Classification of SARS-CoV-2 Variants from Amino Acid Substitution Embeddings Using Supervised Learning

Alessandro S. Silva[1], **Karolayne S. Azevedo**[1] **e Marcelo A. C. Fernandes**[1,2,3]

[1]InovAI Lab, nPITI/IMD, UFRN, 59.078-900, Natal, RN, Brasil

[2]Centro Multiusuário de Bioinformática (BioME) - IMD, UFRN, Natal, Brasil

[3]Departamento de Engenharia da Computação e Automação (DCA), UFRN, Natal, Brasil

asdseng@gmail.com, karolayneazevsantos@gmail.com, mfernandes@dca.ufrn.br

Abstract. *This work proposes a supervised approach for classifying SARS-CoV-2 variants of concern (VOCs) using vector embeddings derived from amino acid substitutions. The pre-trained model* `all-MiniLM-L6-v2` *was employed to generate high-dimensional vectors encoding mutations without requiring sequence alignment. These embeddings served as input to classifiers such as SVM, Random Forest, XGBoost, and k-NN, evaluated through cross-validation and on an external test set of nearly 288,000 samples. XGBoost achieved the best results, with 99.83% accuracy, 99.83% F1 macro, and a logLoss of 0.0068, maintaining high performance on unseen data. Tree-based models outperformed others, particularly in handling the Gamma variant. The proposed approach proves to be robust, accurate, and scalable for application in automated genomic surveillance systems, providing a complementary perspective to clinical and laboratory analyses, and outperforming traditional and hybrid methods from recent literature.*

Resumo. *Este trabalho propõe uma abordagem supervisionada para a classificação de variantes do SARS-CoV-2 (VOCs) a partir de embeddings vetoriais derivados de substituições de aminoácidos. Utilizando o modelo pré-treinado* `all-MiniLM-L6-v2`*, foram gerados vetores de alta dimensionalidade que codificam mutações sem necessidade de alinhamento genômico. Esses embeddings alimentaram classificadores como SVM, Random Forest, XGBoost e k-NN, avaliados por validação cruzada e em um teste externo com quase 288 mil amostras. O XGBoost obteve os melhores resultados, com acurácia de 99,83%, F1 macro de 99,83% e logLoss de 0,0068, mantendo desempenho elevado mesmo em dados não vistos. Os resultados evidenciam que modelos baseados em árvores superam alternativas como SVM, especialmente na variante Gamma. A proposta se mostra robusta, precisa e escalável para aplicação em sistemas automatizados de vigilância genômica, fornecendo uma segunda perspectiva de análise complementar à abordagem clínica e laboratorial, e superando métodos tradicionais e híbridas da literatura recente.*

1. Introdução

Desde o surgimento do SARS-CoV-2 no final de 2019, o acompanhamento de suas variantes tem sido um dos principais desafios da vigilância epidemiológica global.

As mutações acumuladas no genoma viral, especialmente na proteína Spike, impactam diretamente na transmissibilidade, escape imunológico e efetividade de vacinas, tornando essencial a detecção rápida e precisa de variantes de preocupação (VOCs). Nesse contexto, abordagens computacionais capazes de classificar automaticamente essas variantes com base em seus perfis mutacionais representam ferramentas valiosas para subsidiar ações de saúde pública e estratégias de contenção. Neste trabalho, propomos o uso de aprendizado supervisionado aplicado a embeddings gerados a partir de substituições de aminoácidos, sem necessidade de alinhamento genômico, com o objetivo de realizar a classificação de VOCs de forma precisa, escalável e interpretável [Ran et al. 2022].

Diversas abordagens têm sido propostas para a classificação de variantes do SARS-CoV-2, com diferentes fontes de dados e metodologias. O trabalho de [Qin et al. 2024] propõe uma abordagem baseada em espectroscopia Raman com realce por superfície (SERS), combinada com aprendizado de máquina supervisionado, especificamente regressão logística, alcançando 100% de acurácia na distinção entre variantes como Wuhan, Beta, Delta e Omicron em amostras de saliva e swabs nasais. Embora eficaz e promissora para diagnósticos rápidos e não invasivos, essa abordagem depende de infraestrutura laboratorial especializada e de espectrômetros portáteis. De forma semelhante, o estudo de [Fatima and Ahmad 2024] propõe um sistema de detecção de variantes e predição de mortalidade baseado na análise de sintomas clínicos, utilizando algoritmos como Random Forest, XGBoost, SVM e k-NN. A proposta se destaca por mapear sintomas a variantes conhecidas e prever desfechos clínicos com alta acurácia, mas está limitada à disponibilidade e qualidade das informações sintomáticas. Já o trabalho de [Promja et al. 2023] combina qPCR de alta resolução de fusão (HRM) com aprendizado de máquina, alcançando 95,2% de sensibilidade em 167 amostras clínicas, destacando-se como alternativa econômica ao sequenciamento genético, embora ainda dependa de experimentação laboratorial.

No trabalho de [Beduk et al. 2022], os autores desenvolveram uma plataforma portátil de diagnóstico baseada em sensores eletroquímicos funcionalizados com ACE2 e integrados a nanopartículas de ouro e grafeno laser-escrito (LSG), acoplada a um dispositivo point-of-care (PoC) com aprendizado de máquina embarcado. Essa proposta permite a identificação rápida de variantes (Alpha, Beta e Delta) diretamente a partir de swabs nasofaríngeos, com acurácia de 99,37% em menos de 1 minuto, utilizando TinyML em um potenciostato portátil. Apesar da inovação em acessibilidade e resposta em tempo real, o método ainda depende de sensores físicos e experimentação laboratorial. De forma distinta, o estudo de [Ali et al. 2021] propõe uma abordagem computacional baseada em k-mers para representar sequências da proteína spike do SARS-CoV-2, com posterior aplicação de PCA e classificação supervisionada (SVM, Random Forest), atingindo alta acurácia mesmo com 1% dos dados para treinamento. Embora compartilhe o uso da proteína spike com nossa proposta, o trabalho utiliza vetores de frequência de k-mers, enquanto adotamos embeddings contextuais gerados por modelos de linguagem, o que reduz a necessidade de engenharia manual de features e facilita a adaptação a novas variantes. Complementarmente, o estudo de [Chourasia et al. 2023] introduz um modelo baseado em aprendizado federado (Federated Learning – FL) para classificação de variantes, preservando a privacidade dos dados entre instituições. Utilizando modelos locais (Random Forest, XGBoost, Regressão Logística) e um modelo global de rede neural, o sistema alcança acurácia superior a 93%.

Trabalhos como [Singh et al. 2022], propõem uma analogia conceitual, tratando mutações como "palavras" e genomas como "documentos" em um grande "corpus" de evolução viral. Utilizando modelos como Word2Vec e *Dynamic Topic Modeling* (DTM), esses autores conseguiram rastrear assinaturas mutacionais ao longo do tempo e associá-las a linhagens específicas e regiões geográficas distintas. Aspecto também abordado por [Sokhansanj et al. 2022], que propuseram um modelo baseado em atenção interpretável aplicado à sequência da proteína Spike para prever desfechos clínicos com base no padrão. De forma semelhante, os trabalhos apresentados em [Câmara et al. 2022, Azevedo et al. 2024, de Souza et al. 2023, Coutinho et al. 2023] exploram diferentes formas de representações e classificação de sequências usando CNNs, alcançando acurácia superior a 98% na distinção entre SARS-CoV-2 e outros vírus da família *Coronaviridae*, reforçando a eficácia de abordagens baseadas em aprendizado profundo para a classificação genômica.

Assim, este artigo tem como objetivo propor e avaliar uma abordagem supervisionada para a classificação de variantes do SARS-CoV-2 (VOCs) baseada em representações vetoriais (embeddings) derivadas de substituições de aminoácidos. Utilizando o modelo pré-treinado `all-MiniLM-L6-v2`, essas substituições são transformadas em vetores de alta dimensionalidade que preservam aspectos contextuais e estruturais das mutações, eliminando a necessidade de alinhamento genômico, coleta clínica ou experimentação laboratorial. Esses embeddings foram utilizados como entrada para algoritmos clássicos de aprendizado supervisionado, incluindo Máquina de Vetores de Suporte com kernel radial (SVM), Floresta Aleatória (Random Forest), *eXtreme Gradient Boosting* (XGBoost) e k-Vizinhos Mais Próximos (k-NN), treinados com validação cruzada estratificada e avaliados em um conjunto de teste externo composto por dados não vistos durante o treinamento. Para garantir a robustez dos resultados, foram adotadas estratégias de balanceamento de classes e métricas estatísticas consistentes.

2. Metodologia

2.1. Base de Dados e Pré-processamento

As amostras genômicas utilizadas neste estudo foram obtidas da plataforma GISAID [Khare et al. 2021], com extração realizada em 28 de março de 2024. O conjunto original possuía aproximadamente 16,6 milhões de sequências. Foram aplicados filtros de qualidade para reter apenas as sequências completas, com alta cobertura e classificadas como Variantes de Preocupação (VOCs), conforme as diretrizes da Organização Mundial da Saúde. Após essa etapa, restaram 936.638 amostras com metadados contendo a coluna `AA Substitutions`, que representa mutações de aminoácidos no formato `Gene:Substituição` (por exemplo, `Spike:D614G`) [Ran et al. 2022].

Essa notação padronizada preserva a unidade semântica de cada alteração e destaca apenas os eventos mutacionais com potencial impacto funcional, evitando a necessidade de alinhar sequências nucleotídicas completas. Cada amostra foi, então, tratada como um "documento" composto por um conjunto de substituições de aminoácidos, o que favorece a aplicação de técnicas de Processamento de Linguagem Natural para capturar relações contextuais entre mutações. Essas substituições foram processadas com o modelo `all-MiniLM-L6-v2` da biblioteca `Sentence Transformers`, o qual aplica mecanismos de atenção para capturar relações contextuais não lineares entre as

substituições. O resultado foi a geração de vetores de embeddings de 384 dimensões para cada perfil mutacional.

A etapa seguinte consistiu na remoção de duplicatas com base na coluna `AA Substitutions` e na representação vetorial. Para a tarefa de classificação supervisionada, foi realizado um balanceamento por subamostragem com base na menor classe, assegurando distribuição uniforme entre as variantes. As amostras não utilizadas no processo de balanceamento foram reservadas como conjunto de teste externo, de modo a permitir avaliação em dados inéditos não vistos durante a validação cruzada.

Durante o pré-processamento dos dados vetoriais, foi identificado que duas dimensões específicas do espaço de embedding, denominadas aqui como dimensões 224 e 320, apresentavam variância nula, isto é, mantinham exatamente o mesmo valor em todas as amostras analisadas. Tais dimensões invariantes foram removidas antes do treinamento dos modelos, uma vez que não oferecem qualquer contribuição informativa para a separação entre as classes. Além disso, a permanência de atributos invariantes pode prejudicar o desempenho de algoritmos sensíveis à variabilidade dos dados, como SVM e k-NN, ao introduzir redundância ou ruído numérico desnecessário no processo de aprendizado.

2.2. Modelos Supervisionados e Validação Cruzada

Foram avaliados quatro modelos clássicos de aprendizado supervisionado: Máquina de Vetores de Suporte com kernel radial (SVM), Floresta Aleatória (Random Forest), eXtreme Gradient Boosting (XGBoost) e k-Vizinhos Mais Próximos (k-NN). Todos os modelos foram treinados utilizando validação cruzada estratificada do tipo k-fold, com $k = 5$, de modo a assegurar uma avaliação robusta e imparcial do desempenho preditivo.

Os modelos receberam como entrada exclusivamente os vetores de embeddings gerados a partir das substituições de aminoácidos. A tarefa supervisionada consistiu na classificação de cada amostra genômica em uma das cinco variantes de preocupação consideradas no estudo (Alpha, Beta, Delta, Gamma e Omicron). Durante a validação cruzada, foram armazenadas as predições realizadas em cada fold, permitindo a análise consolidada dos resultados sem necessidade de avaliação sobre o conjunto de treino completo.

Durante a validação cruzada, além da avaliação do desempenho preditivo em cada partição, foi realizado o ajuste de hiperparâmetros internos específicos para cada modelo. No caso do SVM, foram testados diferentes valores do parâmetro de penalização C; na Random Forest, diferentes valores de $mtry$; no XGBoost, combinações de profundidade de árvore, taxa de aprendizado, subamostragem e número de iterações; e no k-NN, diferentes valores de k. As melhores configurações foram selecionadas com base na minimização do logLoss, garantindo não apenas acurácia, mas também maior confiabilidade nas probabilidades estimadas. As predições geradas em cada fold foram armazenadas separadamente, permitindo a análise consolidada dos resultados e evitando viés decorrente de avaliação sobre o próprio conjunto de treinamento.

3. Resultados e Discussão

Antes da avaliação do desempenho dos classificadores, é importante apresentar a composição final dos conjuntos utilizados nas análises. O conjunto de treinamento balanceado resultou em 29.325 amostras distribuídas uniformemente entre as VOCs Alpha, Beta, Delta, Gamma e Omicron. Já o conjunto de teste externo foi composto por 287.994 amostras, distribuídas da seguinte forma: Alpha (45.099), Delta (223.466), Gamma (6.937) e Omicron (12.492).

3.1. Desempenho do Classificador SVM

O classificador SVM com kernel radial (RBF) foi treinado com embeddings de 382 variáveis, resultantes da remoção de duas dimensões com variância nula. O treinamento foi conduzido com validação cruzada estratificada do tipo 5-fold, permitindo uma avaliação robusta do desempenho preditivo. Foram testados diferentes valores do parâmetro de penalização C (0,25, 0,50 e 1,00), mantendo-se fixo o parâmetro $\sigma = 0{,}00213$. Os resultados obtidos para cada configuração estão apresentados na Tabela 1.

Tabela 1. Resultados da validação cruzada para o classificador SVM com kernel RBF.

C	logLoss	AUC	prAUC	Acurácia	Kappa	F1 macro	Sens. média
0,25	0,4324	0,9982	0,9933	0,9331	0,9163	0,9296	0,9331
0,50	0,4384	0,9983	0,9936	0,9369	0,9211	0,9339	0,9369
1,00	0,4664	0,9980	0,9926	0,9380	0,9225	0,9351	0,9380

C	Espec. média	Prec. média	NPV média	Acc. balanceada
0,25	0,9833	0,9423	0,9844	0,9582
0,50	0,9842	0,9452	0,9853	0,9606
1,00	0,9845	0,9461	0,9855	0,9612

O melhor desempenho do classificador SVM foi obtido com $C = 0{,}25$, conforme o critério de menor logLoss. Nessa configuração, o modelo alcançou uma acurácia média de 93,31%, F1 macro de 92,96%, AUC de 0,9982 e prAUC de 0,9933. A estatística de Kappa foi de 0,9163, refletindo concordância elevada entre predições e rótulos reais. As métricas de sensibilidade e precisão médias foram de 93,31% e 94,23%, respectivamente, e a acurácia balanceada atingiu 95,82%, evidenciando um desempenho equilibrado na classificação das diferentes variantes. Esses resultados indicam um desempenho robusto e equilibrado na classificação das cinco variantes (Alpha, Beta, Delta, Gamma e Omicron), com alta separabilidade dos embeddings gerados a partir de substituições de aminoácidos. A alta AUC e prAUC refletem a capacidade discriminativa do modelo, mesmo diante de classes com padrões mutacionais semelhantes.

Embora os valores de $C = 0{,}5$ e $C = 1{,}0$ tenham apresentado ligeiras melhorias em métricas como acurácia e F1, o critério de menor logLoss levou à seleção de $C = 0{,}25$, favorecendo maior confiabilidade probabilística nas predições. A combinação de pré-processamento (centralização e normalização), balanceamento estratificado e uso de embeddings contextualizados permitiu que o SVM capturasse padrões não lineares entre os perfis mutacionais das VOCs. Esses achados reforçam o potencial da abordagem para tarefas de vigilância genômica baseada em IA supervisionada.

A Figura 1 apresenta a matriz de confusão percentual do modelo SVM obtida durante a validação cruzada. Observa-se que a maioria das classes apresenta alta taxa de acerto nas diagonais principais, com destaque para as variantes Beta e Alpha, que atingiram 99,9% de acertos. A variante Gamma, por outro lado, apresentou o maior índice de confusão relativa, com 22,7% de suas amostras sendo equivocadamente classificadas como Omicron. Essa confusão entre Gamma e Omicron pode ser atribuída à sobreposição de padrões mutacionais entre essas variantes. De modo geral, a matriz confirma a capacidade do SVM em distinguir corretamente a maioria das variantes, embora com maior dificuldade em contextos de alta similaridade genômica.

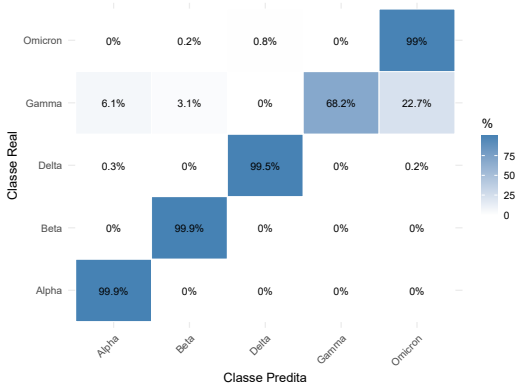

Figura 1. Matriz de confusão percentual para o classificador SVM com kernel radial, obtida durante a validação cruzada (5-fold).

3.2. Desempenho do Classificador Random Forest

O classificador Random Forest foi treinado com os embeddings derivados das substituições de aminoácidos, resultando em 382 preditores. A avaliação foi conduzida com validação cruzada estratificada do tipo 5-fold. Três valores do hiperparâmetro `mtry` foram testados: 2, 192 e 382, correspondendo à quantidade de variáveis consideradas em cada divisão das árvores de decisão.

O melhor desempenho foi obtido com `mtry = 192`, conforme o critério de menor logLoss. Nessa configuração, o modelo alcançou acurácia de 99,71%, F1 macro de 99,71%, AUC de 0,9999 e logLoss de 0,0290. As métricas de Kappa, precisão e recall também apresentaram valores elevados, todos em torno de 99,7%, com acurácia balanceada de 99,82%. Esses resultados estão detalhados na Tabela 2.

Tabela 2. Resultados da validação cruzada para o classificador Random Forest.

mtry	logLoss	AUC	prAUC	Acurácia	Kappa	F1 macro	Sens. média
2	0,0673	0,9999	0,6560	0,9964	0,9955	0,9964	0,9964
192	0,0290	0,9999	0,2305	0,9971	0,9964	0,9971	0,9971
382	0,0319	0,9999	0,1668	0,9946	0,9932	0,9946	0,9946

mtry	Espec. média	Prec. média	NPV média	Acc. balanceada
2	0,9991	0,9964	0,9991	0,9977
192	0,9993	0,9971	0,9993	0,9982
382	0,9986	0,9946	0,9986	0,9966

Além da acurácia elevada, o modelo com `mtry = 192` apresentou desempenho consistente em métricas complementares. O valor de logLoss, igual a 0,0290, indica elevada confiabilidade nas probabilidades preditas, penalizando menos os casos em que o modelo comete erros incertos. A métrica AUC atingiu o valor de 0,9999, evidenciando excelente separabilidade entre as classes, enquanto a prAUC, igual a 0,2305, refletiu a influência do balanceamento das classes sobre a distribuição de precisão e recall. A estatística de Kappa (0,9964) revelou forte concordância entre as predições do modelo e os rótulos reais. A métrica F1 macro foi de 99,71%, indicando equilíbrio entre precisão e sensibilidade, que também atingiram 99,71% e 99,71%, respectivamente. Esses resultados indicam que o modelo manteve um desempenho robusto em todas as dimensões avaliadas, sendo capaz de classificar corretamente as variantes mesmo diante de padrões mutacionais complexos.

A Figura 2 mostra a matriz de confusão percentual obtida para o modelo Random Forest durante a validação cruzada. Observa-se um desempenho excepcionalmente alto, com as classes Delta, Gamma, Alpha e Beta sendo corretamente classificadas em 99,8% dos casos. A variante Omicron, embora ligeiramente mais distribuída, ainda apresentou 99,4% de acertos, com erros residuais mínimos espalhados entre as demais classes. Esses resultados reforçam a robustez da Random Forest na tarefa de classificação das VOCs, com margens de erro quase nulas e alta separabilidade entre os perfis mutacionais representados nos embeddings.

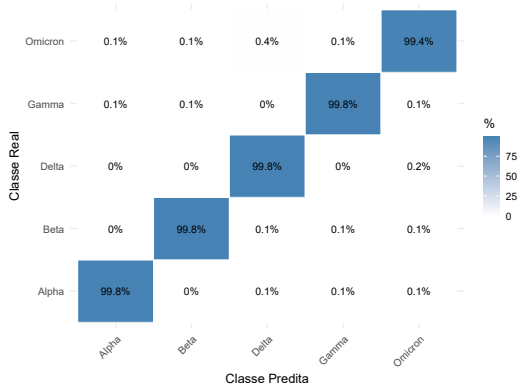

Figura 2. Matriz de confusão percentual para o classificador Random Forest, obtida durante a validação cruzada (5-fold).

3.3. Desempenho do Classificador XGBoost

O classificador XGBoost foi avaliado por meio de validação cruzada estratificada em 5 folds, utilizando embeddings de 382 preditores. Foram testadas diferentes combinações de hiperparâmetros, variando-se o número de iterações (`nrounds`), a profundidade máxima das árvores (`max_depth`) e fração de amostras por árvore (`subsample`. Os demais parâmetros, como taxa de aprendizado (`eta`) e proporção de colunas por árvores (`colsample_bytree`) foram mantidos fixos em seus valores padrão de implementação, 0,3 e 0,6, respectivamente.

A melhor configuração encontrada foi composta por `nrounds = 150`, `max_depth = 2` e `subsample = 0,5`, cujos resultados detalhados encontram-se na Tabela 3.

Tabela 3. Melhores configurações do XGBoost e suas métricas de desempenho.

nrounds	max_depth	subsample	logLoss	Acurácia	F1 macro
150	2	0,5	0,0068	0,9983	0,9983
100	2	0,5	0,0069	0,9982	0,9982
150	1	0,5	0,0069	0,9982	0,9982

nrounds	max_depth	subsample	AUC	prAUC
150	2	0,5	0,99998	0,9883
100	2	0,5	0,99998	0,9963
150	1	0,5	0,99998	0,9971

Com essa parametrização, o modelo obteve logLoss de apenas 0,0068, indicando altíssima confiança nas probabilidades preditas. A acurácia e o F1 macro atingiram 99,83%, refletindo equilíbrio entre precisão e sensibilidade. A AUC de 0,99998 confirma a excelente capacidade discriminativa do modelo, mesmo diante de classes com padrões mutacionais semelhantes. A métrica prAUC, de 0,9883, também reforça a robustez do modelo sob diferentes limiares de decisão. O índice de Kappa (0,9977) e a acurácia balanceada (99,88%) corroboram o desempenho consistente do XGBoost em todo o espectro de classes.

Comparado aos demais classificadores, o XGBoost apresentou desempenho ligeiramente superior, tanto em termos de probabilidade calibrada (menor logLoss), quanto em métricas clássicas de classificação. Sua estabilidade mesmo com profundidades reduzidas de árvore e taxas de subamostragem moderadas evidencia seu potencial para aplicações de classificação genômica em larga escala.

A Figura 3 apresenta a matriz de confusão percentual para o classificador XGBoost, obtida durante a validação cruzada. Observa-se que o modelo alcançou desempenho excepcionalmente elevado em todas as classes, com $99{,}9\%$ de acerto para Alpha, Beta e Gamma, $99{,}8\%$ de acerto em Delta e $99{,}6\%$ para Omicron. A classe Gamma apresentou 0,1% de confusão para Omicron, e Omicron teve 0,3% de suas amostras incorretamente classificadas como Delta. Esses desvios mínimos reforçam a capacidade discriminativa do XGBoost, mesmo em cenários com variantes geneticamente similares. A distribuição dos erros, visualmente concentrada fora da diagonal, é pequena e não compromete o desempenho geral, validando os resultados quantitativos já apresentados.

3.4. Desempenho do Classificador k-NN

O classificador k-Nearest Neighbors (k-NN) foi avaliado por meio de validação cruzada estratificada com 5 folds, utilizando embeddings de 382 preditores previamente centralizados e normalizados. Foram testados cinco valores do hiperparâmetro k: 5, 7, 9, 11 e 13.

A configuração que obteve o menor logLoss foi aquela com $k = 13$, indicando-a como a mais adequada segundo o critério probabilístico. Os resultados detalhados encontram-se na Tabela 4. Com essa configuração, o modelo apresentou logLoss de 0,0303, acurácia e F1 macro de 99,66%, AUC de 0,9996, prAUC de 0,0149 e Kappa de 0,9957. As métricas de precisão, sensibilidade e especificidade médias também foram elevadas, todas acima de 99,6%, com acurácia balanceada de 99,78%.

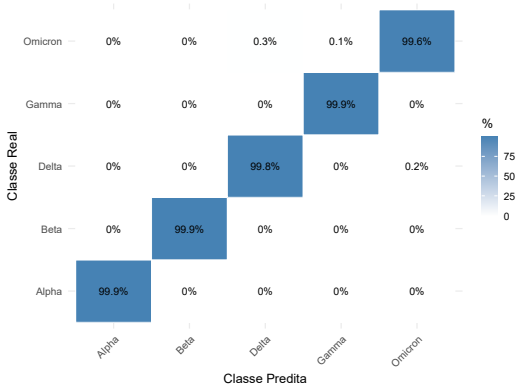

Figura 3. Matriz de confusão percentual para o classificador XGBoost, obtida durante a validação cruzada (5-fold).

Tabela 4. Resultados da validação cruzada para o classificador k-NN.

k	logLoss	AUC	prAUC	Acurácia	Kappa	F1 macro	Sens. média
5	0,0372	0,9994	0,0074	0,9971	0,9964	0,9971	0,9971
7	0,0322	0,9995	0,0098	0,9969	0,9961	0,9969	0,9969
9	0,0316	0,9995	0,0118	0,9968	0,9960	0,9968	0,9968
11	0,0310	0,9996	0,0134	0,9968	0,9960	0,9968	0,9968
13	0,0303	0,9996	0,0149	0,9966	0,9957	0,9966	0,9966

k	Espec. média	Prec. média	NPV média	Acc. balanceada
5	0,9993	0,9971	0,9993	0,9982
7	0,9992	0,9969	0,9992	0,9980
9	0,9992	0,9968	0,9992	0,9980
11	0,9992	0,9968	0,9992	0,9980
13	0,9991	0,9966	0,9991	0,9978

Embora o valor de logLoss tenha sido ligeiramente superior ao observado no XGBoost, o desempenho global do k-NN foi altamente competitivo. O modelo demonstrou estabilidade e elevada capacidade de generalização, mesmo operando em um espaço vetorial de alta dimensionalidade, como o dos embeddings de substituições de aminoácidos.

A Figura 4 apresenta a matriz de confusão percentual do classificador k-NN, obtida durante a validação cruzada. Observa-se que o modelo apresentou desempenho altamente consistente, com todas as classes sendo corretamente classificadas em mais de 99,8% dos casos. A variante Alpha atingiu 99,9% de acerto, com erros praticamente nulos. A classe Gamma foi corretamente classificada em 99,8% das amostras, com 0,2% de confusão para Omicron. Já a classe Omicron apresentou 98,9% de acerto, com 0,8% de suas amostras sendo classificadas como Delta. Apesar dessas pequenas taxas de confusão, o desempenho geral do k-NN manteve-se elevado, confirmando sua eficácia mesmo em cenários de alta similaridade genômica entre variantes.

3.5. Comparação entre os Modelos

A Tabela 5 apresenta um resumo comparativo do desempenho dos quatro classificadores avaliados neste estudo. Entre os modelos analisados, o XGBoost destacou-se em todas as métricas principais: obteve a maior acurácia (99,83%), o maior F1 macro (99,83%), a maior AUC (0,99998), o maior Kappa (0,9977) e a maior acurácia balanceada (99,88%). Além disso, apresentou o menor valor de logLoss (0,0068), indicando

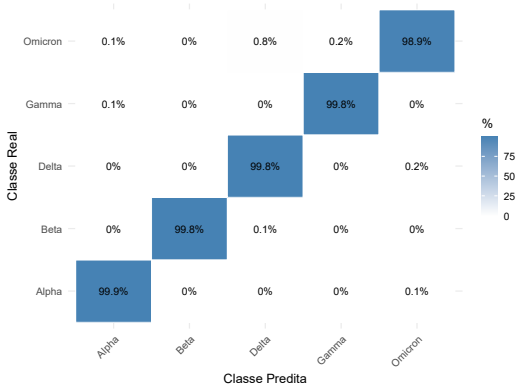

Figura 4. Matriz de confusão percentual para o classificador k-NN, obtida durante a validação cruzada (5-fold).

maior confiabilidade na calibragem das probabilidades preditas.

Tabela 5. Comparativo entre os classificadores com base na validação cruzada (5-fold).

Modelo	Acurácia	F1 macro	logLoss	AUC	Kappa	Acc. balanceada
SVM	0,9331	0,9296	0,4324	0,9982	0,9163	0,9582
Random Forest	0,9971	0,9971	0,0290	0,9999	0,9964	0,9982
XGBoost	**0,9983**	**0,9983**	**0,0068**	**0,99998**	**0,9977**	**0,9988**
k-NN	0,9966	0,9966	0,0303	0,9996	0,9957	0,9978

A Random Forest e o k-NN também demonstraram desempenhos altamente competitivos, com acurácia acima de 99,6%, Kappa superiores a 0,995 e AUC próximas de 1. A Random Forest apresentou resultados ligeiramente superiores ao k-NN, especialmente em logLoss e prAUC. Já o SVM obteve resultados satisfatórios, porém inferiores aos demais modelos, com acurácia de 93,31% e logLoss significativamente mais alto (0,4324), refletindo menor confiança nas probabilidades preditas.

Esses resultados indicam que modelos baseados em árvores de decisão são mais eficazes para a tarefa de classificação de variantes do SARS-CoV-2 a partir de embeddings de substituições de aminoácidos. Tais modelos se mostraram mais robustos, tanto em termos de desempenho absoluto quanto de generalização, mesmo em um espaço vetorial de alta dimensionalidade.

A Tabela 6 apresenta os resultados obtidos na avaliação dos modelos sobre o conjunto de teste externo, composto pelas amostras que não participaram do treinamento ou da validação cruzada. Observa-se que, de forma geral, todos os modelos mantiveram altos níveis de acurácia, indicando boa capacidade de generalização. O classificador XGBoost novamente se destacou, obtendo as maiores acurácias em todas as variantes: 99,91% para Alpha, 99,83% para Delta, 99,75% para Gamma e 99,61% para Omicron.

Os modelos Random Forest e k-NN apresentaram desempenhos muito próximos aos do XGBoost, também com acurácia superior a 99% nas quatro classes. O modelo SVM, por outro lado, teve desempenho comparável nas variantes Alpha, Delta e Omicron, mas apresentou acurácia significativamente inferior na variante Gamma (66,77%), o que confirma a tendência observada na validação cruzada de maior dificuldade do SVM

Tabela 6. Acurácia por modelo e por VOC no conjunto de teste externo.

Modelo	VOC	Acurácia	Total de Amostras
SVM	Alpha	0,9988	45.099
RF		0,9978	
XGB		**0,9991**	
KNN		0,9988	
SVM	Delta	0,9962	223.466
RF		0,9971	
XGB		**0,9983**	
KNN		0,9976	
SVM	Gamma	0,6677	6.937
RF		0,9967	
XGB		0,9975	
KNN		**0,9977**	
SVM	Omicron	0,9892	12.492
RF		0,9922	
XGB		**0,9961**	
KNN		0,9894	

em separar essa classe. Esses resultados reforçam a robustez dos modelos baseados em árvores (XGBoost e Random Forest) e validam a eficácia da estratégia proposta em dados completamente novos.

4. Conclusões

Este trabalho apresentou uma abordagem supervisionada para a classificação de variantes do SARS-CoV-2 (VOCs) utilizando embeddings vetoriais derivados de substituições de aminoácidos, gerados por modelos de linguagem natural, sem a necessidade de alinhamento genômico, coleta clínica ou experimentação laboratorial. Os experimentos demonstraram que classificadores supervisionados aplicados a essas representações, especialmente modelos baseados em árvores, como XGBoost e Random Forest, são capazes de atingir desempenho excepcional, com acurácia superior a 99% e elevada robustez mesmo em cenários de alta similaridade genômica entre variantes. O modelo XGBoost destacou-se com acurácia de 99,83% e logLoss de 0,0068, mantendo alta performance em dados externos. Os resultados validam o potencial da abordagem como uma solução eficiente, escalável e inteiramente digital para vigilância genômica automatizada, superando métodos tradicionais e híbridos da literatura recente.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte e financiamento.

Referências

Ali, S., Sahoo, B., Ullah, N., Zelikovskiy, A., Patterson, M., and Khan, I. (2021). A k-mer based approach for sars-cov-2 variant identification. In *Bioinformatics Research and Applications: 17th International Symposium, ISBRA 2021, Shenzhen, China, November 26–28, 2021, Proceedings 17*, pages 153–164. Springer.

Azevedo, K. S., de Souza, L. C., Coutinho, M. G. F., Barbosa, R. d. M., and Fernandes, M. A. C. (2024). Deepvirusclassifier: a deep learning tool for classifying sars-cov-2 based on viral subtypes within the coronaviridae family. *BMC Bioinformatics*, 25(231).

Beduk, D., de Oliveira Filho, J. I., Beduk, T., Harmanci, D., Zihnioglu, F., Cicek, C., Sertoz, R., Arda, B., Goksel, T., Turhan, K., et al. (2022). 'all in one'sars-cov-2 variant recognition platform: Machine learning-enabled point of care diagnostics. *Biosensors and Bioelectronics: X*, 10:100105.

Chourasia, P., Murad, T., Tayebi, Z., Ali, S., Khan, I. U., and Patterson, M. (2023). Efficient classification of sars-cov-2 spike sequences using federated learning. In *Annual International Conference on Information Management and Big Data*, pages 80–96. Springer.

Coutinho, M. G. F., Câmara, G. B. M., Barbosa, R. d. M., and Fernandes, M. A. C. (2023). Sars-cov-2 virus classification based on stacked sparse autoencoder. *Computational and Structural Biotechnology Journal*, 21:284–298.

Câmara, G. B. M., Coutinho, M. G. F., Silva, L. M. D. d., Gadelha, W. V. d. N., Torquato, M. F., Barbosa, R. d. M., and Fernandes, M. A. C. (2022). Convolutional neural network applied to sars-cov-2 sequence classification. *Sensors*, 22(15):5730.

de Souza, L. C., Azevedo, K. S., de Souza, J. G., Barbosa, R. d. M., and Fernandes, M. A. C. (2023). New proposal of viral genome representation applied in the classification of sars-cov-2 with deep learning. *BMC Bioinformatics*, 24(92).

Fatima, N. and Ahmad, A. (2024). Sars-cov-2 virus variant detection and mortality prediction through symptom analysis using machine learning. *Engineering Applications of Artificial Intelligence*, 130:107743.

Khare, S., Gurry, C., Freitas, L., Schultz, M. B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R. T., Yeo, N. K., Team, C., and Maurer-Stroh, S. (2021). Gisaid's role in pandemic response. *China CDC Weekly*, 3(49):1049–1051.

Promja, S., Puenpa, J., Achakulvisut, T., Poovorawan, Y., Lee, S. Y., Athamanolap, P., and Lertanantawong, B. (2023). Machine learning-assisted real-time polymerase chain reaction and high-resolution melt analysis for sars-cov-2 variant identification. *Analytical Chemistry*, 95(3):2102–2109.

Qin, J., Tian, X., Liu, S., Yang, Z., Shi, D., Xu, S., and Zhang, Y. (2024). Rapid classification of sars-cov-2 variant strains using machine learning-based label-free sers strategy. *Talanta*, 267:125080.

Ran, L., Tan, X., and Zhang, Y. (2022). Precise community-based public health management: Crucial experience responding to covid-19 in wuhan, china. *Risk Management and Healthcare Policy*, pages 171–178.

Singh, R., Nagpal, S., Pinna, N. K., and Mande, S. S. (2022). Tracking mutational semantics of sars-cov-2 genomes. *Scientific Reports*, 12(1):15704.

Sokhansanj, B. A., Zhao, Z., and Rosen, G. L. (2022). Interpretable and predictive deep neural network modeling of the sars-cov-2 spike protein sequence to predict covid-19 disease severity. *Biology*, 11(12).