



## Diagnóstico da chikungunya por inteligência artificial e dados de prontuários

### Chikungunya diagnosis using artificial intelligence and medical record data

### Diagnóstico de chikungunya mediante inteligencia artificial y datos de registros

Cecilia Cordeiro da Silva<sup>1</sup>, Ana Clara Gomes da Silva<sup>2</sup>, Clarisse Lins de Lima<sup>1</sup>, Maíra Araújo de Santana<sup>1</sup>, Juliana Carneiro Gomes<sup>1</sup>, Giselle Machado Magalhães Moreno<sup>1</sup>, Karla Amorim Sancho<sup>1</sup>, Heloísa Ramos Lacerda de Melo<sup>3</sup>, Marcela Franklin Salvador de Mendonça<sup>3</sup>, Wellington Pinheiro dos Santos<sup>1</sup>

1 Doutora, Departamento de Engenharia Biomédica, Universidade Federal de Pernambuco, Recife, Pernambuco, Brasil.

2 Mestra, Programa de Pós-Graduação em Engenharia da Computação, Escola Politécnica da Universidade de Pernambuco, Recife, Pernambuco, Brasil.

3 Doutora, Centro de Ciências Médicas, Universidade Federal de Pernambuco, Recife, Pernambuco, Brasil.

Autor correspondente: Prof. Wellington Pinheiro dos Santos

E-mail: [wellington.santos@ufpe.br](mailto:wellington.santos@ufpe.br)

### Resumo

**Objetivo:** O propósito desta pesquisa é desenvolver e avaliar um modelo de aprendizado de máquina para auxiliar no diagnóstico clínico da febre chikungunya utilizando dados de prontuários de pacientes. **Método:** Os dados foram obtidos do Portal de Dados Abertos da Prefeitura de Recife, englobando 18.881 registros de pacientes. **Resultados:** Após o pré-processamento e a validação cruzada, o modelo *Random Forest* com 100 árvores apresentou a melhor performance, com acurácia de 93,40% e área característica receptor-operador de 0,990. A aplicação do modelo demonstrou alta eficácia na diferenciação entre chikungunya e outras condições. **Conclusão:** Concluímos que o uso de inteligência artificial pode melhorar significativamente o diagnóstico clínico de arboviroses. Futuros trabalhos incluem a expansão da base de dados, a integração do modelo em ambientes clínicos e a exploração de técnicas avançadas de aprendizado de máquina.

**Descritores:** Diagnóstico clínico; Febre chikungunya; Inteligência artificial



## Abstract

**Objective:** The purpose of this research is to develop and evaluate a machine learning model to assist in the clinical diagnosis of chikungunya fever using patient medical records. **Method:** Data were obtained from the Open Data Portal of Recife City Hall, encompassing 18,881 patient records. **Results:** After preprocessing and cross-validation, the Random Forest model with 100 trees showed the best performance, with an accuracy of 93.40% and a receiver-operator characteristic area of 0.990. The model application demonstrated high efficacy in differentiating between chikungunya and other conditions. **Conclusion:** We conclude that the use of artificial intelligence can significantly improve the clinical diagnosis of arboviruses. Future work includes expanding the database, integrating the model into clinical environments, and exploring advanced machine learning techniques.

**Keywords:** Clinical diagnosis; Chikungunya fever; Artificial intelligence

## Resumen

**Objetivo:** El propósito de esta investigación es desarrollar y evaluar un modelo de aprendizaje automático para ayudar en el diagnóstico clínico de la fiebre chikungunya utilizando registros médicos de pacientes. **Método:** Los datos se obtuvieron del Portal de Datos Abiertos del Ayuntamiento de Recife, abarcando 18.881 registros de pacientes. **Resultados:** Después del preprocesamiento y la validación cruzada, el modelo Bosque Aleatorio con 100 árboles mostró el mejor rendimiento, con una precisión del 93,40% y un área bajo la curva característica receptor-operador de 0,990. La aplicación del modelo demostró una alta eficacia en la diferenciación entre chikungunya y otras condiciones. **Conclusión:** Concluimos que el uso de inteligencia artificial puede mejorar significativamente el diagnóstico clínico de las arbovirosis. Los trabajos futuros incluyen la expansión de la base de datos, la integración del modelo en entornos clínicos y la exploración de técnicas avanzadas de aprendizaje automático.

**Descriptor:** Diagnóstico clínico; Fiebre chikungunya; Inteligencia artificial



## Introdução

Arboviroses são doenças causadas por arbovírus, que são vírus transmitidos por artrópodes, como os mosquitos. Entre as arboviroses mais conhecidas estão a dengue, a febre chikungunya e a zika. Todas são transmitidas principalmente pelo mosquito *Aedes aegypti* [1-5]. Essas doenças representam um significativo desafio para a saúde pública, especialmente em regiões tropicais e subtropicais, onde as condições ambientais favorecem a proliferação do vetor [1-5]. Os surtos de arboviroses podem sobrecarregar os sistemas de saúde, causando um aumento expressivo na demanda por serviços médicos e gerando impacto econômico e social substanciais devido à morbidade e mortalidade associadas [1-5].

A febre chikungunya é particularmente difícil de diagnosticar clinicamente devido à semelhança de seus sintomas com os de outras arboviroses, especialmente a dengue [6-8]. Ambos podem causar febre alta, dores nas articulações, erupções cutâneas e outros sintomas inespecíficos, o que complica o diagnóstico diferencial baseado apenas na apresentação clínica [6-8]. Essa dificuldade diagnóstica pode atrasar o tratamento adequado e a implementação de medidas de controle de surtos, exacerbando os impactos negativos na saúde pública [6-8].

A Inteligência Artificial (IA) emerge como uma ferramenta poderosa para auxiliar no diagnóstico clínico de arboviroses, fornecendo diagnósticos mais precisos e rápidos a partir de grandes volumes de dados de prontuários médicos [9-12]. Utilizando técnicas de aprendizado de máquina é possível identificar padrões complexos e sutis nos dados clínicos, que podem passar despercebidos em análises tradicionais. A consideração dos dados de prontuários - incluindo sintomas detalhados e resultados de exames laboratoriais - é crucial para desenvolver modelos de IA robustos [9-12]. Esses dados oferecem uma visão abrangente e rica das condições dos pacientes, permitindo que os algoritmos de IA aprendam e diferenciem com maior precisão entre a febre chikungunya e outras doenças similares [9-12]. Assim, o uso de ferramentas de IA para apoiar o diagnóstico clínico pode melhorar significativamente a eficiência e a precisão dos diagnósticos, resultando em melhores resultados para os pacientes e em uma resposta mais eficaz às epidemias de arboviroses [9-12].

O objetivo desta pesquisa é desenvolver e avaliar um modelo de aprendizado de máquina que possa auxiliar no diagnóstico clínico preciso e precoce da febre



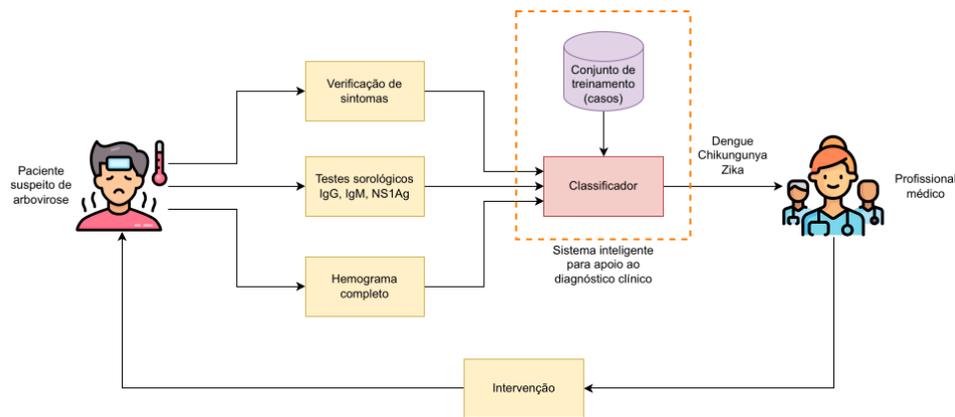
chikungunya utilizando dados de prontuários de pacientes. A partir da análise de sintomas, testes sorológicos e resultados de hemogramas, a pesquisa visa identificar padrões nos dados clínicos para diferenciar eficazmente a chikungunya de outras condições, como a dengue. A intenção é proporcionar uma ferramenta de apoio que melhore a acurácia e a rapidez no diagnóstico, contribuindo assim para a gestão eficiente e o tratamento adequado dos pacientes.

## Métodos

### Proposta

Esta pesquisa teve como objetivo desenvolver e avaliar uma máquina de aprendizado para apoiar o diagnóstico clínico da chikungunya utilizando dados de prontuário de pacientes. A proposta envolveu a aplicação de técnicas de aprendizado de máquina para identificar padrões nos dados clínicos e auxiliar na diferenciação entre casos de dengue e outras condições. Para isso, foram utilizados métodos de classificação, incluindo *Naive Bayes*, *Bayes Net*, *Árvore de decisão J48*, *Random Tree*, *Random Forest*, *Perceptron* multicamadas e *Máquina de Vetor de Suporte (SVM)*. O processo de desenvolvimento e avaliação do modelo foi conduzido em etapas: pré-processamento de dados, seleção de características, modelagem de aprendizado de máquina e avaliação de desempenho por meio de validação cruzada. A máquina de aprendizado resultante foi avaliada quanto à sua eficácia e eficiência no diagnóstico da Chikungunya, visando contribuir para uma abordagem mais precisa e precoce da doença. A Figura 1 a seguir apresenta o diagrama de blocos da proposta de sistema de apoio ao diagnóstico clínico de Dengue, Chikungunya e Zika.

**Figura 1** – Diagrama de blocos da proposta de sistema de apoio ao diagnóstico clínico das arboviroses Dengue, Chikungunya e Zika



Fonte: Elaborado pelos autores.

A ideia geral desta pesquisa - ilustrada pelo diagrama - é que o sistema receba as informações de sintomas, testes sorológicos e hemograma e a partir disso, classifique entre chikungunya e não chikungunya.

#### Base de dados

Os dados utilizados neste estudo foram obtidos por meio do Portal de Dados Abertos da Prefeitura da Cidade do Recife, desenvolvido pela EMPREL, Empresa Municipal de Informática. O objetivo desse portal é disponibilizar de forma pública o acesso e a busca de dados governamentais gerados por secretarias e órgãos da gestão municipal.

A Secretaria Municipal de Saúde disponibiliza 11 conjuntos de dados, incluindo informações sobre os Distritos Sanitários, Vigilância Sanitária, Serviço de Atendimento Móvel de Urgência (SAMU), Unidades de Saúde, Academias da Cidade, além de registros de casos de dengue, zika e chikungunya registrados nas unidades de saúde públicas e privadas. Os dados de registros de casos são atualizados diariamente.

Para o presente estudo foram utilizadas as informações anonimizadas do Sistema de Informação de Agravos de Notificação (SINAN), disponibilizadas para o município de Recife, no período de 2013 a 2021.

A base de dados possui 18.887 registros de pacientes. A distribuição dos casos é a que segue: Dengue clássica (1): 3.499 instâncias; Dengue com complicações (2): 1.389 instâncias; Descartado (5): 5.082 instâncias; Inconclusivo (8): 6 instâncias; e Chikungunya (13): 8.911 instâncias.



Esses dados foram fundamentais para a realização da análise e desenvolvimento do modelo de aprendizado de máquina para apoiar o diagnóstico clínico da chikungunya.

### **Pré-processamento dos dados**

Antes da análise e modelagem, os dados passaram por um processo de pré-processamento para prepará-los para o treinamento dos modelos de aprendizado de máquina. As seguintes etapas foram realizadas:

**Remoção de colunas não relevantes:** Foram retiradas as colunas que continham informações de data, códigos e endereços dos pacientes e das unidades de saúde. Essas informações não eram relevantes para a tarefa de classificação da dengue.

**Remoção de colunas com mais de 90% de dados faltantes:** Colunas que continham mais de 90% de dados faltantes foram eliminadas, pois não contribuíam significativamente para a análise.

**Imputação de dados faltantes:** Os dados faltantes restantes foram estimados pela média ou pela moda, dependendo do tipo de dado e da distribuição dos valores.

Após o pré-processamento, a base de dados resultante continha 18.881 registros de pacientes. Os casos de dengue foram agrupados em duas classes: dengue clássica e dengue com complicações. A distribuição dos casos na base de dados ficou da seguinte forma: Dengue clássica (1): 3.499 instâncias; Dengue com complicações (2): 1.389 instâncias; Descartado (5): 5.082 instâncias; e Chikungunya (13): 8.911 instâncias.

### **Balanceamento de classes**

Para abordar o desequilíbrio entre as classes de diagnóstico da dengue, foi empregada a técnica de subamostragem estratificada, conhecida como "*Spread Subsample*". Essa abordagem busca equilibrar o conjunto de dados removendo aleatoriamente registros das classes majoritárias até que todas as classes tenham o mesmo número de instâncias. O procedimento envolveu a divisão da base de dados em conjuntos de teste e treino, com 20% e 80%, respectivamente, seguido pela aplicação da subamostragem nas classes majoritárias, resultando em um conjunto de dados equilibrado com 4.444 registros.



Para avaliar o desempenho dos modelos de classificação, foi realizada uma validação cruzada com 10 folds, repetida 3 vezes para cada classificador, garantindo uma avaliação robusta. O melhor classificador foi selecionado com base em métricas de desempenho - como a área sob a curva ROC - e submetido a um teste final usando o conjunto de teste reservado. Essas estratégias foram adotadas para mitigar o desbalanceamento das classes e garantir uma avaliação precisa e confiável do modelo de diagnóstico da dengue.

## Classificação

No presente estudo foram utilizados os seguintes classificadores de aprendizado de máquina para apoiar o diagnóstico clínico da dengue, juntamente com seus respectivos parâmetros:

**Naive Bayes:** Um classificador probabilístico simples baseado no teorema de Bayes, que assume independência condicional entre os recursos. Não possui parâmetros ajustáveis.

**Bayes Net:** Uma extensão do Naive Bayes que modela as dependências entre os recursos através de uma rede bayesiana. Não possui parâmetros ajustáveis.

**Árvore de decisão J48:** Um algoritmo de aprendizado de máquina que constrói uma árvore de decisão a partir dos dados de treinamento. Seus principais parâmetros incluem critérios de divisão (como ganho de informação ou índice Gini), profundidade máxima da árvore e número mínimo de exemplos por folha.

**Random Tree:** Uma variação do algoritmo de árvore de decisão que introduz aleatoriedade no processo de construção da árvore, dividindo os dados em subconjuntos aleatórios e selecionando aleatoriamente as características para cada divisão.

**Random Forest:** Uma extensão do Random Tree que constrói múltiplas árvores de decisão e combina suas previsões por votação. Seu principal parâmetro é o número de árvores no conjunto, que varia entre 100, 200, 300, 400 e 500 árvores.

**Perceptron Multicamadas (MLP):** Uma rede neural artificial com múltiplas camadas de neurônios, incluindo camadas de entrada, camadas ocultas e camadas de saída. Seus principais parâmetros incluem o número de neurônios na camada escondida do modelo de duas camadas - variando entre 20, 50 e 100 neurônios - e o



número de neurônios nas camadas escondidas do modelo de três camadas, variando entre 20 e 20, 20 e 50, e 20 e 100 neurônios.

**Máquina de Vetor de Suporte (SVM):** Um classificador que busca encontrar o hiperplano de separação ótimo entre as classes. Seus parâmetros incluem o tipo de kernel (polinomial linear e grau 2, e RBF com  $G = 0,01, 0,25$  e  $0,5$ ).

Esses classificadores foram selecionados devido à sua capacidade de lidar com problemas de classificação e foram configurados com uma variedade de parâmetros para explorar diferentes configurações e encontrar a melhor abordagem para o diagnóstico da dengue.

## Resultado e discussão

A Tabela 1 apresenta os resultados dos métodos de classificação *Naive Bayes*, *Bayes Net*, *Árvore de decisão J48*, *Random Tree* e *Random Forest* (100, 200, 300, 400 e 500 árvores), expressos como acurácia, kappa, sensibilidade, especificidade, área ROC e tempos de treino e teste, para a etapa de validação cruzada com 10 *folds*, rodada 3 vezes, no total de 30 experimentos. A Figura 2 mostra os resultados em boxplots, expressos como área ROC.

A Tabela 2 ilustra os resultados dos métodos de classificação Perceptron Multicamadas, com 20, 50 e 100 neurônios na camada escondida do modelo de duas camadas, e com 20 e 20, 20 e 50, e 20 e 100 neurônios nas camadas escondidas no modelo de três camadas, expressos como acurácia, kappa, sensibilidade, especificidade, área ROC e tempos de treino e teste, para a validação cruzada com 10 *folds*, rodada 3 vezes, no total de 30 experimentos. Na Figura 3 têm-se os resultados da área ROC em boxplots.

Já a Tabela 3 mostra os resultados dos métodos de classificação Máquina de Vetor de Suporte (SVM) de kernel polinomial e RBF, expressos como acurácia, kappa, sensibilidade, especificidade, área ROC e tempos de treino e teste, para a validação cruzada com 10 *folds*, rodada 3 vezes, no total de 30 experimentos. Na Figura 4 podem ser vistos os resultados da área ROC em boxplots.

Após a validação cruzada, foi verificado que as *Random Forests* tiveram os melhores resultados de área ROC, mas que esses resultados foram equivalentes. Então foi escolhido o método *Random Forest* de 100 árvores para teste, dado que



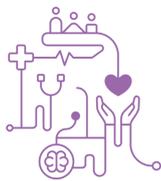
essa foi a configuração com menor número de árvores e, portanto, menos custosa. Os resultados obtidos foram acurácia de 93,40% (0,85%), kappa de 0,912 (0,011), sensibilidade de 0,934 (0,009), especificidade de 0,978 (0,003), e área ROC de 0,990 (0,002).

A Tabela 4 apresenta a Matriz de confusão dos resultados do teste com o melhor modelo (*Random Forest* de 100 árvores) com acurácia de 92,17%, kappa de 0,8828, sensibilidade de 0,922, especificidade de 0,956, e área ROC de 0,982. Dos registros de dengue clássica, 698 foram corretamente classificados, enquanto apenas um foi confundido com descartado e um com chikungunya; todos os resultados de dengue com complicações foram corretamente classificados; dos registros descartados, 904 foram corretamente classificados, enquanto 113 foram confundidos com chikungunya; por fim, dos casos de chikungunya, 1602 foram corretamente identificados, enquanto 178 foram incorretamente descartados e apenas três foram confundidos com dengue com complicações.

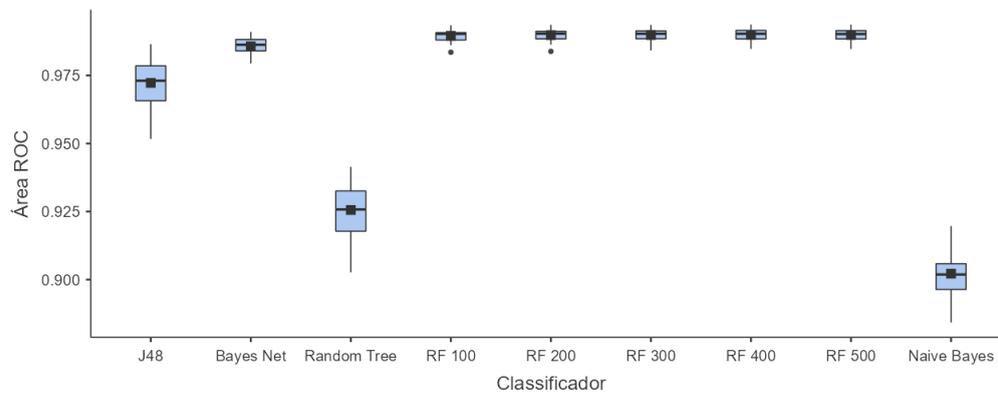
**Tabela 1** – Resultados dos métodos de classificação *Naive Bayes*, *Bayes Net*, Árvore de decisão J48, *Random Tree* e *Random Forest* (100, 200, 300, 400 e 500 árvores), expressos como acurácia, kappa, sensibilidade, especificidade, área ROC e tempos de treino e teste

Estatística Descritiva

	Classificador	Acurácia (%)	Kappa	Sens.	Esp.	Área ROC	Tempo de treino (ms)	Tempo de teste (ms)
Média	J48	92.4	0.899	0.924	0.975	0.972	196	1.56
	Bayes Net	91.9	0.892	0.919	0.973	0.986	79.2	3.13
	Random Tree	88.8	0.851	0.888	0.963	0.926	48.4	1.04
	RF 100	93.4	0.912	0.934	0.978	0.990	1776	41.1
	RF 200	93.4	0.911	0.934	0.978	0.990	3693	93.8
	RF 300	93.5	0.913	0.935	0.978	0.990	5554	136
	RF 400	93.5	0.913	0.935	0.978	0.990	7362	210
	RF 500	93.4	0.912	0.934	0.978	0.990	9585	250
	Naive Bayes	62.8	0.504	0.628	0.876	0.902	21.9	12.0
Desvio-padrão	J48	1.19	0.0158	0.0119	0.00397	0.00841	58.0	4.77
	Bayes Net	1.21	0.0161	0.0121	0.00401	0.00339	57.9	6.36
	Random Tree	1.53	0.0204	0.0153	0.00508	0.0102	25.7	3.96
	RF 100	0.852	0.0114	0.00852	0.00284	0.00224	109	12.0
	RF 200	0.885	0.0118	0.00885	0.00295	0.00231	161	26.0
	RF 300	0.923	0.0123	0.00923	0.00307	0.00223	247	40.4
	RF 400	0.889	0.0119	0.00889	0.00296	0.00219	347	67.1
	RF 500	0.916	0.0122	0.00916	0.00305	0.00216	318	58.0
	Naive Bayes	1.91	0.0256	0.0191	0.00642	0.00809	15.7	10.6



**Figura 2** – Resultados dos métodos de classificação *Naive Bayes*, *Bayes Net*, *Árvore de decisão J48*, *Random Tree* e *Random Forest* (100, 200, 300, 400 e 500 árvores) em *boxplots*, expressos como área ROC

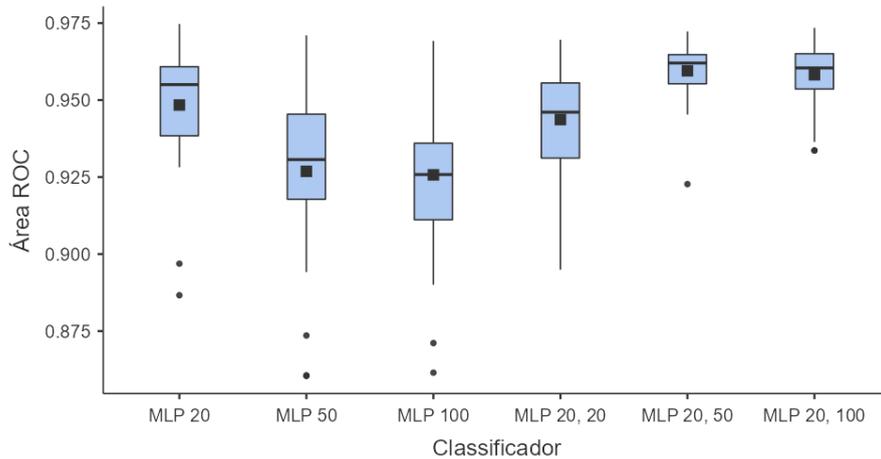


**Tabela 2** – Resultados dos métodos de classificação *Perceptron* Multicamadas, com 20, 50 e 100 neurônios na camada escondida do modelo de duas camadas, e com 20 e 20, 20 e 50, e 20 e 100 neurônios nas camadas escondidas no modelo de três camadas, expressos como acurácia, kappa, sensibilidade, especificidade, área ROC e tempos de treino e teste

Estatística Descritiva

	Classificador	Acurácia (%)	Kappa	Sens.	Esp.	Área ROC	Tempo de treino (ms)	Tempo de teste (ms)
Média	MLP 20	85.1	0.802	0.851	0.950	0.948	46080	8.85
	MLP 50	83.1	0.775	0.831	0.944	0.927	118914	15.6
	MLP 100	83.3	0.777	0.833	0.944	0.926	234772	26.0
	MLP 20, 20	85.8	0.810	0.858	0.953	0.944	79614	7.81
	MLP 20, 50	85.8	0.810	0.858	0.953	0.960	135434	15.6
	MLP 20, 100	85.2	0.803	0.852	0.951	0.958	237532	22.4
Desvio-padrão	MLP 20	2.41	0.0321	0.0241	0.00801	0.0199	2254	7.88
	MLP 50	1.96	0.0262	0.0196	0.00654	0.0275	802	0.00
	MLP 100	2.47	0.0329	0.0247	0.00822	0.0248	1081	7.49
	MLP 20, 20	1.63	0.0217	0.0163	0.00541	0.0177	1184	7.95
	MLP 20, 50	2.35	0.0314	0.0235	0.00787	0.00950	862	0.00
	MLP 20, 100	2.68	0.0357	0.0268	0.00892	0.0104	1082	7.88

**Figura 3** – Resultados dos métodos de classificação *Perceptron* Multicamadas, com 20, 50 e 100 neurônios na camada escondida do modelo de duas camadas, e com 20 e 20, 20 e 50, e 20 e 100 neurônios nas camadas escondidas no modelo de três camadas, em *boxplots*, expressos como área ROC



**Tabela 3 – Resultados dos métodos de classificação Máquina de Vetor de Suporte (SVM) de kernel polinomial e RBF, expressos como acurácia, kappa, sensibilidade, especificidade, área ROC e tempos de treino e teste**

Estatística Descritiva

	Classificador SVM	Acurácia (%)	Kappa	Sens.	Esp.	Área ROC	Tempo de treino (ms)	Tempo de teste (ms)
Média	Poly1 a	56.5	0.420	0.565	0.855	0.732	17989	9.38
	Poly2 a	83.0	0.773	0.830	0.943	0.942	2.46e+6	29542
	Poly3 a	91.5	0.886	0.915	0.972	0.971	6.00e+6	13254
	RBF1 a	29.3	0.0574	0.293	0.765	0.631	3.37e+7	100734
	RBF2 a	60.7	0.477	0.607	0.869	0.866	4.43e+6	63065
	RBF3 a	66.5	0.553	0.665	0.888	0.886	2.69e+6	52434
	Poly1 b	79.3	0.723	0.793	0.931	0.880	114082	13.0
	Poly2 b	89.3	0.858	0.893	0.964	0.964	2.85e+6	17040
	Poly3 b	93.8	0.918	0.938	0.979	0.980	3.00e+7	9840
	RBF1 b	51.0	0.347	0.510	0.837	0.715	8.39e+6	89304
	RBF2 b	83.5	0.779	0.835	0.945	0.944	1.49e+6	31867
	RBF3 b	87.4	0.831	0.874	0.958	0.957	1.06e+6	25074
	Poly1 c	79.8	0.730	0.798	0.933	0.895	718827	7.81
	Poly2 c	91.4	0.886	0.914	0.971	0.971	7.99e+6	11857
	Poly3 c	94.2	0.922	0.942	0.981	0.981	1.06e+8	3733
RBF1 c	81.7	0.756	0.817	0.939	0.912	4.04e+6	68184	
RBF2 c	89.9	0.865	0.899	0.966	0.966	819373	15374	
RBF3 c	90.6	0.875	0.906	0.969	0.968	713649	14563	
Desvio-padrão	Poly1 a	0.744	0.00992	0.00744	0.00248	0.00592	1605	12.7
	Poly2 a	0.638	0.00851	0.00638	0.00213	0.00220	170850	998
	Poly3 a	0.471	0.00628	0.00471	0.00157	0.00159	543321	1139
	RBF1 a	5.81	0.0774	0.0581	0.0193	0.0350	3.78e+6	5113
	RBF2 a	0.737	0.00983	0.00737	0.00245	0.00257	248408	2712
	RBF3 a	0.642	0.00856	0.00642	0.00214	0.00215	157393	2398
	Poly1 b	0.668	0.00891	0.00668	0.00223	0.00515	9543	5.92
	Poly2 b	0.547	0.00729	0.00547	0.00183	0.00187	138126	710
	Poly3 b	0.408	0.00545	0.00408	0.00136	0.00126	3.85e+6	1208
	RBF1 b	0.754	0.0101	0.00754	0.00252	0.00635	996480	6097
RBF2 b	0.619	0.00825	0.00619	0.00207	0.00206	121091	1594	

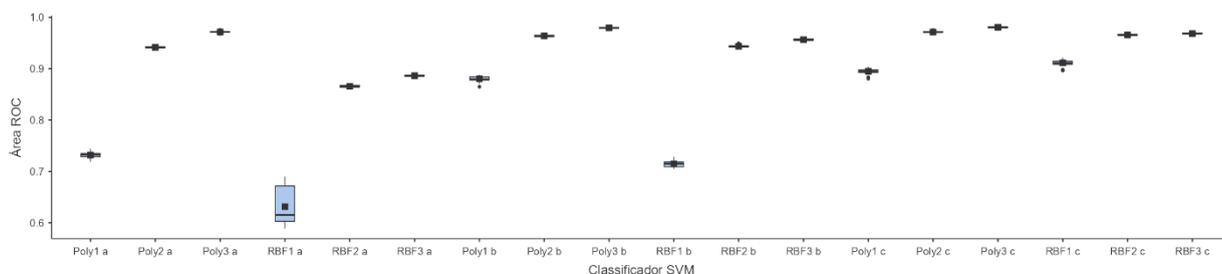


## Estatística Descritiva

Classificador SVM	Acurácia (%)	Kappa	Sens.	Esp.	Área ROC	Tempo de treino (ms)	Tempo de teste (ms)
RBF3 b	0.547	0.00730	0.00547	0.00183	0.00197	64003	695
Poly1 c	0.630	0.00840	0.00630	0.00210	0.00511	73474	7.95
Poly2 c	0.498	0.00663	0.00498	0.00166	0.00164	642400	724
Poly3 c	0.430	0.00573	0.00430	0.00143	0.00140	3.24e+7	1036
RBF1 c	0.719	0.00958	0.00719	0.00240	0.00563	224994	2890
RBF2 c	0.511	0.00682	0.00511	0.00171	0.00169	69362	547
RBF3 c	0.478	0.00638	0.00478	0.00160	0.00166	47571	396

Nota: Poly1: Kernel linear; Poly2: Kernel polinomial de grau 2; RBF1: Kernel RBF,  $G = 0.01$ ; RBF2: Kernel RBF,  $G = 0.25$ ; RBF3: Kernel RBF,  $G = 0.5$ ; a:  $C = 0.01$ ; b:  $C = 0.1$ ; c:  $C = 1.0$ .

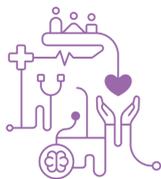
**Figura 4** – Resultados dos métodos de classificação Máquina de Vetor de Suporte (SVM) de kernel polinomial e RBF, em *boxplots*, expressos como área ROC



**Tabela 4** – Matriz de confusão dos resultados do teste com o melhor modelo: Random Forest de 100 árvores, com acurácia de 92,17%, kappa de 0,8828, sensibilidade de 0,922, especificidade de 0,956, e área ROC de 0,982. As linhas representam a verdade *a priori*, enquanto as colunas são a decisão da máquina.

	Dengue clássica	Dengue com complicações	Descartados	Chikungunya
Dengue clássica	698	0	1	1
Dengue com complicações	0	278	0	0
Descartados	0	0	904	113
Chikungunya	0	3	178	1602

A análise dos resultados dos métodos de classificação mostrou que as *Random Forests* apresentaram os melhores desempenhos em termos de área sob a curva ROC, indicando uma boa capacidade de discriminação entre casos de dengue e outras condições. Esses resultados foram consistentes em várias configurações de árvores, mostrando a robustez desse método.



Ao escolher o método *Random Forest* com 100 árvores para testes adicionais, consideramos não apenas a eficácia do modelo, mas também sua eficiência computacional. A configuração com 100 árvores foi selecionada devido ao seu desempenho comparável aos demais experimentos, mas com menor custo computacional em termos de tempo de treinamento e teste.

Os resultados do teste final com o método *Random Forest* de 100 árvores mostraram uma acurácia de 92,17%, kappa de 0,8828, sensibilidade de 0,922, especificidade de 0,956, indicando uma boa capacidade de predição do diagnóstico da dengue e da febre chikungunya. Além disso, a área sob a curva ROC de 0,982 demonstra um alto poder discriminatório do modelo.

## Conclusão

Os resultados gerais indicaram que o modelo *Random Forest* com 100 árvores foi o mais eficaz. A matriz de confusão do teste final demonstrou que o modelo foi capaz de classificar corretamente a grande maioria dos casos de chikungunya, dengue clássica e dengue com complicações, minimizando os erros de classificação entre essas condições. Esses resultados mostram que o modelo desenvolvido pode ser uma ferramenta valiosa para apoiar o diagnóstico clínico, melhorando a eficiência e a precisão na identificação de casos de chikungunya.

Para trabalhos futuros, é essencial explorar alguns pontos de melhoria. Primeiramente, seria benéfico aumentar a diversidade e a quantidade de dados de prontuários para incluir um conjunto mais abrangente de casos de dengue e outras arboviroses. Além disso, a incorporação de novas técnicas de pré-processamento de dados e seleção de características pode potencialmente melhorar ainda mais o desempenho dos modelos de classificação. Também é recomendável realizar testes em ambientes clínicos reais para validar a aplicabilidade prática do modelo e ajustar a sua integração nos sistemas de saúde existentes. Finalmente, a investigação de abordagens de aprendizado de máquina mais avançadas, como redes neurais profundas e aprendizado por reforço, pode proporcionar informações adicionais e aprimorar a capacidade diagnóstica do sistema.



## Agradecimentos

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado de Pernambuco (FACEPE), ao Departamento de Ciência e Tecnologia da Secretaria de Ciência, Tecnologia, Inovação e Insumos Estratégicos em Saúde do Ministério da Saúde (Decit/SCTIE/MS), à Financiadora de Estudos e Projetos (FINEP ref. 2170/22) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo financiamento da pesquisa.

## Referências

1. Lopes N, Nozawa C, Linhares REC. Características gerais e epidemiologia dos arbovírus emergentes no Brasil. *Rev Pan-Amaz Saude*. 2014;5(3):10-10.
2. Donalisio MR, Freitas ARR, Zuben APBV. Arboviroses emergentes no Brasil: desafios para a clínica e implicações para a saúde pública. *Rev Saude Publica*. 2017;51:30.
3. Viana LRDC, Pimenta CJL, Araújo EMNFD, Teófilo TJL, Costa TFD, Costa KNDFM. Arboviroses reemergentes: perfil clínico-epidemiológico de idosos hospitalizados. *Rev Esc Enferm USP*. 2018;52:e03403.
4. Young PR. Arboviruses: a family on the move. *Dengue and Zika: control and antiviral treatment strategies*. 2018:1-10.
5. Thisyakorn U, Thisyakorn C. Diseases caused by arboviruses-dengue haemorrhagic fever and Japanese B encephalitis. *Med J Aust*. 1994;160(1):22-26.
6. Bautista-Reyes E, Núñez-Avellaneda D, Alonso-Palomares L A, Salazar MI. Chikungunya: Molecular aspects, clinical outcomes and pathogenesis. *Rev Investig Clin*. 2018;69(6):299-307.
7. Ahmed M, Mahboob N, Manun KT, Iqbal H. Dilemma in laboratory diagnosis of Dengue, Chikungunya and Zika viruses. *Bangladesh J*. 2019;30:83-92.
8. Cunha RVD, Trinta KS. Chikungunya virus: clinical aspects and treatment-A Review. *Mem Inst Oswaldo Cruz*. 2017;112(8):523-531.
9. Santos MK, Webber CG. Comparing Deep e Transfer Learning na Classificação de Imagens da Membrana Timpânica. *J Health Inform*. 2022;14.
10. Santus E, Marino N, Cirillo D, Chersoni E, Montagud A, Santucciono Chadha A, et al. Artificial intelligence-aided precision medicine for COVID-19: strategic areas of research and development. *J Med Internet Res*. 2021;23(3):e22453.
11. Malik YS, Sircar S, Bhat S, Ansari MI, Pande T, Kumar P, et al. How artificial intelligence may help the Covid-19 pandemic: Pitfalls and lessons for the future. *Rev Med Virol*. 2021;31(5):1-11.



# CBIS'24

XX Congresso Brasileiro de Informática em Saúde  
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

12. da Silva Neto SR, Tabosa Oliveira T, Teixeira IV, Aguiar de Oliveira SB, Souza Sampaio V, Lynn T, et al. Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: A systematic review. PLoS Negl Trop Dis. 2022;16(1):e0010061.