Detecção do risco de Diabetes em estágio inicial utilizando redes ELM e seleção de features baseada em algoritmo genético

Early stage Diabetes risk prediction using ELM and ga-based feature selection

DOI:10.34117/bjdv8n7-339

Recebimento dos originais: 23/05/2022
Aceitação para publicação: 30/06/2022

Lucas Vieira Araujo
Graduando em Ciência da Computação
Instituição: Departamento de Ciência da Computação - Universidade Estadual do Piauí (UESPI)
Endereço: Av. Nossa Senhora de Fátima, 1300, Nossa Senhora de Fátima, Parnaíba – PI
E-mail: lucas.vieira.ar@disroot.org

Matheus Henrique da Silva Miranda
Graduando em Ciência da Computação
Instituição: Departamento de Ciência da Computação - Universidade Estadual do Piauí (UESPI)
Endereço: Av. Nossa Senhora de Fátima, 1300, Nossa Senhora de Fátima, Parnaíba – PI
E-mail: w9m157ht7@mozmail.com

Matheus Henrique de Souza Fontenele
Graduando em Ciência da Computação
Instituição: Departamento de Ciência da Computação - Universidade Estadual do Piauí (UESPI)
Endereço: Av. Nossa Senhora de Fátima, 1300, Nossa Senhora de Fátima, Parnaíba – PI
E-mail: matheusdeveloper.henrique@gmail.com

Odilon Fernandes Damasceno Neto
Graduando em Ciência da Computação
Instituição: Departamento de Ciência da Computação - Universidade Estadual do Piauí (UESPI)
Endereço: Av. Nossa Senhora de Fátima, 1300, Nossa Senhora de Fátima, Parnaíba – PI
E-mail: odilondamasceno@protonmail.com

Josias Guimarães Batista
Mestre em Engenharia de Teleinformática
Instituição: Instituto Federal do Ceará (IFCE)
Endereço: Av. Treze de Maio, 2081, Benfica, Fortaleza - CE, CEP: 60040-531
E-mail: josiasbatista@ifce.edu.br

Alanio Ferreira de Lima
Mestre em Engenharia de Telecomunicações
Instituição: Departamento de Engenharia Elétrica - Universidade Federal do Ceará (UFC)
Endereço: Bloco, 705, CEP: 60455-760, Fortaleza - CE
E-mail: allanio007@gmail.com

Darielson Araújo de Souza
Doutor em Engenharia Elétrica
Instituição: Departamento de Ciência da Computação - Universidade Estadual do Piauí (UESPI)
Endereço: Av. Nossa Senhora de Fátima, 1300, Nossa Senhora de Fátima, Parnaíba - PI
E-mail: darielsondesouza@phb.uespi.br

RESUMO
A diabetes é considerada uma das maiores crises de saúde do século 21, e tem mostrado um crescimento significativo nos últimos anos, de acordo com a Federação Internacional de Diabetes. O diagnóstico em estágios iniciais é de grande importância no controle da doença, mas este se mostra um desafio devido à sutileza com que os sintomas são apresentados no início. O presente trabalho teve como objetivo validar um método de análise automatizada dos sintomas para auxiliar na detecção do risco de diabetes em estágios iniciais. Uma rede neural ELM foi utilizada com o auxílio de seleção de features realizada com algoritmo genético e os resultados foram comparados com os de algoritmos como MLP, SVM e RBF. A acurácia média obtida com a ELM foi de 98,64%

Palavras-chave: Diabetes, estágios iniciais, extreme learning machine.

ABSTRACT
Diabetes is considered one of the greatest health crises of the 21st century, and it has shown a significant growth in recent years, according to the International Diabetes Federation. The diagnosis at early stages is of great importance in controlling the disease, but this is a challenge due to the subtlety with which the symptoms are presented at the beginning. The present work was aimed at validating a method for automated symptom analysis to aid in early-stage diabetes risk detection. An ELM neural network was used with feature selection performed by genetic algorithm, and the results were compared with those of algorithms such as MLP, SVM and RBF. The accuracy obtained with ELM was 98.64%.

Keywords: Diabetes, early stage, extreme learning machine.

1 INTRODUÇÃO

De um ponto de vista médico, Diabetes Mellitus (DM) é o nome coletivo de um grupo de desordens metabólicas associadas à hiperglicemia crônica ocasionadas pela falha na produção e/ou utilização de insulina (EGAN, Aoife M.; DINNEEN, Seán F, 2019). Em 2021, era estimado que 537 milhões de pessoas ao redor do mundo tinham

diabetes, um número que tende a crescer para 643 milhões até 2030 e 783 milhões até 2045 (ATLAS *et al*., 2021).

O diagnóstico de diabetes em estágios iniciais é essencial para o seu controle (ISLAM *et al*., 2020), uma vez que embora seja incurável, esta doença pode ser tratada através de medicações (LARABI-MARIE-SAINTE *et al*., 2019) de forma a evitar as complicações adicionais que surgem em sua decorrência, diminuindo os riscos associados e melhorando a qualidade de vida do paciente. Contudo, os sintomas podem ser tão sutis no início que mesmo médicos experientes não são capazes de identificá-los corretamente (CHAKI *et al.*, 2020).

Algoritmos de Machine Learning permitem a análise inteligente de dados e sua atual tecnologia está bem adaptada para o estudo de dados médicos (MALIK; SINGH; RANA, 2022). Devido a avanços nas técnicas de Inteligência Artificial, o diagnóstico de diabetes por meio de programas automáticos baseados em Aprendizado de Máquina tem se mostrado mais eficiente do que métodos manuais (CHAKI *et al*., 2020).

O presente trabalho busca validar um método automático baseado em Aprendizado de Máquina para análise de sintomas de forma a auxiliar na detecção do risco de diabetes em estágio inicial. Mais especificamente, busca-se: (a) identificar os sintomas que mais contribuem para a detecção de diabetes; (b) treinar um classificador para realizar a detecção automática; e (c) validar o classificador com base na comparação de sua acurácia com a de outros modelos.

Uma rede neural Extreme Learning Machine (ELM) foi treinada utilizando o dataset UCI Early stage diabetes risk prediction dataset, o qual foi pré-processado realizando uma seleção de features por meio de algoritmo genético. Além disso, três outros classificadores, a citar MultiLayer Perceptron (MLP), Support Vector Machine (SVM) e Radial Basis Function (RBF), também foram utilizados para fins de comparação. Para validar o modelo, foi adotada uma validação cruzada do tipo k-fold, onde a acurácia média da ELM com e sem seleção de features foi comparada à dos demais algoritmos previamente citados.

O trabalho está dividido em 5 seções, incluindo esta. A seção 2 trará alguns trabalhos relacionados, reforçando a base teórica consultada na elaboração do artigo. Os métodos e técnicas da pesquisa serão explicados na seção 3. Já na seção 4, os resultados são apresentados e discutidos com base na literatura. Por fim, a seção 5 trará as considerações finais e sugestões de trabalhos futuros.

2 TRABALHOS RELACIONADOS

O trabalho de (ISLAM *et al*., 2020) montou o dataset para detecção de diabetes em estágio inicial e o analisou aplicando Naive Bayes, Logistic Regression e Random Forest, objetivando uma ferramenta para predição do risco da doença. Foi adotada uma validação cruzada estratificada e como melhor resultado foi obtida uma acurácia de 97.4% com a Random Forest.

Em (JULIUS; AYOKUNLE; IBRAHIM, 2021) foi utilizado o software WEKA, comparando os algoritmos KNN, SVM, Functional Tree e Random Forest, buscando avaliar a aplicação destes para o mesmo problema, sendo obtido um resultado de 98.08% de acurácia com o K-NN.

No artigo de (VAKIL *et al*., 2021) foi feita uma análise comparativa dos algoritmos Random Forest, Decision Tree, MLP, K-NN, SVM e XG Boost, aplicando atribuição de features com o método SHAP (SHapley Additive exPlanations) para o mesmo dataset com o objetivo de explicar as predições destes modelos analisando a contribuição de cada atributo para a classificação geral. Em seus resultados é dito que a Random Forest obteve o melhor desempenho dentre os métodos testados, com uma taxa de acerto de 99%, porém sua metodologia para comparação resumiu-se em escolher a melhor acurácia encontrada por cada algoritmo após um número de testes.

Mais recentemente, (ELSAYED; ELSAYED; OZER, 2022) aplicou a ELM com o objetivo de desenvolver um modelo computacionalmente eficiente na detecção de diabetes em estágios iniciais, tendo utilizado a biblioteca Scikit-learn estendida para sua implementação e obteve acurácia de 98.07% com zero falso-positivos.

Embora exista um número de trabalhos anteriores voltados ao mesmo problema, ainda há uma clara lacuna a ser preenchida neste contexto, seja buscando melhorar a acurácia ou mesmo a eficiência da predição do risco de diabetes. Em particular, o uso de redes ELM e suas variantes, assim como a aplicação de seleção de features neste dataset, ainda é praticamente inexplorado.

3 MÉTODOS E TÉCNICAS

3.1 DATASET

O dataset “Early Stage Diabetes Risk Prediction Dataset” tem origem no trabalho de (ISLAM *et al*., 2020) e foi montado com base em questionários diretos aplicados a pacientes do Sylhet Diabetes Hospital em Sylhet (Blangladesh), tendo sido posteriormente disponibilizado no repositório online UCI Machine Learning, de onde foi

obtido para este trabalho. Há 520 instâncias compreendendo pessoas entre 16 e 90 anos de idade, tanto do sexo masculino quanto feminino, das quais 320 possuem diagnóstico positivo para diabetes e o restante possui diagnóstico negativo. Há 16 atributos e uma classe, como ilustrado na Tabela 1.

Tabela 1. Descrição do dataset

Atributos	**Valores**
Age	16-90
Gender	Male/Female
Polyuria	Yes/No
Polydipsia	Yes/No
Sudden weight loss	Yes/No
Weakness	Yes/No
Polyphagia	Yes/No
Genital Trush	Yes/No
Visual blurring	Yes/No
Itching	Yes/No
Irritability	Yes/No
Delayed healing	Yes/No
Partial paresis	Yes/No
Muscle stiffness	Yes/No
Alopecia	Yes/No
Obesity	Yes/No
Class	Positive/Negative

Após a limpeza e eliminação de dados faltantes, ainda restaram 514 instâncias. O dataset foi codificado, com os atributos binários sendo substituídos pelos valores 1 e 0, enquanto a idade foi mantida no seu intervalo original.

3.2 SELEÇÃO DE FEATURES COM ALGORITMO GENÉTICO

A seleção de features foi realizada com o objetivo de identificar os atributos (sintomas) de maior relevância na classificação, de forma a diminuir as dimensões do

dataset e possivelmente aumentar a acurácia do algoritmo utilizado. Considerando que cada um dos 16 atributos pode ser ou mantido (1) ou eliminado (0), o cromossomo que representa uma seleção de features pode ser codificado na forma de um número binário de 16 bits.

Uma vez que o cromossomo foi especificado, é necessário descrever a função objetivo, também chamada função fitness, que indica o aspecto a ser otimizado no problema alvo. Para isso, consideramos a acurácia obtida ao realizar uma classificação, a qual pode ser expressada como:

$$A = \frac{VP + FN}{FP + FN + VP + VN}$$

Onde VP, VN, FP e FN indicam, respectivamente, o número de Verdadeiros Positivos, Verdadeiros Negativos, Falsos Positivos e Falsos Negativos resultantes da classificação. O cálculo de fitness era realizado treinando uma rede neural ELM utilizando apenas as features selecionadas e 80% das entradas do dataset, anotando-se a acurácia obtida ao classificar os 20% restantes. Esse processo era repetido 10 vezes com diferentes permutações do dataset e a média dos valores era utilizada como fitness.

O critério de seleção era baseado no método da roleta, com cada indivíduo tendo chance de reprodução diretamente proporcional ao seu fitness. O cruzamento era realizado escolhendo um ponto aleatório no cromossomo, em torno do qual features eram extraídas de ambos os pais, formando dois novos indivíduos.

Esses indivíduos poderiam então sofrer mutação, o que consistia na inversão do bit em alguma posição aleatória do cromossomo. Por fim, o filho resultante de melhor fitness seria mantido para integrar a população substituindo o pior indivíduo atualmente nela, mas apenas se fosse melhor do que este (elitismo).

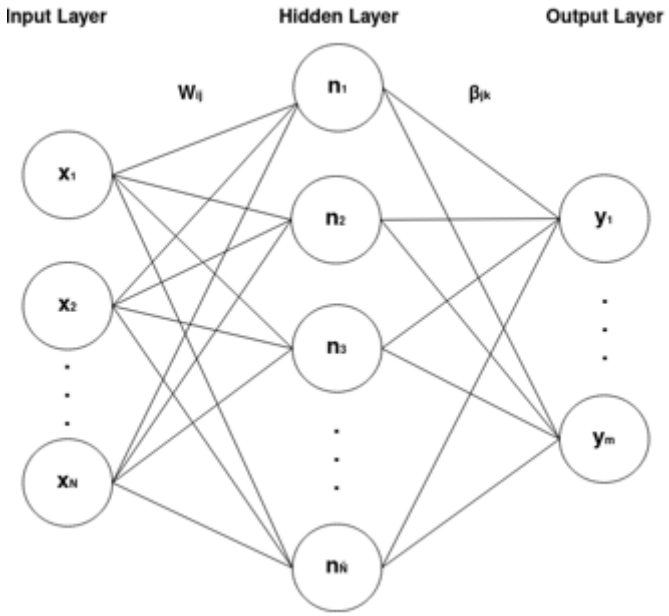
Os parâmetros do GA foram uma população inicial de 500 indivíduos, evoluindo ao longo de 200 gerações com taxa de cruzamento de 0.8 e mutação de 0.5. No final, a seleção de features que apresentou melhores resultados foi: Gender, Polyuria, Polydipsia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, muscle stiffness, Alopecia e Obesity.

3.3 EXTREME LEARNING MACHINE (ELM)

A ELM é um método de Aprendizado de Máquina proposto por (HUANG; ZHU; SIEW, 2004) e consiste em uma Single-Layer Feed-forward Neural Network (SLFN) que não requer ajuste de pesos no processo de treinamento.

Uma vez que os pesos de entrada são escolhidos aleatoriamente, a SLFN pode ser considerada um sistema de equações lineares e os pesos de saída são determinados analiticamente pela solução desse sistema, o que é feito utilizando a inversa generalizada da matriz de saída da hidden layer (HUANG; ZHU; SIEW, 2004). A figura 1 apresenta a arquitetura básica de uma rede ELM com Ñ neurônios:

Figura 1. Arquitetura ELM



Essa arquitetura pode ser expressada como $H\beta = T$, onde

$$H = \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_N \cdot x_N + b_{\tilde{N}}) \\ \vdots & \cdots & \vdots \\ g(w_1 \cdot x_N + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}}$$

é a matriz de saída da hidden layer, β=H†T e T é o vetor de saída da rede. A notação H† representa a pseudo-inversa de Moore-Penrose calculada sobre a matriz H, de forma a encontrar a solução do sistema de equações lineares, minimizando o quadrado da diferença entre a saída aproximada e a pretendida.

Conforme (HUANG; ZHU; SIEW, 2006), os pesos de entrada w e o bias b podem ser atribuídos aleatoriamente desde que a função de ativação g seja infinitamente diferenciável. Assim, a função utilizada neste trabalho foi a curva logística (sigmoidal) definida como

$$g(x) = \frac{1}{1 + e^{-x}}$$

3.4 MULTILAYER PERCEPTRON (MLP)

A Multi-Layer Perceptron (MLP) é considerado o tipo mais básico de rede neural e consiste em um mapeamento não-linear entre os vetores de entrada (Input Layer) e de saída (Output Layer), entre os quais pode haver qualquer número de camadas intermediárias, chamadas de Hidden Layers. O treinamento da MLP é feito ajustando os pesos entre as conexões de cada camada por meio de algum algoritmo de otimização, sendo o método Backpropagation (BP) o mais comum.

BP consiste na propagação do erro em sentido oposto à transmissão de informação entre as Hidden Layers, reajustando os pesos das conexões até que um erro mínimo seja atingido (RUMELHART; HINTON; WILLIAMS, 1985). Assim como outros algoritmos baseados em gradiente em descida, BP é geralmente muito lento devido a etapas de aprendizado impróprias e pode convergir facilmente para mínimos locais, sendo necessárias muitas iterações para obter melhor performance no aprendizado (HUANG; ZHU; SIEW, 2004).

3.5 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) é um algoritmo geral de classificação e regressão que mapeia os vetores de entrada em um espaço multidimensional Z por meio de um mapeamento não-linear escolhido a priori, construindo uma margem decisória máxima que divide os vetores de classes distintas em diferentes hiperplanos (CORTES; VAPNIK, 1995).

3.6 RADIAL BASIS FUNCTION (RBF)

Radial Basis Function (RBF) é uma SLFN que aprende de forma supervisionada em apenas um estágio, sem a necessidade de ajustar parâmetros iterativamente. Redes RBF são baseadas num método de interpolação que funciona construindo um espaço de

funções lineares dependentes das posições de pontos de dados conhecidos (os chamados centróides), de acordo com uma medida arbitrária de distância (BROOMHEAD; LOWE, 1988).

4 RESULTADOS

As simulações foram realizadas em um computador rodando Debian 10, equipado com 4GB RAM e um processador Intel(R) Core(TM) i3-8130U CPU @ 2.20GHz. As implementações da MLP e da SVM são as disponíveis na biblioteca Scikit-Learn versão 1.0.2, enquanto a ELM e RBF foram implementadas pelo próprio autor utilizando a linguagem Python 3.9, com a biblioteca Numpy na versão 1.22.4 para manipulação de matrizes.

A rede MLP possui apenas uma camada oculta, sendo esta composta por 256 neurônios, a mesma quantidade da ELM. A Tabela II apresenta os resultados encontrados utilizando uma validação cruzada estratificada do tipo k-fold com k=10.

Tabela 2. 10-fold stratified cross validation

Fold	ELM (FS)	ELM	MLP	SVM	RBF
1	100%	92.03%	90.03%	94.23%	59.61%
2	100%	90.38%	98.07%	96.15%	80.76%
3	98.07%	94.23%	96.15%	94.23%	84.61%
4	94.23%	90.38%	92.30%	94.23%	84.61%
5	100%	88.23%	94.11%	94.11%	74.50%
6	98.03%	94.11%	90.19%	94.11%	64.79%
7	98.03%	90.19%	94.11%	92.15%	66.66%
8	100%	98.03%	92.15%	94.11%	84.31%
9	98.03%	86.27%	90.19%	90.19%	60.78%
10	100%	98.03%	92.15%	92.15%	74.50%
Média	**98.64%**	92.22%	93.57%	93.57%	73.51%

Onde ELM (FS) representa a ELM quando auxiliada pela seleção de features e Fold representa a parte do dataset utilizada como validação em cada iteração do k-fold. Ao final da tabela, a acurácia média de cada algoritmo é apresentada.

Com base nesses testes, pode-se observar que a rede MLP apresenta bons resultados mesmo ao utilizar apenas uma camada oculta, tendo uma acurácia mínima de 90.03% e uma média de 93.57%.

A SVM teve um desempenho bastante similar ao da rede MLP, com uma acurácia significativamente maior do que a reportada por (VAKIL *et al*., 2021), o que poderia ser atribuído à adoção de um kernel RBF ao invés de um sigmóide.

O desempenho da rede RBF, no entanto, mostrou-se inferior ao que era esperado, atingindo uma máxima acurácia de 84.61% e uma média de 73.51%, tendo portanto os piores resultados encontrados neste estudo.

A rede ELM aplicada ao dataset normal, obteve acurácia inferior à SVM e MLP, com uma média de 92.22%. Estes resultados são inferiores aos de (ELSAYED; ELSAYED; OZER, 2022), porém diferenças na arquitetura e implementação podem ser as responsáveis por este decréscimo.

Já quando auxiliada pela seleção de features, a ELM obteve em média 98.64% de acurácia, a maior dentre os algoritmos testados, tendo em alguns momentos atingido a marca de 100% de acerto, o que serve para ilustrar seu potencial na detecção do risco de diabetes em estágios iniciais.

5 CONSIDERAÇÕES FINAIS

A diabetes tem apresentado um rápido crescimento ao redor do mundo e seu diagnóstico em estágios iniciais é de grande importância para combatê-la, porém devido à sutileza com que os sintomas da doença se apresentam no início, este se mostra um desafio até mesmo para os profissionais mais experientes. Atualmente, avanços nas técnicas de Inteligência Artificial tornaram a detecção de diabetes por meio de modelos de Aprendizado de Máquina mais eficientes do que métodos tradicionais.

Neste trabalho, estudou-se a aplicação de redes neurais ELM auxiliadas por uma seleção de features realizada com Algoritmo Genético na detecção do risco de diabetes em estágio inicial, analisando o dataset originado em (ISLAM *et al*., 2020). A seleção de features possibilitou a redução das dimensões do dataset, além de permitir identificar atributos (sintomas) que possuem maior contribuição na detecção do risco de diabetes, alcançando assim o objetivo (a) da pesquisa.

Ao treinar o classificador ELM e validá-lo por meio da comparação de seus resultados com os de outros modelos, foram atingidos tanto os objetivos (b) quanto (c), sendo possível ilustrar o potencial das redes ELM quando aplicadas a esse problema. A

acurácia média atingida pelo modelo proposto foi de 98.64%, sendo o maior dentre os algoritmos analisados.

Como proposta para trabalhos futuros, poderiam envolver a utilização de diferentes funções de ativação, combinadas com métodos mais comuns de seleção de features, analisando-se a relação entre estas e o número de neurônios na camada oculta da rede. Além disso, variações das redes ELM, tais como a Multilayer ELM (M-ELM) ou Hierarchical ELM (H-ELM), poderiam ser aplicadas ao mesmo dataset, possivelmente melhorando a acurácia da classificação.

REFERÊNCIAS

EGAN, Aoife M.; DINNEEN, Seán F. **What is diabetes?**. Medicine, v. 47, n. 1, p. 1-4, 2019.

ATLAS, Diabetes et al. **International diabetes federation. IDF Diabetes Atlas, 10th edn**. Brussels, Belgium: International Diabetes Federation, 2021.

ISLAM, M. M. et al. **Likelihood prediction of diabetes at early stage using data mining techniques**. In: Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. p. 113-125.

LARABI-MARIE-SAINTE, Souad et al. **Current techniques for diabetes prediction: review and case study**. Applied Sciences, v. 9, n. 21, p. 4604, 2019.

CHAKI, Jyotismita et al. **Machine learning and artificial intelligence based diabetes mellitus detection and self-management: a systematic review**. Journal of King Saud University-Computer and Information Sciences, 2020.

MALIK, Harshit; SINGH, Himanshu; RANA, Ashish. **Disease Detection Using Symptoms Based on Machine Learning**. International Journal of Recent Advances in Multidisciplinary Topics, v. 3, n. 5, p. 143-149, 2022.

JULIUS, Adetunji Olusogo; AYOKUNLE, Ayeni Olusola; IBRAHIM, Fasanya Olawale. **Early Diabetic Risk Prediction using Machine Learning Classification Techniques**. International Journal of Innovative Science and Research Technology, v. 6, n. 9, p. 502, 2021.

VAKIL, Vishal et al. **Explainable predictions of different machine learning algorithms used to predict Early Stage diabetes**. arXiv preprint arXiv:2111.09939, 2021.

ELSAYED, Nelly; ELSAYED, Zag; OZER, Murat. **Early Stage Diabetes Prediction via Extreme Learning Machine**. In: SoutheastCon 2022. IEEE, 2022. p. 374-379.

HUANG, Guang-Bin; ZHU, Qin-Yu; SIEW, Chee-Kheong. **Extreme learning machine: a new learning scheme of feedforward neural networks**. In: 2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541). Ieee, 2004. p. 985-990.

HUANG, Guang-Bin; ZHU, Qin-Yu; SIEW, Chee-Kheong. **Extreme learning machine: theory and applications**. Neurocomputing, v. 70, n. 1-3, p. 489-501, 2006.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. **Learning internal representations by error propagation**. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

CORTES, Corinna; VAPNIK, Vladimir. **Support-vector networks**. Machine learning, v. 20, n. 3, p. 273-297, 1995.

BROOMHEAD, D.; LOWE, D. **Multivariable functional interpolation and adaptive networks, complex systems**, vol. 2. 1988.