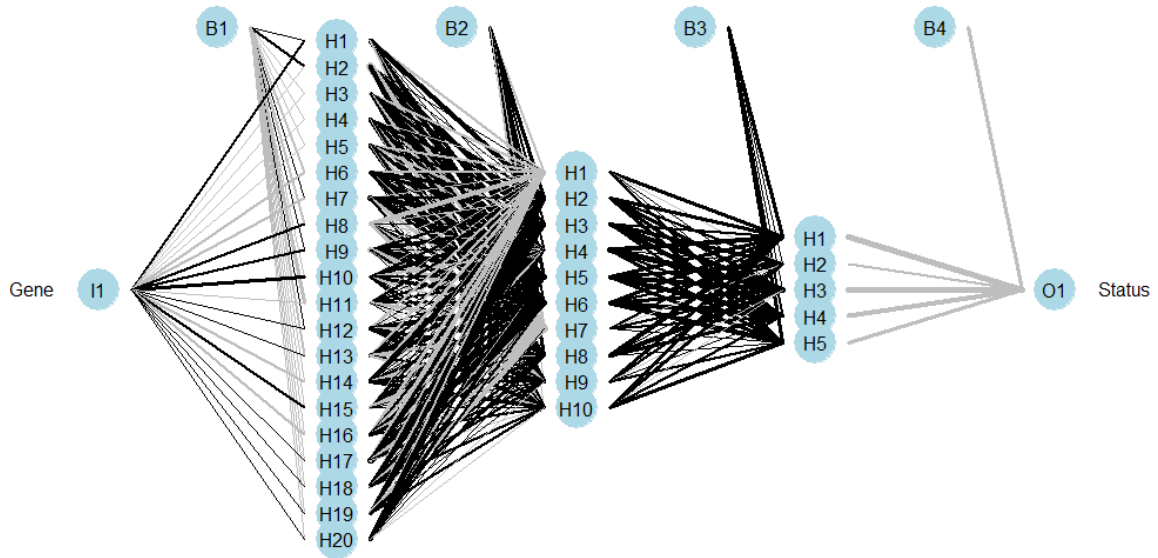
Gráfico 5: Rede neural mista. Círculos com letras I, B, H e O representam respectivamente variáveis de entrada, biases, camadas ocultas, e saída da rede e os números após as letras diferenciam as camadas. Linhas representam as interações entre as camadas. As Linhas pretas representam pesos positivos e as cinzas representam pesos negativos.



5.3 REDE GENÉTICA

A rede neural genética foi construída por meio de um banco de dados que contém todos os genes mutados de todos os pacientes, a tabela de dados que foi utilizada para gerar esta rede continha apenas os genes e o status vital, de forma que as entradas na rede foram realizadas por cada gene mutado individualmente. A finalidade foi de avaliar, por meio do status vital apresentado ao longo do treinamento, se as mutações em um determinado gene seriam capazes de prever o status vital de um paciente. Na implementação dessa rede foram utilizadas 3 camadas ocultas (Gráfico 6) e os grupos de dados de treinamento e de teste apresentaram respectivamente aproximadamente 75% e 25% dos dados.

Gráfico 6: Rede neural genética. Círculos com letras I, B, H e O representam respectivamente variável de entrada, biases, camadas ocultas, e saída da rede e os números após as letras diferenciam as camadas. Linhas representam as interações entre as camadas. As Linhas pretas representam pesos positivos e as cinzas representam pesos negativos.



5.4 AVALIAÇÃO DA PREDIÇÃO DAS REDES

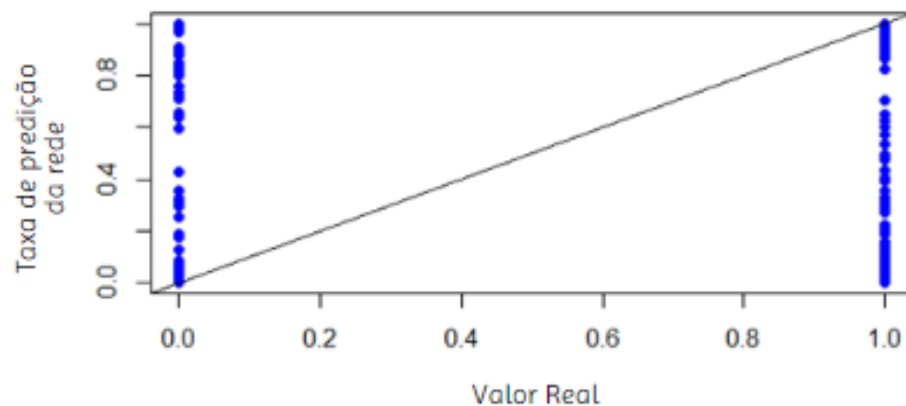
Para verificar a performance da predição das redes neurais implementadas foram calculados os parâmetros classificações corretas e incorretas, acurácia e erro médio quadrático (MSE), presentes no Quadro 1.

Como pode ser visualizado a predição da primeira rede se deu de forma praticamente aleatória apresentando a acurácia de 52,92%. Os resultados desta rede também podem ser observados no Gráfico 7, no qual são representados os dados reais contra os dados preditos, pode-se verificar no gráfico a presença dos pontos azuis (pacientes) possuindo predição $> 0,5$, quando o valor real = 0, bem como pontos possuindo predição $< 0,5$, quando o valor real = 1.

Quadro 1: Performance das redes neurais artificiais. Nas quais as variáveis são para a rede Internacional com 20 variáveis: idade, etnia, gênero, raça, ano de nascimento, classificação de tumores malignos (TNM), localização do tumor primário, morfologia, primeiro diagnóstico, malignidade prévia, tratamento prévio, malignidade sincronizada, local da biópsia, tecido ou órgão de origem, estágio do tumor, ano do diagnóstico, número de cigarros por dia e maços de cigarro por ano. Na rede Internacional com 8 variáveis: idade, etnia, gênero, T, M, malignidade prévia e malignidade sincronizada. Na rede internacional com 13 variáveis: idade, etnia, gênero, raça, ano de nascimento, classificação de tumores malignos (TNM), localização do tumor primário, morfologia, primeiro diagnóstico, estágio do tumor e local da biópsia. Na rede Mista: idade, gênero, raça, classificação de tumores malignos (TNM), localização do tumor primário e morfologia. E na rede Genética: Gene.

Critérios de Avaliação	Redes Clínicas				Rede genética
	Internacional com 20 variáveis	Internacional com 8 variáveis	Internacional com 13 variáveis	Mista	
Classificações corretas	163	104	173	78	28721
Classificações incorretas	145	104	28	93	70422
Acurácia (%)	52,92%	50%	86,06%	45,61%	28,96 %
MSE (%)	47,70%	50%	13,93%	--	71,03%

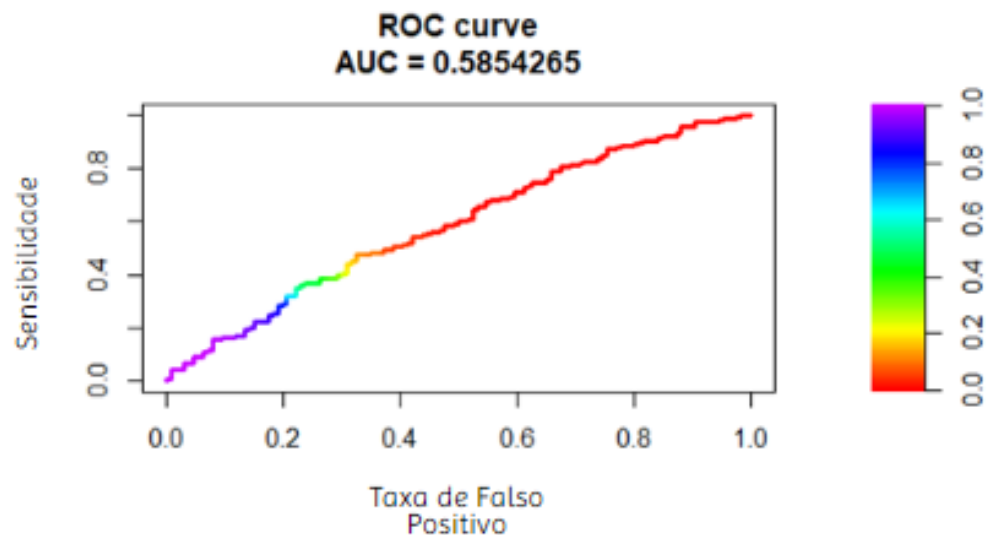
Gráfico 7: Taxa real contra taxa predita pela rede clínica com todas as variáveis. Pontos Azuis representam cada entrada na rede.



Além disso é possível visualizar no Gráfico 8 a curva ROC referente a esta rede, na qual a área sob a curva (AUC) é 0,5854265. De acordo com o valor AUC obtido por esta rede é possível verificar que esta é uma curva ROC não discriminante, pois o resultado da AUC padronizado é 1 para uma curva ROC perfeita e 0,5 para uma curva ROC não discriminante (ROBIN *et al*, 2011). Visualizando a matriz de confusão desta rede no Quadro 2, é evidente que a maior quantidade de imprecisão da rede ocorreu devido a predição de morte para pacientes vivos.

Alguns motivos que podem justificar esse resultado são a quantidade de variáveis de entrada, a arquitetura da rede, devido a presença de apenas uma camada oculta, ou ainda a ausência de dados em algumas variáveis. Segundo GOLDSTEIN e col. (2017) a ausência ou perda de dados em registros de saúde eletrônicos são um grande desafio para implementação de modelos de predição, e tal perda pode ocorrer devido a desorganização dos dados ou mesmo a perda de dados que podem ocorrer, por exemplo, caso um paciente seja readmitido em outros lugares, de maneira que parte dos dados não sejam computados.

Gráfico 8: Curva de características operacionais do receptor da rede clínica com todas as variáveis. AUC refere-se a área sob a curva.



Quadro 2: Matriz de Confusão das redes. Dados em valores reais e em porcentagem. Nas quais as variáveis são para a rede Internacional com 20 variáveis: idade, etnia, gênero, raça, ano de nascimento, classificação de tumores malignos (TNM), localização do tumor primário, morfologia, primeiro diagnóstico, malignidade prévia, tratamento prévio, malignidade sincronizada, local da biópsia, tecido ou órgão de origem, estágio do tumor, ano do diagnóstico, número de cigarros por dia e maços de cigarro por ano. Na rede Internacional com 8 variáveis: idade, etnia, gênero, T, M, malignidade prévia e malignidade sincronizada. Na rede internacional com 13 variáveis: idade, etnia, gênero, raça, ano de nascimento, classificação de tumores malignos (TNM), localização do tumor primário, morfologia, primeiro diagnóstico, estágio do tumor e local da biópsia. Na rede Mista: idade, gênero, raça, classificação de tumores malignos (TNM), localização do tumor primário e morfologia. E na rede Genética: Gene.

REDE NEURAL ARTIFICIAL	Vivo	Morto	Classe Real
Internacional com 20 variáveis	66 (21,45%)	116 (37,66%)	Vivo
	29 (9,41%)	97 (31,50%)	Morto
Internacional com 8 variáveis	28 (13,46%)	94 (45,19%)	Vivo
	10 (4,8%)	76 (36,53%)	Morto
Internacional com 13 variáveis	95 (47,26%)	28 (13,9%)	Vivo
	0	78 (38,8%)	Morto
Mista	30 (17,54%)	47 (27,48%)	Vivo
	46 (26,9%)	48 (28%)	Morto
Genética	0	70422 (71%)	Vivo
	0	28721 (29%)	Morto

Por meio da identificação das possíveis causas da aleatoriedade na predição da primeira rede neural foram selecionadas as oito variáveis mais importantes na primeira predição para realização da segunda rede clínica. De acordo com o Quadro 1, a segunda rede foi capaz de prever 50% dos dados, tendo assim uma acurácia ainda pior do que a obtida na primeira rede, observando-se todos os parâmetros avaliados a segunda rede assemelhou-se muito a primeira rede na curva ROC, a AUC (Gráfico 9) e na matriz de confusão (Quadro 2), na qual o maior erro também foi para pacientes que estão vivos, mas foram preditos como mortos, o que pode ser visualizado no Gráfico 10.

Gráfico 9: Curva de características operacionais do receptor da rede clínica com oito variáveis. AUC refere-se a área sob a curva.

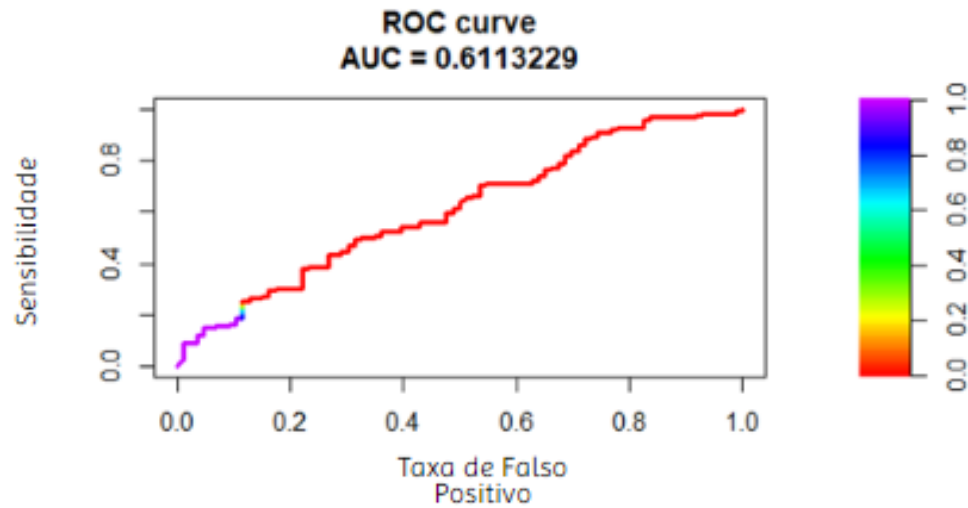
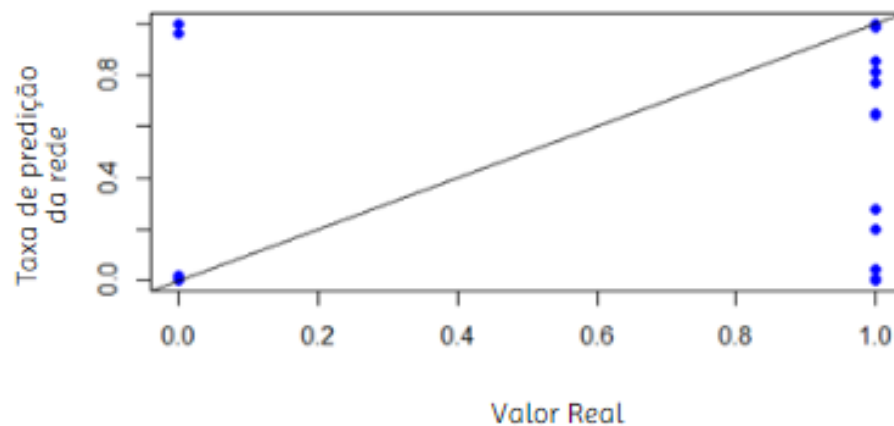


Gráfico 10: Taxa real contra taxa predita pela rede clínica com oito variáveis. Pontos Azuis representam cada entrada na rede.



A escolha das variáveis pela influência na predição da primeira rede não foi capaz de melhorar os resultados da segunda rede, então foi construída uma terceira rede. Nesta rede as variáveis foram escolhidas por meio de teste e observação escolhendo como a rede final o modelo que obtivesse um resultado melhor. Para melhorar o resultado na terceira rede foram removidos pacientes que possuísem variáveis vazias, resultando em um conjunto com 669 indivíduos.

Devido ao refinamento das variáveis e a utilização do conjunto inteiro como dado de treinamento, é possível visualizar uma melhora na performance dessa rede em relação as duas redes anteriores pois a mesma apresentou uma acurácia de 86,06%, de acordo com o Gráfico 11, a predição foi pouco dispersa e não foi observado nenhum caso no qual o real = 0 e o predito >

0,5, tal fato é confirmado pela matriz de confusão (Quadro 2), na qual para esta rede não foi observado falso positivo. Já em relação a curva ROC (Gráfico 12), esta rede apresentou a $AUC = 0,8465708$, ou seja, um valor mais próximo de uma curva ROC perfeita do que de uma curva não discriminatória.

Gráfico 11: Taxa real contra taxa predita pela rede clínica com treze variáveis. Pontos Azuis representam cada entrada na rede.

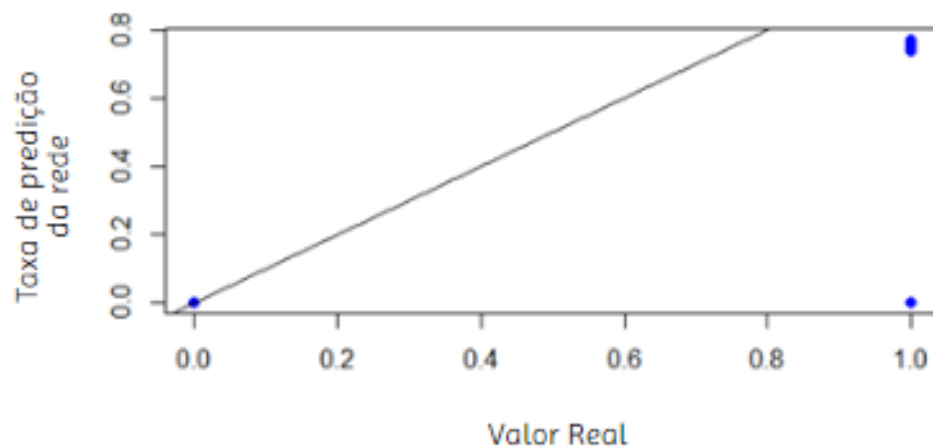
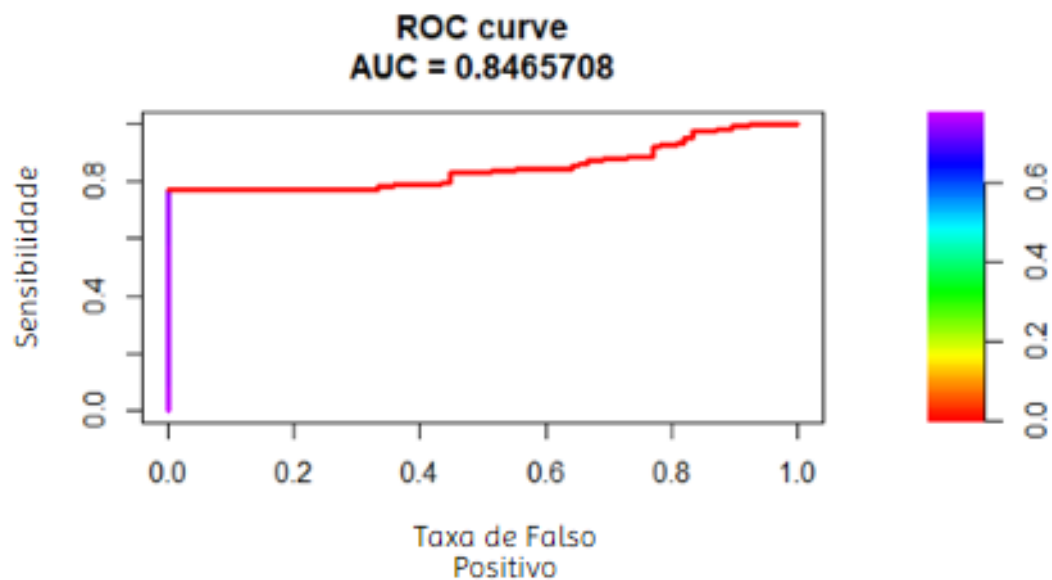


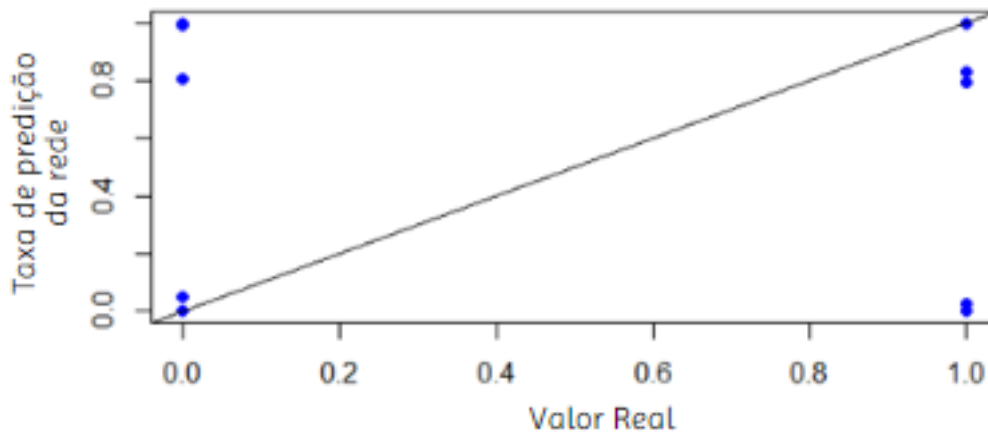
Gráfico 12: Curva de características operacionais do receptor da rede clínica com treze variáveis. AUC refere-se a área sob a curva.



A rede contendo dados internacionais e locais foi a rede clínica que apresentou a pior predição, possuindo a acurácia de 45,61% (Quadro 1), além disso na Quadro 2 é possível observar que as taxas de falso positivo, falso negativo, verdadeiro positivo e verdadeiro negativo foram mais próximas umas das outras, quando comparadas as outras redes clínicas que apresentaram uma maior taxa de falso negativo. Também se observa que diferentemente das outras redes o gráfico de real contra predito (Gráfico 13) não apresentou muita dispersão mesmo que a predição tenha errado em mais de 50% dos casos, o que pode ter ocorrido devido a esta rede ter o menor conjunto de dados teste. Tal resultado, provavelmente, ocorre devido a três principais motivos são eles, a ausência de dados de alguns indivíduos para as variáveis utilizadas, em ambos os grupos de dados, a baixa quantidade variáveis comuns aos dois bancos que possuísem um bom volume de dados presentes e as diferenças genéticas entre as duas populações tendo em vista que os dados do grupo treinamento pertence a um país diferente dos dados do grupo teste. Este último motivo é de grande importância pois existem diferenças importantes na epidemiologia do câncer de pulmão entre países mais desenvolvidos e menos desenvolvidos (YOULDEN *et al*, 2008).

Observou-se que algumas variáveis apresentaram problemas, como por exemplo a variável raça, na qual o banco internacional possuía a classificação negro, branco, asiático e nativo americano, enquanto o banco de dados local classificava-se em negro, branco e pardo, de tal maneira que a classificação nativo americano e pardo não eram comuns entre os grupos. Ou ainda na variável tipos morfológicos do tumor, na qual houve diversos tipos não comuns entre os dois conjuntos de dados, com isso podendo ocorrer por causa das diferentes características genéticas ou de exposição a fatores externos entre as populações.

Gráfico 13: Taxa real contra taxa predita pela rede mista. Pontos Azuis representam cada entrada na rede.



A rede genética foi a rede com pior desempenho no estudo, possuindo a acurácia de 28,96%, a taxa foi baixa devido a rede ter classificado todos os pacientes do grupo teste como mortos, quando na verdade a quantidade de pacientes vivos no conjunto de dados era cerca de 70%. O mesmo resultado pode ser observado no Gráfico 14 e Gráfico 15, os quais apresentam, respectivamente, todos os pacientes com valores de predição abaixo de 0,5, ou seja, que a rede considera como mortos, e uma AUC próximo a 0,5, caracterizando dessa forma uma curva não discriminatória. Portanto, tais dados demonstram que a rede genética sozinha não foi capaz de prever o status vital do paciente. Isso pode ter ocorrido devido ao volume de mutações genéticas recuperadas que foi cerca de 20.000 tipos diferentes de mutações em um conjunto de quase 400.000 entradas, ou pela ausência da influência de outras variáveis, uma vez que nessa rede apenas foram usados como entrada apenas as mutações. No entanto, é possível que quando associada com variáveis clínicas esta rede obtenha um desempenho mais satisfatório do que as redes clínicas e genéticas separadamente, pois ainda que modelos clínicos sejam utilizados para prever o prognóstico de pacientes os avanços no entendimento da base genética pode refinar a capacidade de prever como os pacientes podem ser individualmente impactados pela doença (BEJAR, 2014).

Gráfico 14: Taxa real contra taxa predita pela rede genética. Pontos Azuis representam cada entrada na rede.

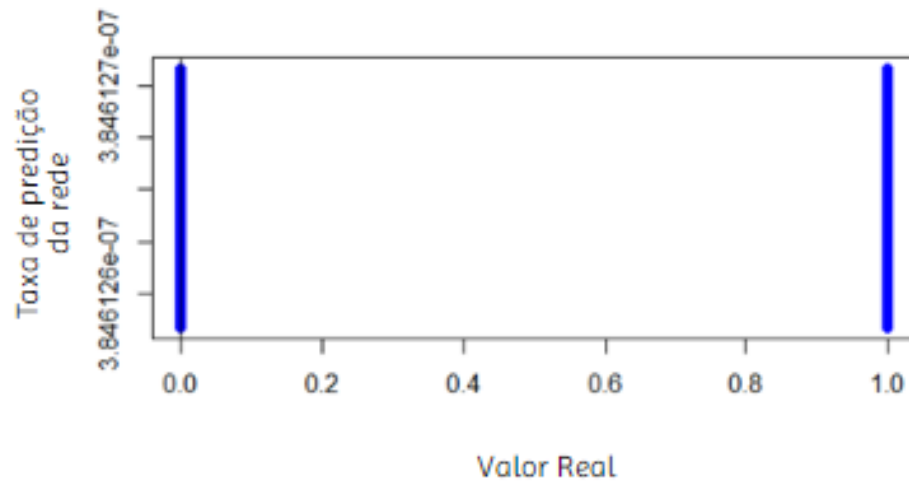
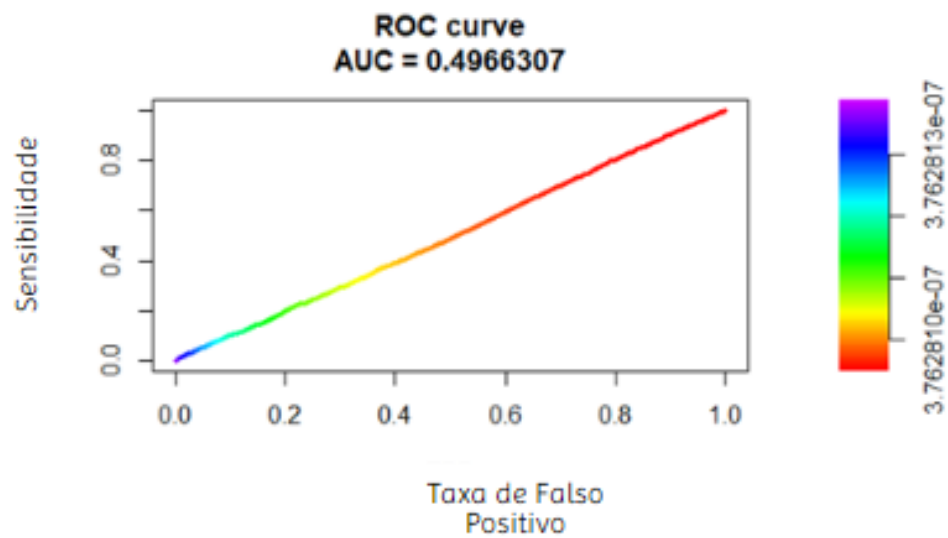


Gráfico 15: Curva de características operacionais do receptor da rede genética. AUC refere-se a área sob a curva.



6. CONSIDERAÇÕES FINAIS

As bases de dados clínicos e genéticos de câncer podem ser utilizadas para diversos estudos que visem a utilização de algoritmos de aprendizado de máquina na predição do status vital dos pacientes. A condensação dos dados clínicos e genéticos mostrou-se uma abordagem útil para a formação de um banco de dados com todas as informações dos pacientes, e poderá ser utilizada para análises futuras utilizando esta base de dados completa sem necessidade de realizar o processo de recuperação e organização desses dados novamente. A metodologia implementada para recuperação de um grande volume de dados genéticos pode ser utilizada para outras neoplasias, bem como é possível realizar a mesma metodologia para a construção de bancos de dados mais completos.

As redes neurais desenvolvidas não apresentaram performances satisfatórias para a predição do status vital dos pacientes, com exceção da rede com treze variáveis que apresentou boa precisão na predição do status vital. Portanto são necessários mais testes para avaliar se os resultados negativos se devem a arquitetura da rede, à ausência de dados influenciando o treinamento, ou se as predições tiveram baixa precisão devido a variáveis que tenham influenciado fortemente a predição, mas que estão menos associadas ao status vital dos pacientes.

Além disso há ainda a possibilidade de futuramente ser realizada a construção de uma rede neural que contenha dados clínicos e genéticos, considerando que sua associação possa ocasionar a obtenção de melhores resultados.

REFERÊNCIAS

- AGRAWAL, S.; AGRAWAL, J. Neural Network Techniques for Cancer Prediction: A Survey. **Procedia Computer Science**. v. 60, p.769-774, 2015.
- BECK M. W. NeuralNetTools: Visualization and Analysis Tools for Neural Networks. **Journal of statistical software**. v.85, n.11, p.1–20, 2018.
- BEJAR, R. Clinical and genetic predictors of prognosis in myelodysplastic syndromes. **Haematologica**. v. 99, n.6, p.956-964, 2014.
- BERTOLACCINI, L.; SOLLI, P.; PARDOLESI, A.; PASINI, A. An overview of the use of artificial neural networks in lung cancer research. **Journal of Thoracic Disease**. v. 9, n. 4, p.924-931, 2017.
- BIBAULT, J.; GIRAUD, P.; BURGUN, A. Big Data and machine learning in radiation oncology: State of the at and future prospects. **Cancer Letters**. v. 382, n. 1, p.110-117, 2016.
- CIABURRO, G.; VENKATESWARAN, B. **Neural Networks with R, Smart models using CNN, RNN, deep learning, and artificial intelligence principles**. Birmingham: Packt Publishing, 2017. 314p.
- CHAURASIA, V.; PAL, S. Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability. **International Journal of Computer Science and Mobile Computing**. v.3, n.1, p.10-22, 2017.
- CHEN, C. L. P.; ZHANG, C. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. **Information Sciences**. v.275, p.314-347, 2014.
- CHEN, Y.; CHANG, Y.; KE, W.; CHIU, H. Cancer adjuvant chemotherapy strategic classification by artificial neural network with gene expression data: An example for non-small cell lung cancer. **Journal of Biomedical Informatics**. v. 56, p. 1-7, 2015.

EL-TELBANY, A.; PATRICK C. M. Cancer Genes in Lung Cancer Racial Disparities: Are There Any?. **Genes Cancer**, v. 3 (7-8), p. 467-480, 2012.

FERLAY, J.; SOERJOMATARAM, I.; DIKSHIT, R.; ESER, S.; MATHERS, C.; REBELO, M.; PARKIN, D. M.; FORMAN, D.; BRAY, F. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. **International Journal of Cancer**. v.136, p. 359-386, 2014.

GDC. **Genomic Data Commons Data Portal**, 2020. Disponível em:

<<https://portal.gdc.cancer.gov/>>. Acesso em: 25 Jan 2020

GOLDSTEIN, B. A.; NAVAR, A. M.; PENCINA, M. J.; ALOANNIDIS, J. P. A. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. **Journal of the American Medical Informatics Association**. v. 24, n. 1, p.198–208, 2017.

GROSSMAN, R. L.; HEATH, A. P.; FERRETTI, V.; VARMUS, H. E.; LOWY, D. R.; KIBBE, W. A.; STAUDT, L. M. Toward a Shared Vision for Cancer Genomic Data. **The New England Journal of Medicine**. v. 375, p. 1109-1112, 2016.

HAYKIN, S. **Redes Neurais: Princípios e prática**. 2.ed. Porto Alegre: Bookman, 2001. 900p.

HUANG, L.; FERNANDES, H.; ZIA, H.; TAVASSOLI, P.; RENNERT, H.; PISAPIA, D.; IMIELINSKI, M.; SBONER, A.; RUBIN, M. A.; KLUK, M.; ELEMENTO, O. The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. **Journal of the American Medical Informatics Association**. v.24, n.3, p.513-519, 2017.

IARC. **World cancer report 2008**. Lyon: International Agency for Research on Cancer, 2008. 511p.

INCA. **ABC do Câncer: Abordagens Básicas para o Controle do Câncer**. 3. ed. Rio de

Janeiro: INCA, 2017. 108p.

INCA. **Câncer: o que é câncer?**, 2019. Disponível em: <<https://www.inca.gov.br/o-que-e-cancer>>. Acesso em: 16 jan. 2019.

JOTHI, N.; RASHID, N. A.; HUSAIN, W. Data Mining in Healthcare - A Review. **Procedia Computer Science**, v.72, p. 306-313, 2015

KOUROU, K.; EXARCHOS, T. P.; EXARCHOS, K. P.; KARAMOUZIS, M. V.; FOTIADIS, D. I. Machine learning applications in cancer prognosis and prediction. **Computational and Structural Biotechnology Journal**. v. 13, p.8-17, 2015.

KRISHNAIAH, V.; NARSIMHA, G.; CHANDRA, S. N. Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. **International Journal of Computer Science and Information Technologies**, v. 4 (1), p. 39-45, 2013.

LEE, C. K.; WU, Y.; DING, P. N.; LORD, S. J.; INOUE, A.; ZHOU, C.; MITSUDOMI, T.; ROSELL, R.; PAVLAKIS, N.; LINKS, M.; GEBSKI, V.; GRALLA, R. J.; YANG, J. C. Impact of Specific Epidermal Growth Factor Receptor (EGFR) Mutations and Clinical Characteristics on Outcomes After Treatment With EGFR Tyrosine Kinase Inhibitors Versus Chemotherapy in EGFR-Mutant Lung Cancer: A Meta-Analysis. **Journal of Clinical Oncology**. v. 33, n. 17, p. 1958-1965, 2015.

LEE, C. K.; MAN, J.; LORD, S.; LINKS, M.; GEBSKI, V.; MOK, T.; YANG, J. C. Checkpoint Inhibitors in Metastatic EGFR-Mutated Non-Small Cell Lung Cancer - A Meta-Analysis. **Journal of Thoracic Oncology**. v. 12, n.2, p. 403-407, 2017.

MINISTÉRIO DA SAÚDE. **Protocolos Clínicos e Diretrizes Terapêuticas em Oncologia**. 1.ed. Brasília: Secretaria de Atenção à Saúde: Ministério da Saúde, 2014. 356 p.

RENAULT, I. Z. Medicina personalizada e câncer: Os avanços na oncogenética. **Inovar Saúde**, Edição Especial.v.1, p.6-10, 2017.

ROBIN, X.; TURCK, N.; HAINARD, A. pROC: an open-source package for R and S+ to analyze and compare ROC curves. **BMC Bioinformatics**. v. 12, n. 77, 2011.

SHOLL, L. M.; AISNER, D. L.; VARELLA-GARCIA, M.; BERRY, L. D.; DIAS-SANTAGATA, D.; WISTUBA, I. I.; CHEN, H.; FUJIMOTO, J.; KUGLER, K; FRANKLIN W. A.; IAFRATE, A. J.; LADANYI, M.; KRIS, M. G.; JOHNSON, B. E.; BUNN, P. A.; MINNA, J. D.; KWIATKOWSKI, D. J. Multi-institutional Oncogenic Driver Mutation Analysis in Lung Adenocarcinoma: The Lung Cancer Mutation Consortium Experience. **Journal of Thoracic Oncology**. v. 10, n.5, p. 768–777, 2015.

TORRE, L. A.; BRAY, F.; SIEGEL, R. L.; FERLAY, J.; LORTET-TIEULENT, J.; JEMAL, A. Global Cancer Statistics, 2012. **CA: A Cancer Journal for Clinicians**. v. 65, n.2, p.87-108, 2015.

TURBAN, E.; SHARDA, R.; ARONSON, J. E.; KING, D. **Business Intelligence: Um enfoque gerencial para a inteligência do negócio**. São Paulo: Bookman, 2009. 245p.

VAVALÀ, T.; LEVRA, G. M.; NOVELLO, S. Lung cancer in never smokers: a different disease. **Current Respiratory Care Reports**. v. 3, n.1, p. 26–34, 2014.

WANG, A.; WANG, H. Y.; LIU, Y.; ZHAO, M. C.; ZHANG, H. J.; LU, Z. Y.; FANG, Y. C.; CHEN, X. F.; LIU, G. T. The prognostic value of PD-L1 expression for non-small cell lung cancer patients: A meta-analysis. **European Journal of Surgical Oncology**. v. 41, n. 4, p. 450-456, 2015.

WANG, J.; LIU, Q.; YUAN, S.; XIE, W.; LIU, Y.; XIANG, Y.; WU, N.; WU, L.; MA, X.; CAI, T.; ZHANG, Y.; SUN, Z.; LI, Y. Genetic predisposition to lung cancer: comprehensive literature integration, meta-analysis, and multiple evidence assessment of candidate-gene association studies. **Scientific Reports**, v. 7, n.8371, 2017.

WHO. **Guide to cancer early diagnosis**. World Health Organization. Geneva, 2017. Disponível

em: <<https://apps.who.int/iris/bitstream/handle/10665/254500/9789241511940-eng.pdf>>. Acesso em: 16 jan. 2019.

YOULDEN, D. R.; CRAMB, S. M.; BAADE, P. D. The International Epidemiology of Lung Cancer: Geographical Distribution and Secular Trends. **Journal of Thoracic Oncology**. v. 3, n. 8, p. 819-831, 2008.