UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE ENGENHARIAS E TECNOLOGIA
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

FRANCISCO LEONÉSIO CARNEIRO DUARTE

ANÁLISE DO EMPREGO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NO DIAGNÓSTICO DO CÂNCER DE MAMA ATRAVÉS DE DADOS COLETADOS EM EXAMES DE ROTINA

PAU DOS FERROS
2019

FRANCISCO LEONÉSIO CARNEIRO DUARTE

ANÁLISE DO EMPREGO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NO DIAGNÓSTICO DO CÂNCER DE MAMA ATRAVÉS DE DADOS COLETADOS EM EXAMES DE ROTINA

Monografia apresentada a Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Pedro Thiago Valério de Souza Prof. Me.

Co-orientadora: Náthalee Cavalcanti de Almeida Lima, Profª. Drª.

PAU DOS FERROS

2019

Ficha catalográfica elaborada pelo Sistema de Bibliotecas
da Universidade Federal Rural do Semi-Árido, com os dados fornecidos pelo(a) autor(a)

D812a Duarte, Francisco Leonésio Carneiro.
Análise do emprego de técnicas de aprendizado de máquina no diagnóstico do câncer de mama através de dados coletados em exames de rotina / Francisco Leonésio Carneiro Duarte. - 2019.
50 f. : il.

Orientador: Pedro Thiago Valério de Souza.
Coorientadora: Náthalee Cavalcanti de Almeida Lima.
Monografia (graduação) - Universidade Federal Rural do Semi-árido, Curso de Engenharia de Computação, 2019.

1. Câncer de Mama. 2. Inteligência Artificial. 3. Classificação de Padrões. 4. Exames de Sangue. 5. Diagnóstico Auxiliado por Computador. I. Souza, Pedro Thiago Valério de, orient. II. Lima, Náthalee Cavalcanti de Almeida, co-orient. III. 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.

FRANCISCO LEONÉSIO CARNEIRO DUARTE

ANÁLISE DO EMPREGO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NO DIAGNÓSTICO DO CÂNCER DE MAMA ATRAVÉS DE DADOS COLETADOS EM EXAMES DE ROTINA

Monografia apresentada a Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Engenharia de Computação.

Defendida em: 07/08/2019.

BANCA EXAMINADORA

Pedro Thiago Valério de Souza
Pedro Thiago Valério de Souza, Prof. Me. (UFERSA)
Presidente

Francisco Carlos Gurgel da Silva Segundo
Francisco Carlos Gurgel da Silva Segundo, Prof. Dr. (UFERSA)
Membro Examinador

Marco Diego Aurélio Mesquita
Marco Diego Aurélio Mesquita, Prof. Me. (UERSA)
Membro Examinador

AGRADECIMENTOS

Agradeço a Deus por ter me ajudado a superar as dificuldades.

A universidade e ao seu corpo docente e administrativo por ter proporcionado um bom ambiente de estudo.

Ao professor e orientador Pedro Thiago Valério de Souza, pelo seu apoio no desenvolvimento e correção do trabalho.

A professora e co-orientadora Náthalee Cavalcanti de Almeida Lima pelo apoio nas correções e escolha da temática do trabalho.

Aos meus pais, pelo amor, incentivo e apoio.

E a todos que, direta ou indiretamente, fizeram parte da minha formação.

RESUMO

O câncer de mama constitui uma perigosa doença, que quando diagnosticada em estágios iniciais possui grandes chances de cura. Desta forma, torna-se necessário investir em pesquisas relacionadas ao diagnóstico desta doença. Nesta vertente, este trabalho tem por objetivo avaliar alguns modelos de classificação: Máquina de Vetores de Suporte, KNN, Rede Neural de Função de Base Radial e Rede *Perceptron* de Múltiplas Camadas, empregados em uma base de dados que apresenta diagnósticos para o câncer de mama juntamente com alguns dados obtidos em exames de rotina. Realizou-se a divisão da base de dados em conjuntos de validação e treinamento de forma a validar a qualidade de previsão de cada modelo. Variou-se também a quantidade de entradas dos modelos, bem como aplicou-se a técnica de análise de componentes principais. Com a finalização do trabalho, foi possível observar que um modelo baseado em rede neural de função de base radial obteve o melhor desempenho, com os seguintes respectivos intervalos de confiança para sensibilidade, especificidade e acurácia com 95% de probabilidade e 34 graus de liberdade: [80,30-87,09], [79,03-86,68], [80,84-85,72]. Realizou-se uma comparação com outros trabalhos da literatura, sendo constatado uma obtenção de desempenho compatível com eles.

Palavras-chave: Câncer de Mama, Inteligência Artificial, Classificação de Padrões, Exames de Sangue, Diagnóstico Auxiliado por Computador.

ABSTRACT

Breast cancer is a dangerous disease, which when diagnosed in the early stages has a great chance of cure. Thus, it is necessary to invest in research related to the diagnosis of this disease. In this section, the objective of this study is to evaluate some classification models: Support Vector Machine, KNN, Radial Base Function Network and Multilayer Perceptron Network, used in a database that presents diagnoses for breast cancer along with some data obtained in routine exams. The database was split into test and training sets in order to validate the predictive quality of each model. The number of entries of the models was also varied, as well as the principal component analysis technique. With the completion of the work, it was possible to observe that a neural network model with a radial base function obtained the best performance, with the following confidence intervals for sensitivity, specificity and accuracy with 95% probability and 34 degrees of freedom: [80.30-87.09], [79.03-86.68], [80.84-85.72]. A comparison with other works of the literature was made, and compatible performance was verified.

Keywords: Breast Cancer, Artificial Intelligence, Pattern Classification, Blood Tests, Computer Assist Diagnostics.

LISTA DE FIGURAS

Figura 1: Representação do Funcionamento do Algoritmo KNN 20
Figura 2: Representação do Modelo da Rede PMC 22
Figura 3: Representação do Modelo da Rede RBF 23
Figura 4: Representação da Classificação Realizada pela SVM 25
Figura 5: Exemplificação da Utilização do PCA 26
Figura 6: Exemplificação da Curva ROC 28
Figura 7: Matriz de Confusão do Melhor Modelo 37
Figura 8: Curva ROC do Melhor Modelo Obtido 38
Figura 9: Curva de Treinamento do Melhor Modelo 39
Figura 10: Tela Inicial do RBFDIAG 42

LISTA DE TABELAS

Tabela 1: Resumo do Estado da Arte 13
Tabela 2: Representação da Qualidade de Um Exame 26
Tabela 3: Associação Entre a Área da Curva ROC e a Qualidade do Modelo 29
Tabela 4: Resultados da PMC 30
Tabela 5: Resultados da MVS 32
Tabela 6: Resultados do KNN 33
Tabela 7: Resultados da RBF Com Normalização mapminmax 34
Tabela 8: Resultados da RBF Com Normalização z-scores 35
Tabela 9: Comparação de Intervalos 36
Tabela 10: Comparação Entre Os Modelos de Melhor Desempenho 39

LISTA DE ABREVIATURAS E SIGLAS

INCA	Instituto Nacional do Câncer
WHO	*World Health Organization*
PMC	*Perceptron* de Múltiplas Camadas
RBF	*Radial Basis Function*
KNN	*K-Nearest Neighbors*
MVS	Máquina de Vetores de Suporte
IMC	Índice de Massa Corporal
HOMA	*Homeostatic Model Assessment*
MCP-1	*Monocyte Chemoattractant Protein-1*
PCA	*Principal Component Analysis*
PAHO	*Pan American Health Organization*
ROC	*Receiver Operating Characteristic*
OMS	Organização Mundial da Saúde

SUMÁRIO

1 INTRODUÇÃO 11

1.1 Estado da Arte 12
1.2 Motivações 14
1.3 Contribuições 14
1.3.1 Contribuição Geral 14
1.3.2 Contribuições Específicas 14
1.4 Metodologia 15
1.5 Organização do Texto 16

2 ARQUITETURA DO CLASSIFICADOR, DAS SUAS ENTRADAS E DAS SUAS MÉTRICAS DE DESEMPENHO 17

2.1 Descrição da Base de Dados 17
2.2 Definição de um Classificador de Padrões 18
2.3 Classificadores Avaliados 19
2.3.1 *K-Nearest Neighbors* 19
2.3.2 *Perceptron* de Múltiplas Camadas 21
2.3.3 Rede Neural de Função de Base Radial 22
2.3.4 Máquina de Vetores de Suporte 24
2.4 Análise de Componentes Principais 25
2.5 Métricas Analisadas 26

3 RESULTADOS E DISCUSSÕES 30

3.1 Análise para a Rede PMC 30
3.2 Análise para a Rede MVS 31
3.3 Análise para o KNN 33
3.4 Análise para a Rede RBF 34
3.5 Análise para o Classificador Proposto 36
3.6 O *Software* RBFDIAG 41

4 CONCLUSÕES 43

4.1 Trabalhos Futuros 43
4.2 Trabalhos submetidos 44
REFERÊNCIAS 45
ANEXO A CONJUNTO DE DADOS DE VALIDAÇÃO PARA O PROGRAMA DESENVOLVIDO 49

1 INTRODUÇÃO

De acordo com INCA (2018), o câncer de mama é um grupo de doenças heterogêneas, possuindo vários tipos de manifestações e consequentemente diferentes respostas terapêuticas. A forma histológica mais comum desta doença é o carcinoma ductal infiltrante, sendo que os principais sintomas que podem ser associados são: aparecimento de nódulo (geralmente indolor, duro e irregular), e tumores (consistência branca, globosos e bem definidos).

A Organização Mundial da Saúde (OMS), WHO (2018), por meio de dados de estimativa padronizada por idade das taxas de incidência e mortalidade de cada tipo de câncer, ressalta que o câncer de mama é o que possui a maior taxa de incidência, se comparado aos outros tipos de câncer. Esta taxa é atualmente de 46,3, com uma respectiva taxa de mortalidade de 13,0, que é inferior apenas à taxa de mortalidade do câncer de pulmão.

Por meio dos dados apresentados anteriormente, verifica-se a gravidade e a importância de se estudar este tipo de câncer. Neste contexto, é relevante ressaltar a necessidade de detectar esta doença o mais cedo possível, uma vez que: “O diagnóstico correto em um estado prematuro do câncer de mama pode auxiliar na tomada de decisões, no planejamento de ações e, evidencialmente, na eficiência do tratamento.” (SILVA *et al*., 2014, p. 1).

Gonçalves (2017) afirma que normalmente os primeiros passos para detecção do câncer de mama são iniciados por meio do autoexame e do exame clínico, sendo ambos procedimentos manuais. Se for detectado alguma anomalia, normalmente o paciente é submetido a uma mamografia, que por sua vez, segundo Borchartt (2013), pode apresentar falhas na identificação da doença em mamas densas, devido a retenção da radiação, o que pode também aumentar o risco do desenvolvimento do câncer. Radiocentro (2011) afirma que a dose média efetiva de radiação em uma mamografia é de 0,4 mSv, o que corresponde a quase metade do limite de radiação anual tolerado desconsiderando a radiação recebida de fontes naturais (1 mSv).

Neste sentido, sabendo-se da complexidade do processo de diagnóstico desta doença, percebe-se a relevância em buscar alternativas e novas metodologia que possam apoiar este processo. Santos *et al*. (2016) aponta que com o surgimento das bases de dados biomédicas, a área de CAD (*Computed-Aided Diagnosis*) vem se desenvolvendo cada vez mais tornando possível a construção de sistemas computacionais de classificação, capazes de apoiar este procedimento de diagnóstico.

Tendo em vista o exposto, este trabalho tem por objetivo avaliar quatro sistemas de classificação desenvolvidos com uma rede neural *Perceptron* de Múltiplas Camadas (PMC), uma Rede Neural de Funções de Base Radial (*Radial Base Function* - RBF), o algoritmo K-

Vizinhos Mais Próximos (*K-Nearest Neighbors* - KNN) e a Máquina de Vetores de Suporte (MVS), os quais com base nos dados de entrada:

1. Idade;
2. Índice de Massa Corporal (IMC);
3. Glicose;
4. Insulina;
5. HOMA (*Homeostatic Model Assessment*);
6. Leptina;
7. Adiponectina;
8. Resistina;
9. MCP-1 (*Monocyte Chemoattractant Protein-1*).

Classificam se o paciente possui ou não câncer de mama. O treinamento e validação dos modelos foi realizado com os dados do *Breast Cancer Coimbra Data set*, disponibilizado por Patrício *et al*. (2018). Será realizado uma comparação de desempenho entre os classificadores empregados, bem como com relação a resultados obtidos em outros trabalhos. De modo a tentar buscar melhores resultados será aplicado a análise de componentes principais (*Principal Component Analysis* - PCA), bem como a seleção de subconjuntos de variáveis de entrada, com o objetivo de diminuir as dimensões do problema.

1.1 Estado da Arte

Na literatura, é possível encontrar várias tentativas de aplicação de técnicas de aprendizado de máquina para o diagnóstico do câncer de mama, com base em alguma fonte de dados.

Como exemplo, Andrade *et al.* (2017) realizaram a classificação de termogramas da mama, em saudáveis ou anômalos. Para tanto, eles extraíram algumas características das imagens e aplicaram os classificadores *K-Star* e máquina de vetores de suporte. O modelo que apresentou o melhor resultado foi o *K-Star*, com uma acurácia de 95,80 %, sensibilidade de 93,60 % e especificidade de 95,90 %.

Holsbach *et al.* (2014) efetuaram a classificação de amostras de células de tumores da mama, em benigno ou maligno. Para realizar esta tarefa eles aplicaram a ferramenta de k-vizinhos mais próximos juntamente com análise discriminante e eliminação de variáveis com menor índice de importância. Os melhores resultados obtidos para acurácia, sensibilidade e especificidade foram respectivamente de 97,77 %, 97,90 % e 98,56 %. Foi utilizado o banco de dados *Wisconsin Brest Cancer Database*.

Silva *et al.* (2014), utilizando-se da mesma base dados utilizada por Holsbach *et al.* (2014) e de uma rede neural ARTMAP-Fuzzy obtiveram em seu melhor resultado uma acurácia de 94,43 %, sensibilidade de 100 % e especificidade de 91,44 %.

Por sua vez, Patrício *et al.* (2018) utilizando-se de dados que podem ser coletados em exames de rotina, criou modelos para diagnóstico do câncer de mama. Os modelos testados foram: máquina de vetores de suporte, regressão logística e florestas aleatórias. O melhor modelo foi a máquina de vetores de suporte cujas entradas eram glicose, resistina, idade e índice de massa corporal. Este modelo apresentou os seguintes intervalos de confiança, com 95% de probabilidade para acurácia, sensibilidade e especificidade: [87, 91] %, [82, 88] %, [84, 90] %.

Os trabalhos apresentados nesta seção, foram de crucial importância para o desenvolvimento do presente estudo. A partir deles foi possível conhecer as métricas de desempenho comumente empregadas nos trabalhos desta área, bem como os valores destas que normalmente são atingidos.

Por fim, é importante ressaltar que os trabalhos destacados colaboraram para a definição da metodologia de pesquisa deste trabalho, bem como forneceram uma base de dados para ser avaliada. No quadro abaixo, encontra-se um resumo dos resultados obtidos pelos autores citados no estado da arte.

Tabela 1: Resumo do Estado da Arte

Autor	**Sensibilidade (%)**	**Especificidade (%)**	**Acurácia (%)**	**Fonte de Dados**	**Modelo**
Patrício *et al.* (2018)	82,00-88,00	84,00-90,00	87,00-91,00	Exames de Rotina	Máquina de Vetores de Suporte
Andrade *et al.* (2017)	93,60	95,90	95,80	Termogramas	K-Star
Holsbach *et al.* (2014)	97,90	98,56	97,77	Amostras de células de tumores da mama	KNN
Silva *et al.* (2014)	100,00	91,44	94,43	Amostras de células de tumores da mama	Rede Neural ARTMAP-Fuzzy

Fonte: Elaborado Pelo Autor

1.2 Motivações

De acordo com a Organização Pan-americana de Saúde PAHO (2018), cerca de 627 mil mulheres morreram em 2018 devido ao câncer de mama em todo o mundo. Ainda segundo Oncoguia (2017), quando o câncer de mama é detectado precocemente as chances de cura são de 95%, por sua vez, se o câncer for descoberto mais tarde, a taxa de cura cai para 50%.

Com base no exposto, verifica-se a grande capacidade destrutiva desta doença, bem como a necessidade de um diagnóstico precoce de modo a trazer mais chances de sobrevivência para o paciente.

Neste contexto, esse trabalho encontra sua justificativa uma vez que caso os modelos classificatórios desenvolvidos obtenham desempenho adequado, tem-se a possibilidade de adicionar um diagnóstico de câncer de mama aos exames de rotina realizados nos laboratórios de maneira barata e não invasiva, o que provavelmente poderia acarretar em uma maior possibilidade de que as mulheres pudessem identificar esta doença em estágios iniciais.

É importante ressaltar que neste trabalho foram empregados modelos que ainda não foram testados pelos outros trabalhos da literatura que utilizaram o banco de dados adotado, o que traz a possibilidade de novas descobertas e até mesmo da identificação de um modelo de classificação mais adequado para o problema.

Por fim, é importante destacar que a metodologia proposta, não tem como finalidade substituir as técnicas clássicas de detecção do câncer de mama, mas sim orientar o paciente a procurar um especialista e realizar os exames adequados, caso o modelo indique a possibilidade de existência do câncer de mama.

1.3 Contribuições

1.3.1 Contribuição Geral

- Descobrir qual dos modelos (MVS, PMC, KNN ou RBF) e sua respectiva configuração, que maximiza a qualidade de classificação do conjunto de dados de validação relacionados a detecção do câncer de mama por intermédio de exames de rotina.

1.3.2 Contribuições Específicas

- Elaborar rotinas automatizadas para a criação de modelos MVS, PMC, KNN e RBF, por meio do MATLAB;
- Aplicar diferentes combinações de variáveis de entrada retiradas do banco de dados alvo do estudo, nos modelos criados no MATLAB, com o objetivo de avaliar qual a combinação de variáveis de entrada que maximiza o desempenho de cada modelo.

- Aplicar a análise de componentes principais no treinamento dos modelos, de forma a reduzir a dimensão do conjunto de dados de entrada;
- Avaliar o desempenho de cada modelo treinado anteriormente por meio de sua aplicação no conjunto de validação;
- Comparar os melhores resultados obtidos com cada um dos testes efetuados;
- Desenvolver *software* de simulação de diagnóstico do câncer de mama, com base no modelo de melhor desempenho.

1.4 Metodologia

O trabalho foi iniciado com a busca de um banco de dados sobre o câncer de mama, sendo que foi escolhido o banco de dados de Patrício *et al*. (2018) por ele ser recente e tratar de uma vertente nova da literatura (detecção do câncer de mama por meio de exames de sangue). Este banco de dados é composto por 116 instâncias, das quais 52 não possuem o câncer de mama e 64 possuem esta doença.

Durante o processo de ajuste dos modelos, foi separado 70,69% dos dados para treinamento (47 doentes e 35 não doentes) e 29,31% para validação (17 doentes e 17 não doentes), sendo a separação entre os conjuntos de treinamento e validação realizada de maneira aleatória.

De forma a manter a aleatoriedade e garantir uma quantidade igual de doentes e não doentes no conjunto de validação, separou-se o banco de dados em dois grupos, o de doentes e o de não doentes. Em seguida, sorteou-se destes conjuntos 47 doentes e 35 não doentes de forma a construir o conjunto de treinamento, por meio da função **randperm** disponibilizada pelo MATLAB. Por fim, os dados não selecionados para o conjunto de treinamento foram transferidos para o conjunto de validação.

Os modelos de classificação foram implementados por meio do software MATLAB. Para cada configuração do modelo e combinação de entradas foram realizadas 35 execuções, de modo a gerar um intervalo de confiança com 95% de probabilidade, por meio da distribuição normal para as seguintes medidas de desempenho: acurácia, sensibilidade e especificidade, utilizando o *software* Excel.

Em cada modelo, serão variadas a quantidade de atributos de entrada, de acordo com a relevância de cada preditor, sendo também aplicada uma análise de componentes principais, com 2 até 8 componentes principais.

Para todos os modelos, aplicou-se normalização nos dados de entrada, fazendo-os ficar entre -1 e 1. Para a rede neural RBF foi também aplicado a normalização *z-scores*.

Após a realização das simulações, foram separados os melhores resultados de cada configuração de entrada em cada modelo, com base na medida de acurácia média. Os melhores resultados foram tabelados de modo a permitir a identificação de qual modelo obteve o melhor desempenho, bem como, qual a influência das diferentes configurações de entradas para com a acurácia do modelo.

Por fim, com a definição do melhor modelo de classificação, serão apresentadas a curva ROC e a matriz de confusão do mesmo, bem como será apresentado um *software* desenvolvido no MATLAB que com base neste modelo e nos dados de entrada realiza a simulação do diagnóstico do câncer de mama.

1.5 Organização do Texto

O presente trabalho é dividido em 5 capítulos, com organização descrita conforme a seguir:

- No Capítulo 2, os classificadores empregados no trabalho são apresentados, bem como as suas métricas de desempenho. Além disso, é apresentado uma descrição da base de dados utilizada;
- No Capítulo 3, os resultados da pesquisa são apresentados juntamente com o *software* de simulação de diagnóstico do câncer de mama desenvolvido;
- No Capítulo 4, são apresentadas as considerações finais do trabalho.

2 ARQUITETURA DO CLASSIFICADOR, DAS SUAS ENTRADAS E DAS SUAS MÉTRICAS DE DESEMPENHO

Este capítulo destina-se a apresentar a descrição da base de dados utilizada, bem como introduzir de maneira genérica os classificadores empregados neste trabalho para o diagnóstico do câncer de mama, sendo destacado também a motivação para o uso de cada um deles. Também será apresentado o conceito associado a análise de componentes principais, bem como será descrito as métricas para aferição de desempenho dos classificadores empregados.

2.1 Descrição da Base de Dados

A descrição da base de dados, será realizada conforme as informações disponibilizadas pelo organizador da mesma. O conjunto de pessoas doentes, constituem de mulheres que foram diagnosticadas recentemente com relação ao câncer de mama, sendo elas recrutadas do Departamento de Ginecologia do Centro Hospitalar Universitário de Coimbra no período de 2009 até 2013.

O diagnóstico destas foi realizado por meio de uma mamografia com resultado positivo, confirmada histologicamente. Todas as amostras foram coletadas antes da realização de qualquer tratamento ou cirurgia. Algumas mulheres saudáveis voluntárias foram selecionadas para participar do estudo.

Todos os indivíduos da base de dados não haviam passado por qualquer tipo de tratamento prévio contra o câncer, e também estavam livres de qualquer outra infecção ou doença, que não seja o câncer de mama, no momento em que se inscreveram para participar do estudo.

As amostras sanguíneas foram coletadas no mesmo horário do dia, após um jejum durante a noite. Por sua vez, os dados antropométricos foram coletados de todos os pacientes em condições semelhantes, pela mesma equipe de pesquisa durante a primeira consulta.

As possíveis variáveis de entrada disponibilizadas pelo banco são Glicose (V1) [mg/dL], Resistina (V2) [ng/mL], Idade (V3) [anos], Índice de Massa Corporal (V4) [kg/m^2], HOMA (V5), Leptina (V6) [ng/mL], Insulina (V7) [µU/mL], Adiponectina (V8) [µg/mL] e MCP-1 (V9) [pg/dL], sendo estas apresentadas em ordem decrescente de importância de acordo com uma análise multivariada realizada pelo publicador da base de dados. A base também dispõe de um parâmetro binário que informa se a amostra de dados é respectiva a uma pessoa doente ou não.

Ainda com relação aos parâmetros disponibilizados pelo banco de dados, alguns deles necessitam de uma descrição para que o leitor possa compreender corretamente a que eles se referem.

O índice de massa corporal é expresso na unidade (kg/m^2) e calculado pela Equação 2.1:

$$IMC = \frac{Massa}{Altura^2} \tag{2.1}$$

O parâmetro HOMA corresponde a um modelo de avaliação de homeostase, sendo útil para quantificar a resistência à insulina, calculado por meio da Equação 2.2:

$$HOMA = \log\big((NI)x\ (NG)\big)\ /22{,}5 \tag{2.2}$$

em que NI refere-se ao nível de insulina em jejum e NG refere-se ao nível de glicose em jejum.

O MCP-1 é uma proteína que recruta monócitos, células T de memória e células dendríticas para locais onde ocorrem inflamações, sejam elas causadas por lesão ou por infecção em tecidos.

A Leptina é um hormônio muitas vezes associado a obesidade, tendo como principal efeito o controle do apetite.

A Resistina corresponde a uma proteína que tem como principal finalidade o bloqueio da ação da Leptina.

A Adiponectina é um hormônio produzido no tecido adiposo, o qual tem papel crucial na supressão de alguns eventos metabólicos associados a algumas doenças como a diabetes e a obesidade.

O estudo que produziu esta base de dados foi aprovado pelo comitê de ética do Centro Hospitalar Universitário de Coimbra, sendo realizado em conformidade com a declaração de Helsinki. Todos os pacientes apresentaram um consentimento por escrito antes de participarem do estudo.

2.2 Definição de um Classificador de Padrões

O processo de classificação consiste na identificação e diferenciação de objetos em classes, de acordo com as suas características mensuráveis. Estas características dos objetos são

responsáveis pela formação de um espaço multidimensional, que é conhecido como o espaço de características, aonde cada objeto pode ser representado como um ponto no espaço.

Nesta perspectiva, de acordo com Eduardo (2018c) um classificador de padrões refere-se a uma função discriminante responsável por determinar as semelhanças de uma nova amostra com relação a alguma das classes em estudo. Para tanto utiliza-se as informações de um conjunto de treinamento constituído por amostras cujas respectivas classes são previamente conhecidas.

2.3 Classificadores Avaliados

2.3.1 *K-Nearest Neighbors*

Lima (2012) define que o *K-Nearest Neighbor* (KNN) é um dos mais simples algoritmos de classificação, sendo utilizado para classificar elementos utilizando com critério a proximidade destes no espaço das características.

Para aplicar o algoritmo é necessário a existência de um conjunto de treinamento. Em seguida, deve-se escolher uma métrica para calcular as distâncias entre os elementos do conjunto de dados (LIMA, 2012). Além disso, é necessário definir um valor denominado *k*, o qual indica o número de vizinhos mais próximos que serão considerados para definir a classe de um determinado elemento do conjunto de validação.

O KNN pode ser descrito em três etapas:

1. Inicialmente, supondo a existência de um elemento desconhecido que não esteja no conjunto de treinamento, deve-se calcular a distância deste elemento desconhecido até todos os outros elementos do conjunto de treinamento.

$$distancias = calcula\_distancia(elemento, conjuntoTreinamento) \quad (2.3)$$

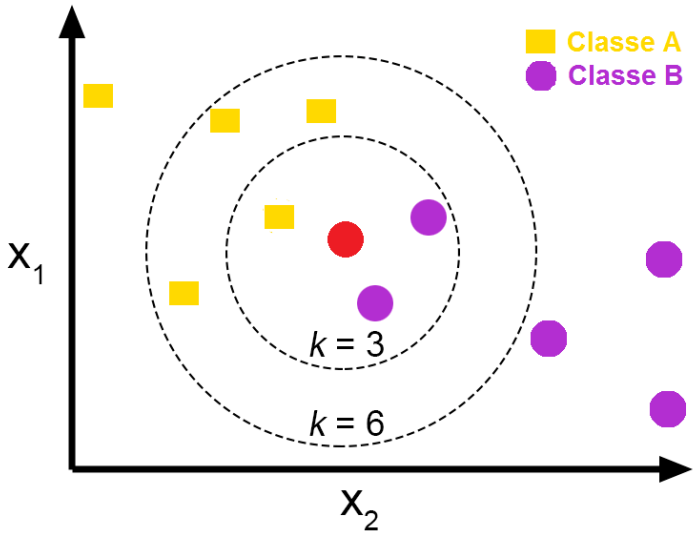
2. Em seguida, realiza-se a identificação dos *k* elementos do conjunto de treinamento que possuem a menor distância até a amostra desconhecida.

$$indice\_k\_elementos = indices(\min(distancias, k)) \quad (2.4)$$

3. Por fim, a classe que será atribuída ao elemento desconhecido será a classe predominante nos k elementos mais próximos.

Na Figura abaixo, tem-se uma representação do funcionamento do algoritmo KNN. Neste caso, busca-se classificar se o elemento de cor vermelha pertence à classe A ou B. Verifica-se que a adoção do número de vizinhos mais próximos k=3, e considerando-se uma métrica de distância euclidiana, o elemento será classificado como pertencente a classe B, uma vez que a classe predominante nos 3 elementos mais próximos é a classe B.

Figura 1: Representação do Funcionamento do Algoritmo KNN



Fonte: https://miro.medium.com/max/1400/0*jqxx3-dJqFjXD6FA

Por sua vez a adoção de um k=6, implica que o elemento de cor vermelha seria classificado como pertencente a classe A, uma vez que ela é a classe que predomina nos 6 elementos mais próximos.

Normalmente, é necessário aplicar uma normalização nos dados de treinamento de modo a evitar que um atributo domine as medidas de distâncias. Em adição, deve-se realizar testes com o parâmetro *k* de modo a determinar o seu valor sub-ótimo. Este modelo foi selecionado para ser avaliado neste trabalho, por ser um dos algoritmos mais simples para um processo de classificação, apresentando bons resultados em alguns casos práticos.

Em adição as motivações para uso deste modelo, Cavalcanti (2010) apresenta as seguintes vantagens com relação a adoção do mesmo: rápido treinamento, capacidade de aprendizado de funções complexas e não desperdício de informação. Em contraste, o mesmo autor apresenta as seguintes desvantagens para o modelo: lentidão para realizar uma consulta e facilidade de engano devido a um atributo irrelevante.

Neste trabalho, realizou-se testes com as métricas de distância: chebychev, euclidiana e mahalanobis. Além disso, para cada distância o número de vizinhos mais próximos utilizado para classificação foi variado de 1 até 35.

2.3.2 *Perceptron* de Múltiplas Camadas

Haykin (2001) define a rede neural como uma máquina, que deve ser projetada para modelar a forma como o cérebro humano realiza determinadas tarefas de interesse. Ela pode ser implementada utilizando-se componentes eletrônicos ou por meio de programação em um computador digital.

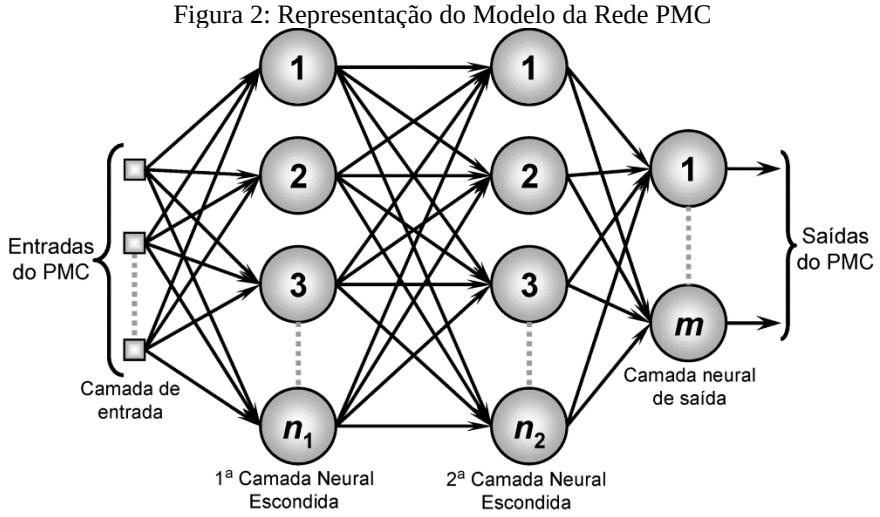
A rede neural artificial inspira-se na rede neural biológica, sendo então composta por unidades de processamentos simples, os neurônios (HAYKIN, 2001). Eles são responsáveis por simular o armazenamento e disponibilização do conhecimento, com base em seus pesos sinápticos. Esta estrutura computacional se assemelha ao cérebro humano, na medida em que o conhecimento é adquirido por meio do seu ambiente através de um processo de aprendizagem, proporcionando assim valores para os pesos sinápticos (HAYKIN, 2001).

De acordo com Silva, Spatti e Flauzino (2010), um dos tipos de redes neurais é a *Perceptron* de Múltiplas Camadas (PMC). Ela é caracterizada pela existência de no mínimo duas camadas de neurônios, sendo uma delas a camada intermediária e a outra a camada de saída.

A PMC pode ser classificada como pertencente a arquitetura *feedforward* de camadas múltiplas, sendo seu treinamento realizado de forma supervisionada, geralmente por meio do algoritmo *backpropagation* ou de alguma de suas variações. Este tipo de rede neural pode ser aplicado para resolver diversos tipos de problemas, de diferentes áreas do conhecimento. Uma das suas aplicabilidades é o reconhecimento de padrões, a qual será explorada neste trabalho.

Uma representação do modelo da PMC é exibida na Figura 2. Utilizou-se neste trabalho uma única camada escondida, com função de ativação tangente sigmoide hiperbólica. A camada de saída possui 2 neurônios com a função de ativação *softmax*. A quantidade de neurônios da camada escondida foi variada de 5 até 100 neurônios, aumentando-se 5 neurônios por vez. A configuração das entradas foi variada conforme é descrito em **1.4**.

Figura 2: Representação do Modelo da Rede PMC



Fonte: Silva, Spatti e Flauzino (2010, p. 92)

O algoritmo de treinamento utilizado foi o *Levenberg-Marquardt backpropagation*, com a condição de parada de 10^{-7} para o gradiente de performance ou de um número máximo de 10000 interações.

Este modelo foi selecionado para análise, por ser um dos mais clássicos e importantes modelos no contexto de classificação de padrões. Além disso, Batista (2003) apresenta as seguintes vantagens para este modelo: simplicidade de implementação e boa capacidade de generalização. Por sua vez, as desvantagens citadas são: dificuldade em justificar as respostas, significativo custo computacional e baixa velocidade de aprendizado.

2.3.3 Rede Neural de Função de Base Radial

Silva, Spatti e Flauzino (2010) destacam que a rede RBF, assim como a PMC, pertence a arquitetura *feedforward* de múltiplas camadas. A RBF pode ser empregada em praticamente todos os problemas que podem ser resolvidos pela PMC, sendo que algumas das suas aplicabilidades mais comum são a de aproximação de funções e classificação de padrões.

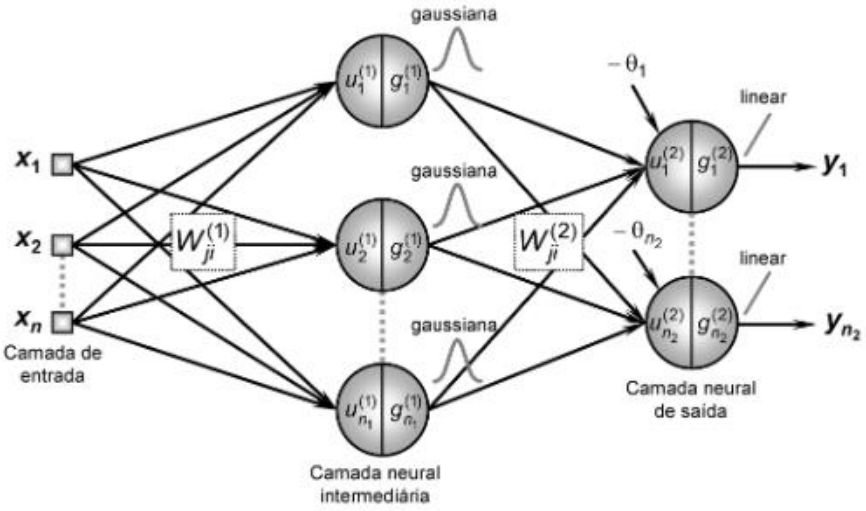
A RBF, diferencia-se da PMC, na medida em que a primeira, em sua estrutura típica possui apenas uma camada intermediária, na qual a função de ativação é do tipo gaussiana (SILVA; SPATTI; FLAUZINHO, 2010). Um outro ponto de diferenciação diz respeito a estratégia de treinamento, na RBF o treinamento é dividido em duas partes descritas a seguir.

1. Na primeira parte do treinamento da RBF os pesos dos neurônios da camada intermediária são ajustados de maneira não-supervisionado com base em algum método de aprendizagem auto-organizado (exemplo: *K-means*), que depende apenas das características dos dados apresentados na entrada.

2. Na segunda parte, os pesos da camada de saída são ajustados utilizando-se uma metodologia semelhante a utilizada na última camada do PMC, ou seja, é utilizado um treinamento supervisionado baseado na regra delta.

Na Figura 3, o modelo da RBF é representado graficamente. Neste trabalho, variou-se a quantidade de neurônios da camada escondida de 5 até 50, com o passo de cinco neurônios por vez. Ajustou-se também o desvio padrão da gaussiana, testando-se valores de 0,1 até 10, com o passo de 0,1 por vez. Na camada escondida é utilizado a função de ativação de base radial, sendo na camada de saída utilizada uma função de ativação linear.

Figura 3: Representação do Modelo da Rede RBF



Fonte: Silva, Spatti e Flauzino (2010, p. 173)

Este modelo foi selecionado, por ser clássico, bastante conhecido e com possibilidades bastante promissores para aplicações na área de classificação de padrões. Eduardo (2018a) também ressalta que, a rede RBF pode apresentar desempenho melhor que a PMC em tarefas difíceis de classificação, e além disso o treinamento é bem mais rápido que o da PMC.

Por sua vez, com relação as desvantagens, o mesmo autor ressalta que a RBF é mais sensível a outliers e além disso normalmente possui um maior número de parâmetros a serem ajustados que a PMC.

Neste trabalho foi utilizado a implementação da RBF do MATLAB. De acordo com MathWorks (2019) a RBF do Matlab é implementada conforme o algoritmo descrito a seguir: Inicialmente a camada escondida não possui neurônios, sendo executado os procedimentos enumerados abaixo até que a condição de parada do algoritmo seja atingida:

1. A resposta da rede a cada entrada é testada;
2. O vetor de entradas com o maior erro é encontrado;

3. Um neurônio é adicionado a camada escondida com pesos iguais ao vetor encontrado no passo 2;
4. Os pesos da camada de saída são ajustados de forma a minimizar o erro.

2.3.4 Máquina de Vetores de Suporte

De acordo com MQL5 (2014), as máquinas de vetores de suporte constituem uma metodologia que permite classificar dados de entrada entre duas categorias. Para realizar tal tarefa, a MVS mapeia as entradas do conjunto de treinamento em um espaço multidimensional de características, sendo posteriormente utilizado uma regressão para encontrar um hiperplano de separação ótimo entre as duas classes.

Após o treinamento da máquina, e com base no hiperplano de divisão, torna-se possível a classificação de novas entradas. Haykin (2001) destaca que o treinamento deste tipo de máquina constitui um problema de programação quadrática, sendo então atrativo devido a eficiência do procedimento computacional, bem como devido a garantia de que um extremo global da superfície de erro seja encontrado.

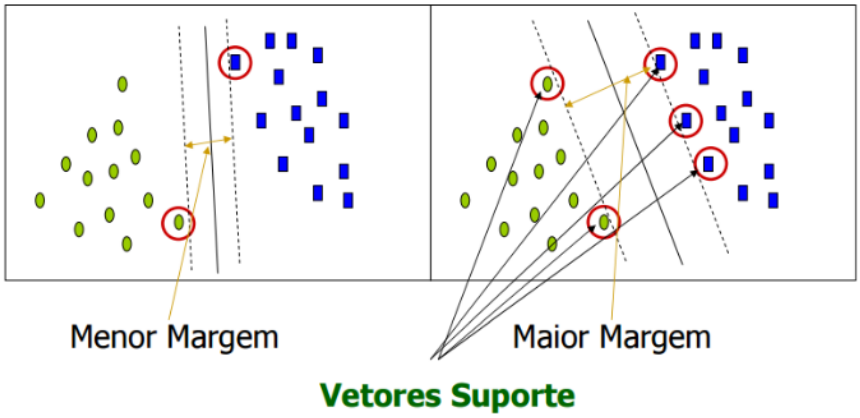
Eduardo (2018b) apresenta também algumas das seguintes vantagens relacionadas ao uso do MVS. A primeira vantagem está associada ao fato de que o modelo sempre encontra a melhor solução possível para o problema de classificação em questão. Ele também é um dos mais eficientes classificadores para problemas de dimensões elevadas.

O mesmo autor destaca as seguintes desvantagens com relação ao uso do MVS. A primeira está relacionada ao fato da MVS ser um classificador do tipo caixa-preta, não permitindo assim a interpretação da estratégia de decisão que está sendo tomada. Além disso, o modelo é voltado apenas para atributos numéricos, trazendo a necessidade de realizar conversão para que se possa trabalhar com atributos discretos.

Além do que foi exposto anteriormente, a motivação que levou a escolha da realização de teste com este modelo, foi o fato de que Patrício *et al*. (2018), obtiveram seu melhor resultado com um modelo deste tipo. Desta forma, decidiu-se empregá-lo também neste estudo.

Para a este modelo foram realizados testes com os seguintes tipos de *kernel*: linear, polinomial, gaussiano e sigmoide. Na Figura abaixo tem-se uma representação de uma classificação de retângulos e círculos utilizando máquina de vetores de suporte. Verifica-se que a MVS definirá os seus vetores de suporte e maximizará a margem de separação das classes, em busca de atingir o hiperplano de separação ótimo apresentado no lado direito da figura.

Figura 4: Representação da Classificação Realizada pela SVM



Fonte: Araújo (2015, p. 17)

2.4 Análise de Componentes Principais

Consultoria (2017) define a Análise de Componentes Principais como uma técnica de análise multivariada, cujo objetivo é condensar as informações contidas nas várias variáveis originais. Esta condensação gera um conjunto menor de variáveis estatística, as quais são denominadas componentes. Na realização deste procedimento, busca-se minimizar a perda de informações.

Sabe-se que o projeto e implementação de classificadores para dados de grandes dimensões é bastante complexo e oneroso, visto que nem sempre é possível obter bom desempenho de classificação e além disso o tempo de treinamento e validação do modelo aumenta proporcionalmente com a dimensionalidade dos dados.

Neste contexto, verifica-se que a compressão dos dados, com a preservação das principais informações contidas neles, pode tornar-se uma boa maneira de melhorar o desempenho e tempo de treinamento dos classificadores propostos neste trabalho. Nesta perspectiva, justifica-se a utilização da técnica de PCA.

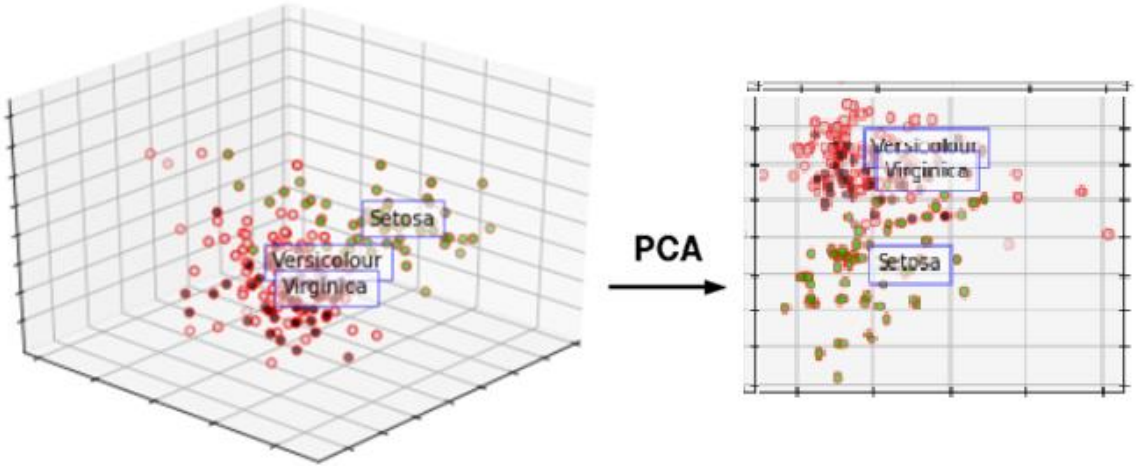
Zuben e Attux (2010) ressaltam que uma das maneiras de se realizar a PCA, é por meio de uma transformação linear **A** que consiste de uma matriz de dimensões M x N, a qual a partir de uma série de vetores de entrada com dimensão **N**, projeta estes dados de forma a gerar vetores com dimensão **M**. Por sua vez, a dimensão do espaço **M** é menor que a dimensão **N** do espaço dos dados originais.

Os mesmos autores ressaltam que a matriz de projeção **A** pode ser obtida de acordo com os seguintes passos: Primeiro é necessário fazer com que os dados passem a ter média amostral nula. Em seguida, realiza-se a estimativa da matriz de autocorrelação dos dados. Agora, define-se um valor de **M** correspondente a nova dimensão que se deseja alcançar.

Logo em seguida, escolhe-se os **M** autovetores associados aos **M** maiores autovalores da matriz de autocorrelação, para serem as direções principais de projeção. Por fim, monta-se a matriz **A** concatenando-se os **M** vetores selecionados anteriormente.

Na Figura abaixo tem-se uma exemplificação da utilização do PCA. Percebe-se nela a existência de um problema de classificação tridimensional envolvendo três espécies da planta íris (setosa, virginica, versicolor). Com a aplicação do PCA o problema foi simplificado para um espaço de duas dimensões.

Figura 5: Exemplificação da Utilização do PCA



Fonte: https://www.deeplearningitalia.com/wp-content/uploads/2018/01/1.png

2.5 Métricas Analisadas

Nesta seção, serão apresentadas as métricas de desempenho que foram empregadas na comparação dos modelos testados neste trabalho. Os conceitos de sensibilidade, especificidade e acurácia, serão apresentados com base em Moreira (2011).

Para que estes conceitos possam melhor ser ilustrados, observe a Tabela 2, a qual corresponde a uma forma de se representar o desempenho da aplicação de algum exame. Nas colunas, tem-se a condição real do paciente, por sua vez nas linhas tem-se o resultado dos testes.

Tabela 2: Representação da Qualidade de Um Exame

Condição do Paciente / Resultado do exame	**Doentes**	**Não Doentes**	**Total**
Positivo	VP	FP	VP+FP
Negativo	FN	VN	FN+VN
Total	VP+FN	FP+VN	VP+FP+FN+VN

Fonte: Moreira (2011, p.1)

Desta forma, o cruzamento entre as linhas e colunas, apresentam algumas informações relevantes sobre a qualidade deste exame. Por exemplo, o valor **VP** refere-se a quantidade de pessoas submetidas ao exame que estavam doentes, e que o resultado do exame confirmou a presença da doença (valor verdadeiro positivo).

O valor **VN** indica a quantidade de pessoas que não estavam doentes e o resultado do exame confirmou esta situação (valor verdadeiro negativo). **FN** refere-se as pessoas que estavam doentes, mas o exame falhou em identificar esta condição (valor falso negativo). Por fim, o valor **FP** remete-se a quantidade de pessoas que não estavam doentes, mas o exame apresentou um resultado contrário (valor falso positivo).

Com base nestes valores apresentados anteriormente, é possível definir a acurácia, sensibilidade e especificidade. A sensibilidade consiste na probabilidade de ocorrência de um verdadeiro positivo, sendo calculada por meio da expressão abaixo.

$$sensibilidade = \frac{VP}{VP + FN} \tag{2.5}$$

A especificidade é a probabilidade de ocorrência de um verdadeiro negativo, sendo calculado pela expressão abaixo.

$$especificidade = \frac{VN}{FP + VN} \tag{2.6}$$

Por sua vez, a acurácia é a probabilidade de o exame fornecer resultados corretos, podendo ser obtida por intermédio da expressão abaixo.

$$acurácia = \frac{VP + VN}{VP + FP + FN + VN} \tag{2.7}$$

Câmara (2009) define a curva *Receiver Operating Characteristic* (ROC) como uma medida da capacidade de um modelo classificar corretamente um dado. Para o desenvolvimento de modelos de classificação busca-se maximizar os acertos e minimizar os erros, desta forma a

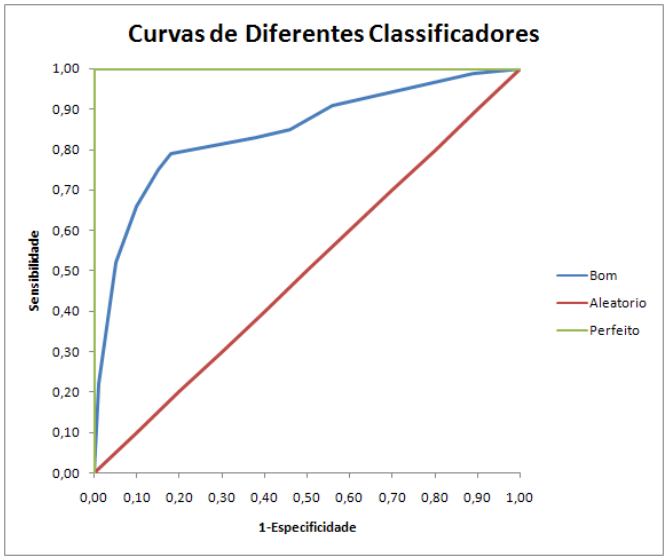
curva ROC representa uma forma conveniente de verificar se estas premissas estão sendo atendidas.

A ROC prova-se eficiente para verificar este tipo de comportamento no modelo, pois apresenta todos os valores de sensibilidade (acertos positivos) no eixo **y** e associa esta informação ao eixo **x** correspondente a proporção de falsos acertos (1 – especificidade). (CÂMARA, 2009).

Desta forma, quando se observa uma concentração da curva ROC na parte superior esquerda do gráfico, pode-se concluir que a sensibilidade do modelo é alta e ao mesmo tempo a proporção de falso positivo é baixa, indicando assim que o modelo é preciso.

Na Figura 6 tem-se uma exemplificação de três curvas ROC, cada uma sendo avaliada com relação a qualidade de classificação do respectivo modelo. Pode-se perceber na Figura que conforme foi apresentado anteriormente as curvas que se concentram na parte superior esquerda do gráfico possuem bom desempenho, sendo que à medida que ela se afasta desta localidade, o desempenho do modelo respectivo é menor.

Figura 6: Exemplificação da Curva ROC



Fonte: http://crsouza.com/wp-content/uploads/2009/07/example-roc-curves_thumb-5B6-5D.png

Uma forma simples de se avaliar a eficácia de um modelo classificativo por meio da curva ROC, consiste em calcular a área sob a mesma. Câmara (2009) através da tabela abaixo apresenta a associação entre a faixa do valor da área da curva ROC e a respectiva qualidade do modelo com uma área da ROC pertencente ao intervalo. É importante destacar que a área da curva ROC corresponde a acurácia do modelo.

Tabela 3: Associação Entre a Área da Curva ROC e a Qualidade do Modelo

Valor da Área da ROC	**Qualidade do Modelo**
>0,9	Excelente
>0,8 e ≤ 0,9	Bom
>0,7 e ≤ 0,8	Regular
>0,6 e ≤ 0,7	Ruim
>0,5 e ≤ 0,6	Reprovado

Fonte: Câmara (2009, p.1)

3 RESULTADOS E DISCUSSÕES

Este capítulo tem por finalidade apresentar os resultados obtidos com cada um dos modelos empregados para diagnóstico do câncer de mama. Além disso, apresenta-se de maneira detalhada o melhor modelo encontrado, realizando-se comparações com a literatura. Por fim, apresenta-se um sistema desenvolvido para realizar o diagnóstico do câncer de mama com base neste melhor modelo encontrado.

3.1 Análise para a Rede PMC

Na Tabela 4, encontra-se um resumo dos melhores resultados de execução para a PMC, em diferentes configurações de variáveis de entrada adotadas. Os resultados são apresentados em intervalos de confiança com 95% de probabilidade e 34 graus de liberdade para a sensibilidade, especificidade e acurácia.

A coluna referente a **configuração melhor** apresenta a quantidade de neurônios na camada escondida que gerou a melhor acurácia média. Para esta configuração específica, foram retirados os dados para geração dos intervalos de confiança.

As entradas da Tabela 4 com a forma **V1-VY**, indicam que para o treinamento e validação do modelo foram utilizados o conjunto de variáveis V1, V2, V3, V4, ..., VY. Cada variável do banco de dados tem sua representação reduzida na forma **VX**, que pode ser observada na seção **2.1**.

Para a Tabela 4, ainda é importante ressaltar que as entradas descritas como **PCAX**, referem-se aos testes realizados com o emprego da técnica de análise de componentes principais para **X** componentes, variando estas de 2 até 8 componentes.

Tabela 4: Resultados da PMC

Entradas	Configuração Melhor	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
V1-V2	5	62,75-71,70	58,74-66,64	62,65-67,27
V1-V3	15	74,69-81,95	65,35-71,80	71,59-75,30
V1-V4	**35**	**78,24-84,45**	**65,60-73,89**	**72,78-78,31**
V1-V5	80	73,90-81,40	67,17-74,01	71,63-76,60
V1-V6	40	77,09-83,92	64,97-71,50	72,31-76,43
V1-V7	100	74,45-82,52	64,30-72,17	70,84-75,88
V1-V8	75	73,48-80,13	66,62-73,89	71,88-75,18
V1-V9	100	70,89-77,68	68,38-75,48	71,23-74,99
PCA2	65	56,28-63,72	45,62-53,87	52,40-57,35

PCA3	20	65,46-73,36	46,42-53,41	58,85-62,47
PCA4	80	64,23-71,23	54,90-62,75	60,47-66,09
PCA5	85	72,69-78,24	56,78-64,56	65,74-70,40
PCA6	75	73,31-81,99	64,85-71,95	71,06-75,00
PCA7	35	70,94-77,97	66,29-73,88	70,00-74,54
PCA8	100	69,43-77,46	67,48-74,03	70,07-74,14

Fonte: Elaborado Pelo Autor

Analisando a Tabela 4 e considerando-se os valores do melhor intervalor de acurácia, verifica-se que a melhor configuração de entradas para este modelo é uma rede PMC com 35 neurônios na camada oculta, com variáveis de entrada de glicose, resistina, idade e IMC. Este fato mostrou-se condizente com o estudo de Patrício *et al*. (2018), uma vez que estes também detectaram que esta seria a melhor configuração de entradas.

O emprego do PCA de 5 até 8 componentes mostrou-se sempre ser melhor que ao menos a configuração de entradas **V1-V2** (pior combinação de entradas que não são componentes). Além disso, é possível observar que apenas a configuração de 6 componentes apresentou resultados um pouco melhor que a combinação de entradas **V1-V9** (segunda pior configuração de entradas que não são componentes).

Desta forma, é possível concluir que o emprego do PCA no modelo da rede neural PMC não trouxe melhoria de desempenho considerável, uma vez que o melhor resultado do PCA conseguiu ser apenas melhor que a segunda pior combinação de entradas com variáveis do banco de dados.

Conforme era de se esperar, na medida em que se vai diminuindo muito a quantidade de componentes principais, o modelo vai perdendo desempenho, já que a alta diminuição de dimensionalidade do problema, acaba gerando uma alta perda de informação das variáveis de entrada. É possível constatar, conforme os dados dos próximos modelos, que este comportamento se replica na MVS, KNN e RBF. Além disso, será possível verificar que o PMC foi o modelo com pior desempenho neste trabalho.

3.2 Análise para a Rede MVS

Na Tabela 5, tem-se o resumo dos resultados para a máquina de vetores de suporte. A coluna configuração melhor, refere-se ao *kernel* que gerou a melhor acurácia média. Verifica-se que o melhor resultado ocorreu para a configuração de variáveis de entrada: glicose, resistina e idade (V1-V3) com o *kernel* gaussiano. A MVS apresenta um ganho de acurácia de 3,21 % com relação ao melhor resultado da PMC. Para calcular tal ganho o valor do intervalo superior

para a acurácia da MVS com entradas **V1-V3** foi subtraído do respectivo valor para o modelo da PMC com entradas **V1-V4**.

Tabela 5: Resultados da MVS

Entradas	Configuração Melhor	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
V1-V2	Gaussiano	85,15-89,97	55,05-62,93	70,97-75,59
V1-V3	**Gaussiano**	**88,12-92,38**	**66,19-72,30**	**77,98-81,52**
V1-V4	Gaussiano	87,62-92,88	65,57-72,91	77,99-81,50
V1-V5	Gaussiano	85,56-90,91	64,80-73,69	76,79-80,69
V1-V6	Gaussiano	86,78-92,04	57,40-64,85	73,01-77,58
V1-V7	Linear	66,28-73,55	75,93-83,40	72,26-77,32
V1-V8	Linear	65,66-72,82	73,70-80,92	70,89-75,67
V1-V9	Linear	66,83-73,67	71,21-79,38	70,22-75,33
PCA2	Sigmoide	55,27-65,40	37,51-44,51	48,50-52,85
PCA3	Polinomial	76,46-82,53	35,58-43,41	57,14-61,85
PCA4	Linear	70,04-76,17	56,00-62,32	63,91-68,36
PCA5	Polinomial	67,01-75,85	60,67-67,73	65,52-70,11
PCA6	Linear	65,38-71,43	69,60-77,63	68,82-73,20
PCA7	Linear	70,37-76,85	67,29-73,88	70,20-74,00
PCA8	Linear	63,99-70,80	68,94-76,27	67,72-72,28

Fonte: Elaborado Pelo Autor

A melhor configuração obtida para as variáveis de entrada foi um pouco diferente da obtida por Patrício *et al*. (2018), a qual também incorporava o índice de massa corporal. Porém, verifica-se que no presente modelo, as entradas **V1-V3** apresentaram resultados quase idênticos aos de **V1-V4**. Desta forma, pode-se considerar que a escolha de configuração de entradas ótimas mostrou-se próxima a do trabalho de Patrício *et al*. (2018).

Ainda comparando com Patrício *et al*. (2018), verifica-se que mesmo sendo empregado a mesma técnica, o melhor resultado de Patrício *et al*. (2018), mostrou-se aproximadamente 9,48 % melhor com relação ao valor superior do intervalo de confiança para a acurácia. Isto indica que o modelo empregado neste trabalho ainda é passível de otimização. Uma possível forma de se obter melhores resultados seria otimizar os parâmetros escala do kernel e restrição da caixa.

Com relação ao *kernel* que gerou o melhor resultado, observa-se que, nas configurações de 2 até 6 variáveis de entrada, o *kernel* mais adequado mostrou-se ser o **gaussiano**, por sua vez com o aumento da dimensionalidade das variáveis de entrada, o *kernel* **linear** mostrou resultados mais adequados.

Com relação ao emprego da técnica de PCA, pode-se observar que ela não trouxe nenhuma melhoria para o desempenho do modelo, sendo percebido que na verdade esta técnica gerou uma perda de desempenho, uma vez que nenhuma configuração de PCA apresentou melhor intervalo de confiança para a acurácia, que a pior configuração de variáveis de entrada deste modelo, que corresponde as entradas **V1-V9**.

Analisando-se os *kernels* que forneceram melhores resultados com o PCA, é possível concluir que para uma grande quantidade de componentes principais o *kernel* **linear** mostrou-se mais adequado, porém para uma pequena quantidade de componentes os *kernels* **sigmoide** e **polinomial** mostraram-se mais adequados.

3.3 Análise para o KNN

Na Tabela 6, encontra-se o resumo dos resultados para o algoritmo K-vizinhos mais próximos. A coluna configuração melhor apresenta o seguinte padrão: **distância**-k=**N**, onde **distância**, refere-se a métrica de distância empregada e **N** refere-se ao número de vizinhos mais próximos utilizados para classificar uma amostra do conjunto de validação.

Tabela 6: Resultados do KNN

Entradas	Configuração Melhor	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
V1-V2	euclidiana-k=29	77,37-84,31	68,67-74,52	74,07-78,37
V1-V3	**euclidiana-k=9**	**87,15-92,00**	**65,72-72,76**	**77,39-81,58**
V1-V4	mahalanobis-k=5	83,68-89,09	69,52-75,69	77,57-81,42
V1-V5	euclidiana-k=9	82,08-85,99	70,27-77,96	77,02-81,13
V1-V6	euclidiana-k=8	76,60-83,40	76,47-81,51	77,61-81,38
V1-V7	euclidiana-k=8	71,48-80,11	74,71-80,59	74,48-78,96
V1-V8	euclidiana-k=7	76,56-83,10	65,17-73,65	71,88-77,36
V1-V9	euclidiana-k=5	70,19-77,71	71,05-76,51	71,86-75,87
PCA2	mahalanobis-k=11	70,96-77,62	37,91-45,12	55,79-60,01
PCA3	mahalanobis-k=10	68,25-74,94	50,48-57,75	60,77-64,94
PCA4	mahalanobis-k=4	54,97-63,01	67,54-72,97	62,05-67,19
PCA5	mahalanobis-k=4	56,94-64,74	79,11-86,27	69,15-74,38
PCA6	mahalanobis-k=1	75,38-80,59	76,14-81,17	76,64-80,00
PCA7	euclidiana-k=1	72,83-78,43	68,23-74,63	71,64-75,42
PCA8	euclidiana-k=14	64,39-70,73	72,72-78,88	69,72-73,64

Fonte: Elaborado Pelo Autor

Verifica-se para este modelo, que a configuração de entradas ótima foi a mesma que a da MVS, sendo que novamente a configuração de entradas **V1-V4** mostrou desempenho

bastante próximo da configuração ótima com distância euclidiana e o valor de **k** sendo 9. O KNN mostrou-se ser apenas 0,06 % melhor que a SVM, considerando-se a medida superior do intervalo de confiança da acurácia para as melhores configurações de entradas.

Analisando-se agora as distâncias que mais aparecem na coluna de melhor configuração, é possível constatar que, quando as entradas do modelo são os dados brutos a distância euclidiana aparenta ser a mais adequada. A única exceção para esta observação é a configuração de entrada **V1-V4**, onde pode-se observar que o melhor resultado foi obtido a partir da distância de mahalanobis.

Por sua vez, quando as entradas do algoritmo são obtidas por uma análise de componentes principais, a distância de mahalanobis mostrou melhor desempenho, excerto no caso em que muitas componentes principais eram empregadas (7 e 8), sendo mais adequado a distância euclidiana nestes dois casos.

Observa-se também que para este modelo a aplicação do PCA foi a que gerou resultados mais próximos da melhor configuração de entradas brutas, comparando-se com os outros modelos, uma vez que, o PCA de 6 componentes ficou com o valor do intervalo superior da acurácia apenas 1,58 % menor que a configuração **V1-V3** deste modelo. Isto era de se esperar uma vez que, o KNN normalmente apresenta dificuldades para resolver problemas de grandes dimensões, devido ao crescimento arbitrário das distâncias entre os elementos.

3.4 Análise para a Rede RBF

Na Tabela 7, é apresentado o resumo dos resultados para o modelo de rede neural de função de base radial cujos dados de entrada foram normalizados entre -1 e 1 usando a função do MATLAB **mapminmax**. A coluna configuração melhor apresenta a seguinte padronização: **X**-**Y**, em que **X** corresponde ao número de neurônios e **Y** corresponde ao desvio padrão da gaussiana das configurações que geraram a maior acurácia média.

Tabela 7: Resultados da RBF Com Normalização mapminmax

Entradas	Configuração Melhor	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
V1-V2	10-7,6	75,24-81,40	71,68-77,90	74,87-78,24
V1-V3	5-3,4	86,54-90,94	72,22-78,71	80,17-84,03
V1-V4	**5-1,4**	**80,30-87,09**	**79,03-86,68**	**80,84-85,72**
V1-V5	10-4	82,16-87,93	79,41-83,95	81,90-84,92
V1-V6	10-5,5	81,53-86,87	75,10-81,88	79,21-83,48
V1-V7	15-7,8	78,78-84,92	78,03-83,65	79,10-83,59
V1-V8	15-8,9	80,13-85,59	74,09-81,54	77,96-82,71

V1-V9	15-2,4	78,05-85,31	72,50-79,43	76,60-81,04
PCA2	5-3,9	68,85-77,03	38,53-45,84	55,58-59,54
PCA3	10-2,3	76,37-82,29	50,01-57,89	64,09-69,19
PCA4	5-3,4	74,90-81,40	59,39-68,34	68,54-73,48
PCA5	15-2	71,08-78,50	66,50-74,68	70,56-74,82
PCA6	15-1,6	78,96-85,07	66,27-74,57	74,14-78,29
PCA7	20-2,7	74,12-81,52	71,06-77,52	73,63-78,47
PCA8	15-5,4	77,61-84,40	66,73-73,10	73,43-77,49

Fonte: Elaborado Pelo Autor

A melhor configuração de entradas para este modelo foi o de entradas **V1-V4**, possuindo um desempenho 4,14% superior ao do KNN e por isto sendo o melhor modelo encontrado neste trabalho. Analisando-se o intervalo de confiança para a acurácia, desta configuração, e comparando-se com as informações da Tabela 3, é possível concluir que foi obtido um resultado que pode ser avaliado como bom.

Com relação ao uso do PCA, verificou-se que não houve ganho de desempenho relevante com relação a sua aplicação neste modelo, com a normalização **mapminmax**, uma vez que apenas as configurações de 6 e 7 componentes principais, apresentaram desempenho ligeiramente melhor que a pior configuração de variáveis de entrada: **V1-V2**.

Na Tabela 8, encontra-se o resumo dos resultados para o modelo RBF cujas entradas foram normalizadas por meio da metodologia z-scores. Percebe-se que para este caso a melhor configuração correspondeu a de **V1-V5**, porém é possível concluir que o emprego da normalização z-scores produz em geral modelos com menor acurácia que o **mapminmax**, considerando esta base de dados e modelo, uma vez que a diferença no limite superior de acurácia para os melhores resultados foi de 0,83 %.

Tabela 8: Resultados da RBF Com Normalização z-scores

Entradas	**Configuração Melhor**	**Sensibilidade (%)**	**Especificidade (%)**	**Acurácia (%)**
V1-V2	5-3,1	79,26-85,11	67,63-75,90	74,68-79,27
V1-V3	10-3,6	86,69-91,46	73,90-79,37	81,25-84,46
V1-V4	15-8,6	78,84-86,20	79,48-86,57	80,69-84,86
V1-V5	**15-3,5**	**83,67-88,77**	**77,01-82,99**	**81,33-84,89**
V1-V6	10-4,3	79,65-86,06	75,76-81,89	79,17-82,51
V1-V7	15-9	81,07-85,99	74,29-81,00	78,77-82,41
V1-V8	15-3,5	79,71-85,67	72,94-80,67	77,83-81,67
V1-V9	20-7,4	74,13-82,17	72,00-79,93	74,71-79,41
PCA2	10-1,7	68,99-77,90	40,43-47,30	56,41-60,90

PCA3	10-6,6	76,93-84,08	47,99-55,87	63,91-68,53
PCA4	10-5,6	73,51-80,77	57,14-65,22	66,72-71,60
PCA5	15-4,2	73,01-79,26	68,51-75,02	72,07-75,83
PCA6	25-8,1	72,65-78,28	71,47-79,46	73,50-77,42
PCA7	15-7,8	78,38-85,31	69,58-76,31	75,55-79,24
PCA8	15-5,5	76,16-82,50	72,53-80,07	75,41-80,22

Fonte: Elaborado Pelo Autor

Analisando neste momento o emprego do PCA com relação a RBF submetida a este último tipo de normalização, é possível observar que a técnica também não trouxe melhorias significativas para o sistema de classificação, uma vez que apenas a configuração de 8 componentes principais obteve resultado um pouco melhor que a configuração de entradas **V1-V2**.

3.5 Análise para o Classificador Proposto

Esta seção, tem por objetivo comparar os intervalos de confiança para acurácia, sensibilidade e especificidade do melhor resultado deste trabalho com o melhor resultado de Patrício *et al.* (2018). Além disso, o modelo com melhor desempenho individual, retirado de umas das 35 execuções do modelo de melhor configuração, será apresentado de maneira detalhada de forma a possibilitar uma comparação com os trabalhos que apresentaram seus resultados na forma do melhor desempenho individual obtido.

Na Tabela 9, estão os melhores intervalos de confiança obtidos neste trabalho em paralelo com os melhores obtidos por Patrício *et al*. (2018), permitindo assim realizar uma comparação entre os dois trabalhos.

Tabela 9: Comparação de Intervalos

Trabalho	**Sensibilidade (%)**	**Especificidade (%)**	**Acurácia (%)**	**Fonte de Dados**	**Modelo**
Este Trabalho	80,30-87,09	79,03-86,68	80,84-85,72	Exames de Rotina	Máquina de Vetores de Suporte
Patrício *et al*. (2018)	82,00-88,00	84,00-90,00	87,00-91,00	Exames de Rotina	Rede Neural de Função de Base Radial

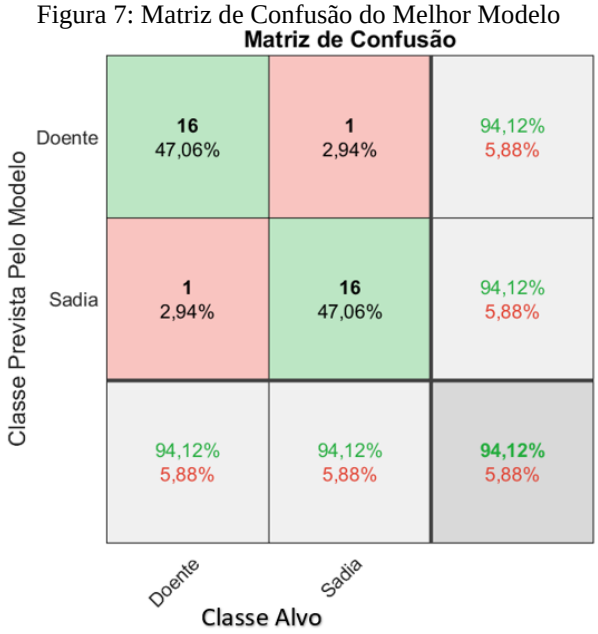
Fonte: Elaborado Pelo Autor

Ainda com relação a Tabela 9, verifica-se que os intervalos obtidos no outro trabalho são próximos aos obtidos neste, mostrando assim que a RBF conseguiu mostrar um desempenho

próximo ao da MVS empregada por Patrício *et al*. (2018). Com relação a sensibilidade observa-se que a maior parte do intervalo obtido neste trabalho está contido no intervalo do outro trabalho. Considerando a especificidade, observa-se algo semelhante, mas com diferenças mais acentuadas. Por sua vez, com relação a acurácia, os intervalos mostraram-se disjuntos, porém, a diferença entre os valores superiores do intervalo foi de apenas 5,28 %.

Na Figura 7, apresenta-se a matriz de confusão do melhor modelo treinado neste trabalho. Nesta matriz, observa-se que a sensibilidade do modelo foi de 94,12 %, implicando que apenas 5,88 % dos indivíduos doentes (1 indivíduo), acabariam sendo classificados como não doente. A ocorrência deste evento é bastante problemática, uma vez que este indivíduo não seria indicado a fazer um exame mais aprofundado, porém sabe-se que não existe modelos de classificação totalmente perfeitos, sendo esta percentagem de acerto considerada aceitável.

Figura 7: Matriz de Confusão do Melhor Modelo



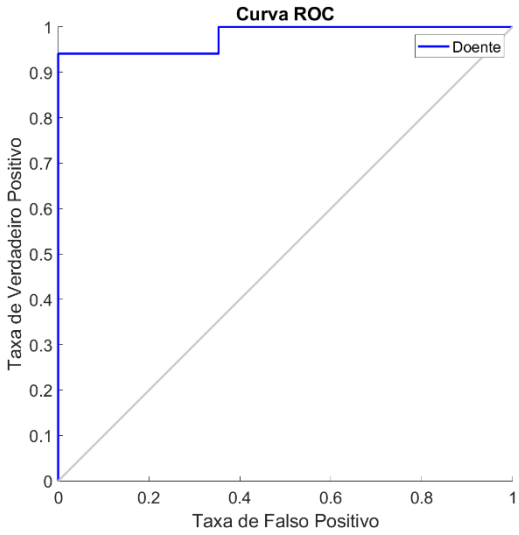
Fonte: Elaborado Pelo Autor

Ainda com relação a Figura anterior, observa-se também neste modelo a existência de uma especificidade de 94,12 %, significando assim que apenas 5,88 % das pessoas sadias seriam classificados como doentes. O acontecimento deste evento, principalmente com esta percentagem de frequência é menos preocupante, visto que o indivíduo apenas seria encaminhado a realizar exames mais aprofundados para diagnóstico da doença e com certeza acabaria descobrindo que não está com câncer de mama.

Considerando agora o desempenho geral do modelo, observa-se a existência de uma acurácia de 94,12%, ou seja, de todo o conjunto de validação, o modelo só errou no diagnóstico de 5,88 % dos indivíduos, o que corresponde a apenas duas pessoas.

Na Figura 8, tem-se a curva ROC do modelo em estudo.

Figura 8: Curva ROC do Melhor Modelo Obtido

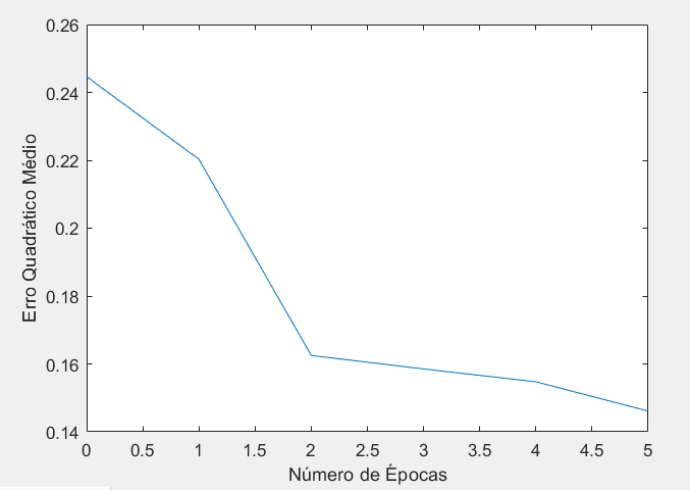


Fonte: Elaborado Pelo Autor

Analisando-se o gráfico da Figura 8, observa-se que a curva ROC do modelo em estudo, apresenta um bom desempenho, visto que tem-se uma baixa probabilidade de falso positivo e uma alta taxa de verdadeiro positivo.

Na Figura 9 tem-se a curva que relaciona o erro quadrático médio ao longo do número de épocas de treinamento do modelo em análise. O processo de treinamento foi finalizado após 5 épocas, de acordo com o critério de parada estabelecido pelo MATLAB, o qual treina o modelo por um número de épocas equivalente a quantidade de neurônios da camada escondida.

Figura 9: Curva de Treinamento do Melhor Modelo



Fonte: Elaborado Pelo Autor

Na Tabela 10, tem-se uma comparação de desempenho deste último modelo em estudo, com os encontrados por outros autores.

Tabela 10: Comparação Entre Os Modelos de Melhor Desempenho

Trabalho	**Sensibilidade (%)**	**Especificidade (%)**	**Acurácia (%)**	**Fonte de Dados**	**Modelo**
Este Trabalho	94,12	94,12	94,12	Exames de Rotina	Rede Neural de Função de Base Radial
Patrício *et al.* (2018)	82,00-88,00	84,00-90,00	87,00-91,00	Exames de Rotina	Máquina de Vetores de Suporte
Andrade *et al.* (2017)	93,60	95,90	95,80	Termogramas	K-Star
Holsbach *et al.* (2014)	97,90	98,56	97,77	Amostras de células de tumores da mama	KNN
Silva *et al.* (2014)	100,00	91,44	94,43	Amostras de células de tumores da mama	Rede Neural ARTMAP-Fuzzy

Fonte: Elaborado Pelo Autor

Comparando-se o melhor modelo obtido neste trabalho, com os intervalos de confiança de Patrício *et al*. (2018), verifica-se que tanto a sensibilidade, especificidade, com também a

acurácia deste melhor modelo foram superiores aos respectivos intervalos de confiança deste outro trabalho.

É importante salientar que na Tabela 9 foi possível verificar que o intervalo de confiança obtido por Patrício et al. (2018) foi melhor que o obtido neste trabalho. Desta forma provavelmente o melhor resultado de execução individual de seu modelo foi melhor que o deste trabalho, mas como Patrício et al. (2018) não apresentou seu melhor resultado individual, foi mantido o intervalo de confiança para comparação na Tabela 10.

Neste sentido, a diferença entre a Tabela 9 e a Tabela 10, diz respeito ao fato de que a Tabela 9 compara todos os trabalhos analisados que apresentam resultados na forma de intervalos de confiança. Por sua vez a Tabela 10 realiza uma comparação de todos os trabalhos analisados com enfoque na melhor execução individual do modelo, sendo considerado no caso de Patrício et al. (2018) os intervalos de confiança devido a não apresentação do melhor resultado individual.

Realizando-se agora a comparação com os outros trabalhos, que empregaram formas diferentes de identificar o câncer de mama, verifica-se que com relação ao trabalho de Andrade *et al*. (2017), o melhor modelo obtido neste trabalho apresentou uma acurácia, sensibilidade e especificidade, respectivamente: 1,68 % menor, 0,52% maior e 1,78 % menor.

Comparando-se agora com o trabalho de Holsback *et al.* (2014), verifica-se que a acurácia, sensibilidade e especificidade deste trabalho foram respectivamente: 3,65 % menor, 3,48 % menor e 4,44 % menor. Por fim, considerando agora o trabalho de Silva *et al*. (2014) tem-se que a acurácia, sensibilidade e especificidade foram respectivamente: 0,31 % menor, 5,88 % menor e 2,68 % maior.

Com base no exposto, é possível concluir que a metodologia de diagnóstico do câncer de mama com intermédio da análise de dados obtidos em exames de rotina é bastante promissora, uma vez que obteve desempenho compatível com os das outras metodologias comumente adotadas, como por exemplo, a análise de imagens dos termogramas e a análise de dados de amostras de tumor.

É importante ressaltar que os resultados obtidos neste trabalho ainda podem ser melhorados. Algumas das formas de conseguir melhores resultados seria buscar formas de otimização para cada um dos modelos empregados, utilizando por exemplo, técnicas inteligentes para otimização dos parâmetros de cada modelo.

Além disso, é possível alterar alguns comportamentos dos modelos empregados, o que pode melhorar o seu desempenho. Por exemplo, com relação a MLP, é possível trocar o algoritmo de aprendizado, os parâmetros de treinamento, as funções de ativação e etc. Em

adição, também é possível empregar outras técnicas de classificação, como por exemplo as Máquinas de Comitê de Decisão e o Algoritmo *K-means*.

Uma ampliação da quantidade de instancias da base de dados, com adição de outras variáveis podem também colaborar com a obtenção de melhores resultados, uma vez que ter-se-á a possibilidade de treinar os modelos com mais instâncias, e ter uma boa quantidade de dados disponível para validar o desempenho deles.

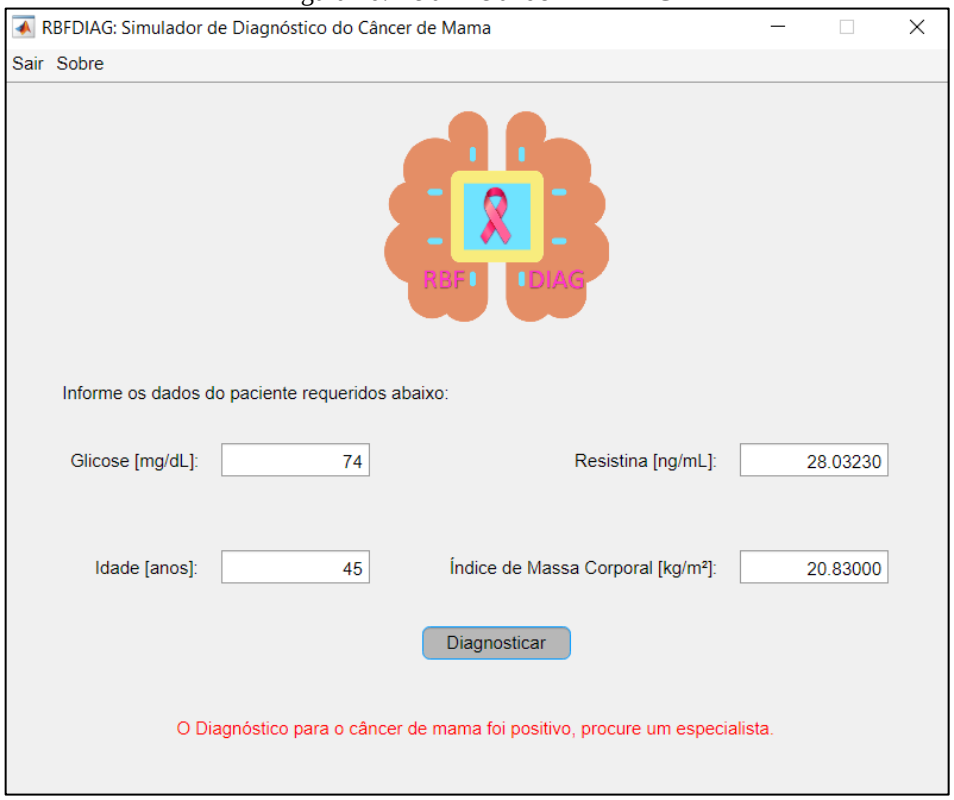
3.6 O *Software* RBFDIAG

Desenvolveu-se utilizando o MATLAB, um *software* denominado RBFDIAG, o qual com base no modelo proposto, realiza o diagnóstico do câncer de mama tendo como informações de entradas o nível de Glicose, Resistina, Idade e Índice de Massa Corporal do paciente a ser analisado.

O programa pode ser baixado através do seguinte *link*: (https://drive.google.com/open?id=1_-W0ZiTBLzVlJJb6xgGRA3FuoxZZG9Hw). No *link* apresentado é possível fazer o download de duas versões do programa. A primeira versão corresponde ao arquivo com o nome **rbfdiag_matlab.mlappinstall**, a qual requer a existência do *software* MATLAB previamente instalado na máquina. Com o MATLAB aberto, é possível executar o arquivo mencionado anteriormente, fazendo com que o sistema desenvolvido seja instalado e disponibilizado na aba APPS. A segunda versão corresponde ao arquivo de nome **rbfdiag_standardalone.exe**, a qual é destinada para as máquinas com sistema operacional Windows 7 ou posterior, com arquitetura de 64 bits e que não possuem o MATLAB previamente instalado.

Na Figura 10, é possível encontrar a tela inicial da aplicação desenvolvida. Basicamente, tem-se uma interface simples com a apresentação da marca do programa, sendo logo em seguida solicitado o preenchimento dos dados do paciente. Com a finalização do preenchimento dos dados, o usuário deve clicar no botão **Diagnosticar**.

Figura 10: Tela Inicial do RBFDIAG



Fonte: Elaborado Pelo Autor

Após a realização do clique no botão **Diagnosticar**, o sistema apresentará logo abaixo deste botão o resultado do diagnóstico com base no modelo discutido. Caso o diagnóstico seja positivo, é apresentado a mensagem: “O Diagnóstico para o câncer de mama foi positivo, procure um especialista.”.

Caso o resultado do diagnóstico seja negativo, é apresentado a mensagem: “O Diagnóstico para o câncer de mama foi negativo.”. Além do que foi debatido, é possível verificar na Figura 10 a existência de um menu na parte superior do programa. O botão **Sair** deste menu finaliza o programa em execução.

Por sua vez, o botão **Sobre** abre uma nova janela com uma breve descrição do programa. No Anexo I, encontram-se alguns dados de pacientes que podem ser utilizados para testar o programa.

4 CONCLUSÕES

A partir da análise realizada neste trabalho, foi possível perceber que a área de diagnóstico do câncer de mama a partir da realização de exames de rotina é bastante promissora, na medida em que foi possível obter bons resultados estatísticos para a classificação do conjunto de validação a partir de uma rede neural de função de base radial.

Com relação ao melhor modelo, considerando-se resultados de execução individual, foi possível encontrar um modelo que errou apenas 2 classificações no conjunto de validação, o que corresponde a apenas 5,88 % deste conjunto, apresentando assim sensibilidade, especificidade e acurácia próximas a 100%, e também próximo aos resultados dos trabalhos presentes na literatura.

Analisando-se agora os preditores dispostos na base de dados utilizada, é possível concluir que as variáveis: glicose, resistina, idade e índice de massa corporal são as que se mostraram mais adequadas para o problema de classificação da presença do câncer de mama no paciente, uma vez que todos os resultados ótimos de cada modelo estavam associados a grande parte destas 4 variáveis. Com relação ao uso da técnica PCA, foi possível constatar que de maneira geral ela não trouxe melhorias significativa para o desempenho dos modelos.

4.1 Trabalhos Futuros

Para trabalhos futuros, sugere-se a tentativa de otimização dos resultados de cada modelo empregado, bem como o teste de outros modelos com potencialidades de obtenção de melhores resultados. Além disso, é de crucial importância investir na ampliação da base de dados utilizada de forma a conseguir mais instancias para validação e treinamento, bem como avaliar o impacto de outras variáveis que podem ser coletadas em exames de rotina e que possam contribuir com o diagnóstico do câncer de mama. Testar todos as combinações de variáveis de entrada pode também acarretar em uma descoberta de uma configuração de entradas mais adequada.

Outro ponto importante a ser analisado é acrescentar outros tipos de doenças na base de dados, uma vez que os indivíduos que foram utilizados no treinamento dos modelos ou possuíam apenas a doença do câncer de mama, ou não possuíam nenhuma doença. Este fato, pode acarretar em um baixo desempenho do modelo ao ser aplicado em uma população não controlada, uma vez que existe a possibilidade de o modelo confundir o câncer de mama com outras doenças, ou até mesmo não detectar a presença do câncer de mama devido a presença de uma outra doença qualquer.

4.2 Trabalhos submetidos

DUARTE, F. L. C.; SOUZA, P. T. V. - Detecção do Câncer de Mama Utilizando Redes RBF e Exames de Rotina. IX Conferência Nacional em Comunicações, Redes e Segurança da Informação. Petrolina/PE – 18, 19 e 20 de Outubro.

REFERÊNCIAS

ANDRADE, F.; PAIVA, A.; CORREA, A. **Análise de Imagens de Termografia Dinâmica para Classificação de Alterações na Mama Usando Séries Temporais**. In: SIBGRAPI17. 2017.

ARAÚJO, J. M. F. R. **Inteligência Artificial**. 2012. Disponível em: http://www.dsc.ufcg.edu.br/~joseana/IAPos_NA17.pdf. Acesso em: 11/08/2019.

BATISTA, L. B. **Redes Neurais**. 2003. Disponível em: http://www.dsc.ufcg.edu.br/~hmg/disciplinas/graduacao/rn-2014.2/RedesNeurais-luana.ppt. Acesso em: 22/07/2019.

BORCHARTT, T. B. **Análise de imagens termográficas para a classificação de alterações na mama**. Tese (Doutor em Computação Visual) - Universidade Federal Fluminense. Niterói, p. 116. 2013.

CÂMARA, F. P. C. **PSIQUIATRIA E ESTATÍSTICA V: VALIDAÇÃO DE PROCEDIMENTOS DIAGNÓSTICA PELA CURVA R.O.C**. 2009. Disponível em: http://www.polbr.med.br/ano09/cpc0409.php. Acesso em: 23/07/2019

CAVALCANTI, G. D. C. **k-Nearest Neighbor**. 2010. Disponível em: http://200.17.137.109:8081/novobsi/Members/giordano/aulas/2014.1/computacao-inteligentes-sistemas-inteligentes/4.kNN.pdf. Acesso em: 22/07/2019.

EDUARDO. **Redes Neurais**. 2018a. Disponível em: http://www.facom.ufu.br/~backes/pgc204/Aula07-RedesNeurais.pdf. Acesso em: 22/07/2019.

EDUARDO. **Support Vector Machine – SVM**. 2018b. Disponível em: http://www.facom.ufu.br/~backes/pgc204/Aula08-SVM.pdf. Acesso em: 22/07/2019.

EDUARDO. **Classificadores Elementares**. 2018c. Disponível em: http://www.facom.ufu.br/~backes/pgc204/Aula03-ClassificadoresElementares.pdf. Acesso em: 13/08/2019.

CONSULTORIA, A. **Análise de Componentes Principais**. 2017. Disponível em: http://www.abgconsultoria.com.br/blog/analise-de-componentes-principais/. Acesso em: 22/07/2019.

GONÇALVES, C. B. **Detecção de câncer de mama utilizando imagens termográficas**. Trabalho de Conclusão de Curso (Bacharel em Ciência da Computação) - UFU. Uberlândia, p. 59. 2017.

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

HOLSBACH, N.; FOGLIATTO, F. S.; ANZANELLO, M. J. **Método de mineração de dados para identificação de câncer de mama baseado na seleção de variáveis**. Ciência & Saúde Coletiva, v. 19, p. 1295-1304, 2014.

INCA. **Conceito e Magnitude do câncer de mama**. Rio de Janeiro: INCA, 2018. Disponível em: https://www.inca.gov.br/controle-do-cancer-de-mama/conceito-e-magnitude. Acesso em: 03/05/2019.

LIMA, E. S. **Aula 16 – K-Nearest Neighbor (KNN)**. 2012. Disponível em: http://edirlei.3dgb.com.br/aulas/ia_2012_1/IA_Aula_16_KNN.pdf. Acesso em: 08/05/2019.

MATHWORKS. **newrb**. 2019. Disponível em: https://www.mathworks.com/help/deeplearning/ref/newrb.html. Acesso em: 13/08/2019.

MOREIRA, W. B. **Capítulo 5: Artigos sobre Testes Diagnósticos**. 2011. Disponível em: https://www.sboc.org.br/leitura-critica-de-artigos-cientificos?task=callelement&format=raw&item_id=99&element=f85c494b-2b32-4109-b8c1-083cca2b7db6&method=download&args[0]=58c19ae9051e72993587cd31b2ebee90. Acesso em: 23/07/2019.

MQL5. **APRENDIZAGEM DE MÁQUINA: COMO AS MÁQUINAS DE VETORES DE SUPORTE PODEM SER UTILIZADAS NAS NEGOCIAÇÕES**. 2014. Disponível em: https://www.mql5.com/pt/articles/584. Acesso em: 10/05/2019.

ONCOGUIA. **Câncer de mama tem 95% de chance de cura se diagnosticado precocemente.** 2017. Disponível em: http://www.oncoguia.org.br/conteudo/cancer-de-mama-tem-95-de-chance-de-cura-se-diagnosticado-precocemente/11292/7/. Acesso em: 05/05/2019.

PAHO. **Folha informativa – Câncer**. 2018. Disponível em: https://www.paho.org/bra/index.php?option=com_content&view=article&id=5588:folha-informativa-cancer&Itemid=1094. Acesso em: 03/05/2019.

PATRÍCIO, M.; PEREIRA, J.; CRISÓSTOMO, J.; MATAFOME, P.; GOMES, M.; SEIÇA, R.; CARAMELO, F. **Using Resistin, glucose, age and BMI to predict the presence of breast cancer**. BMC cancer, v. 18, n. 1, p. 29, 2018.

RADIOCENTRO, B. **Informes sobre Raios X aos pacientes**. 2011. Disponível em: http://www.radiocentro.com.br/blog/informacoes-sobre-raios-x-aos-pacientes. Acesso em: 11/08/2019.

SANTOS, R. C.; CASAGRANDE, L. C. S.; CROTTI, Y.; MARCELINO, R.; GRUBER, V. **Rede neural artificial otimizada para classificação de câncer de mama**. In: XV Congresso Brasileiro de Informática em Saúde. Goiânia, 2016.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. **Redes neurais artificiais para engenharia e ciências aplicadas**. São Paulo: Artliber, 2010.

SILVA, J. C.; LIMA, F. P. A.; LOPES, M. L. M.; MINUSSI, C. R. **Utilizando uma Rede Neural Artificial ARTMAP-Fuzzy para Realizar o Diagnóstico Clínico de Amostras de Câncer de Mama**. Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 2, n. 1, 2014.

WHO. **Estimated age-standardized incidence and mortality rates (World) in 2018, worldwide, both sexes, all ages**. Disponível em: https://gco.iarc.fr. Acesso em: 03/05/2019.

ZUBEN, F. J. V.; ATTUX, R. R. F. **Análise de Componentes Principais**. 2010. Disponível em:

ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia004_1s10/notas_de_aula/topico5_IA004_1s2010.pdf. Acesso em: 22/07/2019.

ANEXO A
CONJUNTO DE DADOS DE VALIDAÇÃO PARA O PROGRAMA DESENVOLVIDO

Glicose [mg/dL]	Resistina [ng/mL]	Idade [anos]	IMC [kg/m²]	Diagnóstico Real
74	28,0323	45	20,83	doente
93	5,57055	51	19,13265	doente
95	15,73606	38	22,49964	doente
86	10,34455	54	24,21875	doente
114	4,62	44	19,56	doente
105	4,82	72	23,62	doente
112	42,7447	71	25,5102	doente
98	53,6717	42	29,29688	doente
87	24,24591	52	30,80125	doente
131	11,50005	60	31,23141	doente
70	20,76801	49	29,77778	doente
99	23,03306	44	27,88762	doente
104	49,24184	71	27,91552	doente
108	16,48508	69	28,44444	doente
88	18,35574	74	28,65014	doente
90	15,55625	45	29,38476	doente
152	11,73	75	30,48	doente
91	9,27715	82	23,12467	sadio
77	12,9361	89	22,7	sadio
97	6,28445	73	22	sadio
82	5,14	25	22,86	sadio
75	9,35	38	23,34	sadio
84	3,32	47	22,03	sadio
87	5,62592	34	31,97501	sadio
84	16,43706	35	30,27682	sadio
83	8,70448	45	37,03561	sadio
87	21,44366	28	35,85581	sadio
76	17,2615	77	35,58793	sadio
83	8,04375	76	29,21841	sadio
85	7,5767	75	27,3	sadio
93	11,78796	69	32,5	sadio
102	4,2989	71	30,3	sadio
90	6,7052	66	27,7	sadio

96	9,6135	85	26,6	sadio

Fonte: Patrício *et al.* (2018)