UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE DO CAMPUS ARARANGUÁ
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

Yan Wiverson Lavor Bentes

# MusicalGestures: Reconhecimento Gestual utilizando Visão Computacional e Inteligência Artificial aplicado à Composição Musical

Araranguá
2024

Yan Wiverson Lavor Bentes

# MusicalGestures: Reconhecimento Gestual utilizando Visão Computacional e Inteligência Artificial aplicado à Composição Musical

Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia de Computação submetido ao Centro de Ciências, Tecnologias e Saúde do Campus Araranguá da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Engenharia de Computação.
Orientador: Prof. Antonio Carlos Sobieranski, Dr.
Coorientador: Prof. Tobias Rossi Müller

Araranguá
2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Bentes, Yan Wiverson Lavor
MusicalGestures: Reconhecimento Gestual utilizando Visão Computacional e Inteligência Artificial aplicado à Composição Musical / Yan Wiverson Lavor Bentes ; orientador, Antonio Carlos Sobieranski, coorientador, Tobias Rossi Müller, 2024.
27 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá, Graduação em Engenharia de Computação, Araranguá, 2024.

Inclui referências.

1. Engenharia de Computação. 2. Visão Computacional. 3. Reconhecimento de Gestos. 4. Música. I. Sobieranski, Antonio Carlos. II. Müller, Tobias Rossi. III. Universidade Federal de Santa Catarina. Graduação em Engenharia de Computação. IV. Título.

Yan Wiverson Lavor Bentes

**MusicalGestures: Reconhecimento Gestual utilizando Visão Computacional e Inteligência Artificial aplicado à Composição Musical**

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel em Engenharia de Computação e aprovado em sua forma final pelo Curso de Graduação em Engenharia de Computação.

Araranguá, 03 de Dezembro de 2024.


Documento assinado digitalmente
**Jim Lau**
Data: 17/12/2024 17:15:55-0300
CPF: ***.464.702-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Jim Lau, Dr.
Coordenador do Curso

**Banca Examinadora:**


Documento assinado digitalmente
**Antonio Carlos Sobieranski**
Data: 17/12/2024 18:48:18-0300
CPF: ***.305.809-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Antonio Carlos Sobieranski, Dr.


Documento assinado digitalmente
**TOBIAS ROSSI MULLER**
Data: 17/12/2024 16:59:13-0300
CPF: ***.669.569-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Tobias Rossi Müller
Coorientador


Documento assinado digitalmente
**RODRIGO VINICIUS MENDONCA PEREIRA**
Data: 17/12/2024 20:38:10-0300
CPF: ***.663.779-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Rodrigo Pereira, Dr.
Avaliador
Universidade Federal de Santa Catarina


Documento assinado digitalmente
**Jim Lau**
Data: 17/12/2024 17:16:17-0300
CPF: ***.464.702-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Jim Lau, Dr.
Avaliador
Universidade Federal de Santa Catarina

# MusicalGestures: Reconhecimento Gestual utilizando Visão Computacional e Inteligência Artificial aplicado à Composição Musical

Yan Wiverson Lavor Bentes*

2024, DEZEMBRO

## Resumo

O avanço das tecnologias de reconhecimento de gestos e visão computacional tem permitido o desenvolvimento de interfaces musicais inovadoras que oferecem novas formas de interação entre músicos e instrumentos. Este trabalho propõe uma interface musical baseada no reconhecimento de gestos em tempo real, através da detecção das mãos e um modelo de aprendizado de máquina treinado para classificar os gestos. A interface traduz os gestos reconhecidos em notas musicais. Ao final do trabalho uma demonstração prática do sistema foi realizada para validar qualitativamente sua capacidade de traduzir movimentos gestuais em notas musicais e usabilidade em tempo real. Os resultados obtidos destacam o potencial e possíveis melhorias da interface como ferramenta interativa para performances musicais tanto para pessoas com conhecimento musical prévio ou não, oferecendo novas possibilidades de criação e expressão artística.

**Palavras-chaves**: Visão Computacional, Reconhecimento de Gestos, Música.

*yanwiverson@gmail.com

# MusicalGestures: Reconhecimento Gestual utilizando Visão Computacional e Inteligência Artificial aplicado à Composição Musical

Yan Wiverson Lavor Bentes*

2024, DECEMBER

## Abstract

The advancement of gesture recognition and computer vision technologies has enabled the development of innovative musical interfaces that offer new forms of interaction between musicians and instruments. This work proposes a musical interface based on real-time gesture recognition, through hand detection and a machine learning model trained to classify gestures. The interface translates the recognized gestures into musical notes. At the end of the work, a practical demonstration of the system was carried out to qualitatively validate its ability to translate gestural movements into musical notes and real-time usability. The results obtained highlight the potential and possible improvements of the interface as an interactive tool for musical performances for both people with and without prior musical knowledge, offering new possibilities for creation and artistic expression.

**Key-words**: Computer Vision, Gesture Recognition, Music.

*yanwiverson@gmail.com

# 1 Introdução

Os avanços tecnológicos têm desempenhado um papel transformador no mundo da música, promovendo inovações que ampliam significativamente as possibilidades de interação entre músicos e instrumentos. Tradicionalmente, a prática musical era limitada ao uso de instrumentos convencionais, como piano, violão entre outros, instrumentos esses cuja maestria muitas vezes exige anos de estudo e prática constante devido à alta curva de aprendizado. Além disso, não é fácil tocar música em todos os momentos e lugares, alguns instrumentos musicais, como piano e bateria, não podem ser transportados facilmente, por essas razões, a performance musical pode acabar não se tornando uma tarefa tão acessível (SHANG; WANG, 2022). Nesse cenário, novas interfaces musicais emergem como alternativas cada vez menos custosas e mais acessíveis, trazendo soluções tecnológicas que permitem performances musicais em qualquer lugar e por qualquer pessoa, democratizando o acesso à expressão artística. Essas inovações não apenas podem tornar a criação musical mais acessível, mas também incentivam novas formas de experimentação e criatividade (BUSCHERT, 2012).

Com a produção de sensores de alta precisão, surgimento de sistemas de realidade aumentada e algoritmos avançados de reconhecimento de gestos, tem sido possível criar instrumentos musicais digitais que transcendem a simples imitação de instrumentos tradicionais. Essas interfaces não apenas replicam a funcionalidade dos instrumentos convencionais, mas também introduzem formas completamente novas de interação, transformando gestos corporais em expressões sonoras de maneira intuitiva e inovadora, como é o exemplo das MiMU Gloves que através de uma luva com sensores embutidos é capaz de conectar diferentes movimentos à parâmetros musicais (HUGHES; FIGUEROA, 2019).

O reconhecimento de gestos destaca-se como uma tecnologia versátil e de fácil acesso, visto que sua utilização requer apenas um computador equipado com câmera. Apesar de amplamente explorada em outras áreas, como reconhecimento de linguagem de sinais, robótica e jogos (YASEN; JUSOH, 2019), sua aplicação no campo musical ainda é limitada, representando uma oportunidade significativa para inovações criativas. Essa tecnologia permite que músicos se expressem de forma dinâmica, traduzindo movimentos corporais ou gestos das mãos diretamente no controle de parâmetros musicais e na síntese sonora, ampliando as possibilidades de performance artística (NIETO; SHASHA, 2013).

Apesar de seu grande potencial, a implementação de uma interface musical baseada no reconhecimento de gestos apresenta desafios consideráveis. O desempenho dessa tecnologia pode ser significativamente influenciado por variáveis ambientais, como iluminação, fundo, distância entre o usuário e a câmera, além de características individuais, como a cor da pele e a orientação das mãos. Além disso, a precisão na detecção dos gestos em tempo real e a capacidade de resposta do sistema são aspectos cruciais para garantir uma experiência satisfatória. Outro ponto a ser considerado é a usabilidade, pois os usuários precisam aprender o significado de uma ampla gama de gestos, o que pode se tornar uma barreira, especialmente para pessoas com diferentes níveis de experiência em música ou familiaridade com tecnologia (YASEN; JUSOH, 2019). Esses desafios exigem soluções tecnológicas que equilibrem inovação e acessibilidade, garantindo que a interface seja eficiente e intuitiva tanto para amadores quanto para profissionais.

Com o objetivo de superar os desafios mencionados, este trabalho apresenta o desenvolvimento de uma interface musical, que utiliza a tecnologia de reconhecimento de gestos em tempo real como principal meio de interação entre o músico e o instrumento digital. A proposta busca criar uma ferramenta que seja simultaneamente de fácil utilização

e precisa, traduzindo gestos em comandos musicais com fluidez e confiabilidade. Para alcançar esse objetivo, a solução combina diferentes tecnologias de forma integrada: captura de vídeo em tempo real, processamento de gestos e controle de síntese sonora para geração de notas musicais. Esse sistema visa não apenas atender às demandas de músicos profissionais que buscam novas formas de expressão artística, mas também ser acessível para amadores e iniciantes, promovendo uma experiência inclusiva e enriquecedora. Com isso, espera-se ampliar as possibilidades criativas na música e explorar os limites da interação entre tecnologia e arte.

O trabalho está organizado da seguinte maneira: na Seção 2 é realizada a apresentação de conceitos teóricos essenciais relacionados ao desenvolvimento do trabalho. A Seção 3 aborda trabalhos correlatos e as técnicas utilizadas pelos autores para a resolução de problemas similares. A Seção 4 apresenta qual o método proposto e o detalhamento do desenvolvimento de cada etapa. Na Seção 5 são apresentados os resultados obtidos e, por fim, a Seção 6 apresenta as considerações finais e trabalhos futuros.

# 2 Fundamentação Teórica

## 2.1 Teoria Musical

A teoria musical compreende o estudo detalhado dos elementos estruturais e expressivos que fundamentam a música, investigando como esses componentes interagem para criar composições harmônicas, melódicas e rítmicas. Elementos como tom, duração, timbre e dinâmica (ou volume) são amplamente reconhecidos como essenciais na estruturação musical. Além disso, outros aspectos, como a frequência e a textura sonora, também são frequentemente destacados por estudiosos como pilares fundamentais na construção de obras musicais (FEEZELL, 2011).

A teoria musical também se preocupa com a estruturação e organização dos sons dentro de um contexto musical, incluindo aspectos como escalas, intervalos, progressões harmônicas e formas musicais. Estes conceitos proporcionam a base para que compositores e intérpretes possam comunicar suas ideias musicais de forma clara e precisa. Além disso, a teoria musical serve como uma ferramenta essencial para músicos ao permitir uma compreensão mais aprofundada dos mecanismos que regem a criação e a execução de peças musicais.

### 2.1.1 Notas

As notas musicais representam a base de qualquer composição musical, sendo os blocos essenciais para a estruturação de melodias e harmonia. Cada nota é definida por uma vibração sonora em uma frequência específica, organizada dentro de escalas musicais que proporcionam ordem e contexto às composições. As notas naturais incluem Dó (C), Ré (D), Mi (E), Fá (F), Sol (G), Lá (A) e Si (B). Além dessas, existem variações conhecidas como sustenidos (#) e bemóis (♭), que alteram a altura das notas em meio tom para cima ou para baixo, respectivamente. Por exemplo, Dó sustenido (C#) corresponde a um meio tom acima de Dó, enquanto Ré bemol (D♭) está um meio tom abaixo de Ré. Curiosamente, algumas dessas notas compartilham a mesma frequência sonora, sendo chamadas de notas enarmônicas. Isso significa que, embora soem iguais, como C# e D♭, sua notação varia de acordo com o contexto musical e a intenção do compositor.

Tabela 1 – Tabela de Frequências, Períodos e Comprimentos de Onda

| Nº | Nota | Frequência (Hz) | Período (s) | Comprimento de Onda (m) |
|---|---|---|---|---|
| 60 | C4 | 523.251099 | 0.001911 | 0.657428 |
| 61 | C#4 | 554.365234 | 0.001804 | 0.620529 |
| 62 | D4 | 587.329529 | 0.001703 | 0.585702 |
| 63 | D#4 | 622.253906 | 0.001607 | 0.552829 |
| 64 | E4 | 659.255127 | 0.001517 | 0.521801 |
| 65 | F4 | 698.456482 | 0.001432 | 0.492515 |
| 66 | F#4 | 739.988403 | 0.001352 | 0.464872 |
| 67 | G4 | 783.990845 | 0.001276 | 0.438781 |
| 68 | G#4 | 830.609375 | 0.001205 | 0.414154 |
| 69 | A4 | 880.000000 | 0.001136 | 0.390909 |
| 70 | A#4 | 932.327576 | 0.001073 | 0.368969 |
| 71 | B4 | 987.766602 | 0.001012 | 0.348262 |
| 72 | C5 | 1046.502075 | 0.000956 | 0.328714 |

**Fonte:** (Iazzetta, Fernando, 2024)

### 2.1.2 Tom

O tom refere-se à percepção subjetiva da altura de um som, sendo uma característica fundamental que permite aos ouvintes distinguir notas musicais e compreender a melodia e harmonia de uma peça. Essa percepção está intimamente ligada à frequência das vibrações sonoras: sons com frequências mais altas são percebidos como tons mais agudos, enquanto frequências mais baixas resultam em tons mais graves. Embora a frequência seja o principal determinante do tom, outros fatores, como a amplitude e a riqueza harmônica do som, também influenciam a forma como o tom é percebido. Por exemplo, dois sons com a mesma frequência, mas com diferentes conteúdos harmônicos, podem ser percebidos de maneiras distintas devido às diferenças na qualidade sonora (YOST, 2009).

### 2.1.3 Timbre, Dinâmica e Duração

O timbre é a qualidade sonora que distingue diferentes fontes sonoras, mesmo quando elas produzem notas de igual altura e intensidade. Essa característica é essencial para identificar e diferenciar instrumentos musicais, vozes ou outros sons. Estudos em percepção auditiva demonstram que o timbre está relacionado à distribuição de energia nas frequências espectrais e às variações temporais dessa distribuição, que juntas formam a base da nossa percepção da "personalidade"sonora de um instrumento. Por exemplo, enquanto um violino e uma flauta podem executar a mesma nota, as propriedades acústicas únicas de cada instrumento criam timbres distintos, permitindo que os ouvintes identifiquem facilmente sua origem sonora (WESSEL, 1979).

A dinâmica (ou volume) na música refere-se à intensidade ou força com que uma nota musical é tocada. As dinâmicas não apenas influenciam o volume perceptível de uma nota musical, mas também desempenham um papel crucial na expressão emocional da música. Já a duração é uma medida do comprimento de tempo de uma nota, pausa, ou qualquer outro som. A duração é essencial para definir o ritmo e o andamento de uma peça musical, contribuindo para a estrutura rítmica e a fluidez de uma composição (FEEZELL, 2011).

## 2.2 Melodia

A melodia é amplamente reconhecida como um dos elementos mais marcantes e memoráveis da música, sendo frequentemente considerada a essência de uma composição. Ela pode ser descrita como uma sequência linear de notas que se destaca na textura musical, proporcionando coesão e significado emocional à obra. Embora sua estrutura possa parecer simples à primeira vista, a melodia é altamente subjetiva e depende tanto da percepção dos ouvintes quanto do contexto em que está inserida. Como tal, ela desempenha um papel crucial na forma como as pessoas experienciam e se conectam com a música (POLINER et al., 2007).

## 2.3 Processamento Digital de Imagens

O processamento digital de imagens é uma área fundamental da computação que se dedica à manipulação e análise de imagens digitais por meio de algoritmos avançados. Essa disciplina inclui técnicas como filtragem de imagens, segmentação e transformação de pixels, que são empregadas para realçar características visuais, identificar objetos em uma cena e extrair informações relevantes. Essas operações permitem aplicações que vão desde a melhoria da qualidade visual de fotografias até o reconhecimento de padrões complexos em tempo real, como placas de trânsito, rostos humanos e gestos manuais.

Uma aplicação essencial dessa área é o reconhecimento de objetos, como mãos, rostos e corpos humanos, em imagens ou vídeos. Esse processo é realizado através de algoritmos que identificam características específicas e extraem informações relevantes para análises posteriores ou interações. No contexto deste trabalho, a identificação de pontos de referência ou pontos-chave (landmarks), que representam articulações ou extremidades importantes, é uma das técnicas mais avançadas para rastrear e interpretar movimentos com precisão.

### 2.3.1 Visão Computacional e Reconhecimento de Gestos

A visão computacional, uma subárea do processamento de imagens, visa capacitar sistemas computacionais a interpretar e compreender informações visuais do mundo real. No reconhecimento de gestos, a visão computacional permite que sistemas detectem, rastreiem e analisem movimentos humanos, traduzindo-os em ações ou comandos. Esse processo é facilitado por uma combinação de algoritmos de visão computacional e aprendizado de máquina, que pode ser classificada em métodos baseados em sensores e métodos baseados em visão.

Métodos baseados em sensores utilizam dispositivos especializados que capturam diretamente os movimentos físicos. Exemplos incluem luvas de dados, que medem flexão dos dedos, rotação do punho e aceleração do movimento por meio de sensores embutidos, e dispositivos como o Leap Motion, que utiliza câmeras infravermelhas para rastrear mãos em um espaço tridimensional (LOKHANDE; PRAJAPATI; PANSARE, 2015). Esses métodos oferecem alta precisão, mas dependem de hardware dedicado e, muitas vezes, custoso (GUZSVINECZ; SZUCS; SIK-LANYI, 2019).

Por outro lado, métodos baseados em visão processam imagens ou vídeos capturados por câmeras comuns, como webcams ou câmeras de dispositivos móveis, para interpretar gestos. Esses métodos são não invasivos, pois não exigem que os usuários usem dispositivos físicos. A flexibilidade e acessibilidade tornam essas abordagens amplamente populares, especialmente para aplicações que exigem interação em tempo real.

Entre as técnicas mais eficazes para reconhecimento de gestos está a detecção de pontos de referência. Essa abordagem identifica articulações específicas das mãos (como pontas dos dedos e a base da palma), fornecendo informações detalhadas para análise de posição e movimento. Ferramentas como o MediaPipe Hands, que emprega redes neurais convolucionais (CNNs), são amplamente utilizadas devido à sua precisão na detecção de pontos de referência em três dimensões (ZHANG et al., 2020). Além disso, o rastreamento de movimento (motion tracking), utilizando algoritmos como Optical Flow, complementa o reconhecimento ao analisar a direção, velocidade e trajetória dos gestos ao longo do tempo (HONGWEI et al., 2015).

## 2.4 Aprendizado de Máquina

Segundo (MITCHELL, 1997) o aprendizado de máquina é o estudo de algoritmos computacionais que melhoram automaticamente através da experiência. Esses algoritmos são aplicados a uma ampla variedade de problemas, desde a descoberta de padrões gerais em grandes volumes de dados até sistemas de recomendação que aprendem as preferências dos usuários. Embora essa definição tenha evoluído com o tempo, ela ainda captura a essência do aprendizado de máquina como um pilar da inteligência artificial.

Atualmente é possível dividir o aprendizado de máquina em três categorias principais: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. No aprendizado supervisionado, o algoritmo é treinado com um conjunto de dados rotulados, onde o objetivo é prever as saídas com base nas entradas especificadas. No aprendizado não supervisionado, o sistema tenta identificar padrões e relações nos dados sem rótulos explícitos, enquanto o aprendizado por reforço envolve o treinamento de um agente que interage com um ambiente e aprende segundo recompensas e penalidades de acordo suas ações (GERON, 2019).

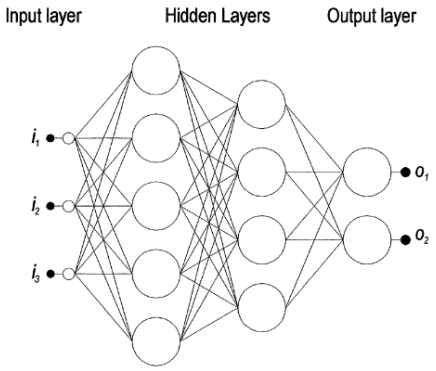
No contexto do reconhecimento de gestos, o aprendizado supervisionado desempenha um papel crucial. Modelos como Redes Neurais Convolucionais (CNNs) são frequentemente empregados para processar imagens ou vídeos de mãos e identificar padrões visuais que correspondem a diferentes gestos. Durante o treinamento, esses modelos aprendem a partir de grandes quantidades de dados rotulados, permitindo que façam previsões precisas após identificar características visuais específicas em novos dados.

### 2.4.1 Perceptron Multicamadas

O Perceptron Multicamadas (MLP, do inglês Multilayer Perceptron) é uma arquitetura clássica de redes neurais artificiais amplamente utilizada em tarefas de classificação e regressão. Caracterizado por sua estrutura feed-forward, onde sinais são propagados da camada de entrada para a de saída sem conexões cíclicas, o MLP consiste em uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada neurônio na rede utiliza funções de ativação não lineares, como ReLU ou sigmoide, para transformar dados de entrada em representações mais complexas (POPESCU et al., 2009).

Um diferencial do MLP em relação ao perceptron de camada única é sua capacidade de resolver problemas não linearmente separáveis. As camadas ocultas permitem ao modelo criar fronteiras de decisão complexas ao combinar os pesos de entrada com funções de ativação. Essa característica torna o MLP ideal para tarefas que exigem detecção de padrões sofisticados, como o reconhecimento de gestos (RUCK; ROGERS; KABRISKY, 1990).

Figura 1 – Perceptron Multicamadas




**Fonte:** (GARDNER; DORLING, 1998)

O perceptron multicamadas é treinado principalmente usando o algoritmo de retropropagação do erro (backpropagation), que ajusta os pesos das conexões neurais para minimizar o erro entre a saída prevista e o valor esperado (RUCK; ROGERS; KABRISKY, 1990). O treinamento começa com a inicialização aleatória dos pesos, seguido pela propagação dos dados de entrada pela rede até a camada de saída, onde é calculado o erro. O algoritmo então propaga esse erro de volta pelas camadas, ajustando os pesos de cada conexão para minimizar o erro em iterações subsequentes. Esse processo é repetido por várias épocas até que o erro da rede seja suficientemente pequeno (RUCK; ROGERS; KABRISKY, 1990). Embora eficiente, a retropropagação pode ser lenta, especialmente para problemas complexos que requerem muitas iterações de treinamento. Métodos para acelerar o aprendizado, como o uso de taxa de aprendizado adaptativa e o método do momento, são frequentemente implementados para melhorar o tempo de convergência (POPESCU et al., 2009).

## 3 Trabalhos Correlatos

Em buscas realizadas na literatura foram encontrados diversos trabalhos que tratam de reconhecimento de gestos em tempo real, alguns destes trabalhos possuíam aplicação a música também. O método mais comum de reconhecimento de gestos utilizado nessas pesquisas envolve o uso de visão computacional, com destaque para técnicas que incluem redes neurais convolucionais e algoritmos de aprendizado profundo, que permitem uma análise detalhada e precisa dos movimentos capturados por câmeras ou sensores. As buscas foram realizadas nos portais ScienceDirect®, IEEE Xplore®, Web of Science® e Scopus®, utilizando combinações das seguintes palavras-chave: “computer vision”, “gesture recognition”, “music performance”.

A seguir, detalham-se os estudos mais relevantes encontrados durante a revisão bibliográfica. Estes trabalhos foram selecionados por sua relevância tecnológica, metodológica e aplicação prática, fornecendo uma visão abrangente das possibilidades atuais e futuras no campo do reconhecimento de gestos para performance musical e outras interações humano-computador.

O artigo Zhang e Zhang (2020) propõe um sistema inovador para tocar o Guqin, um instrumento tradicional chinês, utilizando gestos reconhecidos pelo sensor Kinect. O

objetivo do sistema é estimular o interesse pelo Guqin, permitindo uma interação natural e sem contato direto com o instrumento real. Através de uma interface virtual intuitiva e tecnologias de Realidade Aumentada, os usuários podem manipular a interface gráfica do Guqin com gestos corporais, gerando sons autênticos do instrumento. O sistema utiliza o Kinect para rastrear os movimentos do usuário, traduzindo os gestos em comandos musicais. Testes com 20 participantes demonstraram que o sistema aumenta o interesse pelo Guqin e facilita a exploração musical da cultura chinesa, mesmo para usuários sem conhecimento prévio do instrumento. O método é de fácil implementação em espaços públicos e pode ser expandido para funções educacionais e exploração musical mais complexa no futuro.

Shang e Wang (2022) propõem um método de performance musical baseado no reconhecimento de gestos, permitindo que usuários controlem parâmetros musicais como a nota, oitava, e alterações como sustenidos e bemóis através de gestos personalizados. O sistema utiliza uma câmera e algoritmos de visão computacional, como o YOLOv3 e o RexNetV1, para detectar pontos-chave das mãos em tempo real e determinar qual gesto está sendo performado.

Reconhecimento de gestos já foi utilizado também para o design e desenvolvimento de um sistema de player de música inteligente que utiliza o reconhecimento de gestos para interação com o usuário. De acordo com o trabalho Xing e Lei (2022) o sistema permite que os usuários controlem o player através de gestos como pausar, reproduzir, trocar músicas e ajustar o volume. A abordagem utiliza descritores de Fourier para extrair contornos dos gestos e um perceptron multicamadas para reconhecer gestos com alta precisão e em tempo real. O sistema é implementado em dispositivos Android, funcionando de forma eficiente e intuitiva, o que foi comprovado por meio de testes que demonstraram sua eficácia e responsividade. Além disso, o trabalho foi premiado em competições de inteligência artificial, destacando-se pela inovação em interfaces de interação baseadas em gestos.

O método Kodály é uma abordagem pedagógica para educação musical desenvolvida pelo compositor e educador húngaro Zoltán Kodály. Neste contexto, o método utiliza sinais de mão específicos para representar notas musicais, facilitando o ensino e aprendizado de melodia, entonação e percepção musical. No trabalho Sari e Utomo (2023) os autores propuseram um sistema que reconhece esses gestos de mão para classificar tons musicais de forma automatizada, com o objetivo de permitir a interação musical por meio de gestos manuais, focando especialmente no instrumento tradicional angklung. Para isso, uma Rede Neural Convolucional (CNN) foi treinada com um dataset de 20.000 imagens de sinais de mão Kodály, alcançando uma precisão média de 98,29%. A metodologia incluiu etapas de pré-processamento das imagens (como segmentação e conversão para tons de cinza), seguida pelo treinamento de uma CNN com cinco camadas convolucionais e max pooling, culminando em uma simulação de classificação de gestos em tempo real para uso em performances musicais.

O trabalho Chang et al. (2015) descreve um sistema para controlar instrumentos musicais virtuais em tempo real usando redes neurais e o dispositivo Leap Motion, um dispositivo que suporta movimentos de mãos e dedos como entrada, de forma análoga a um mouse, mas não requer contato ou toque manual para reconhecimento de gestos manuais. O método permite que usuários toquem diferentes instrumentos simulando suas interações com gestos naturais, sem precisar de instrumentos físicos. O Leap Motion captura os movimentos das mãos e a rede neural processa essas informações para controlar instrumentos virtuais, acionando eventos como ativar acordes ou ajustar tons. O objetivo é fornecer uma experiência acessível e personalizada para tocar música.

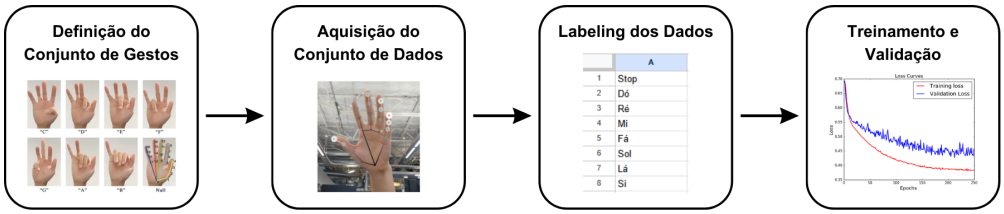
A aplicação desenvolvida no trabalho Begum et al. (2017) apresenta uma aplicação para Android que permite aos usuários tocar um piano virtual usando uma representação das teclas do piano desenhadas em uma folha de papel. O aplicativo usa a câmera do smartphone para capturar uma imagem do teclado de piano desenhado e processa essa imagem em tempo real para detectar a posição dos dedos. Quando os dedos tocam as teclas desenhadas, o aplicativo reproduz o som correspondente, simulando a experiência de tocar um piano real. A aplicação converte a imagem da câmera para o espaço de cor HSV, separando informação de cor e intensidade para detectar teclas e dedos com precisão.

# 4 Abordagem Proposta

O desenvolvimento deste trabalho está estruturado em dois fluxos principais: (i) fluxo de dados, que compreende a construção do modelo de reconhecimento de gestos, representado na Figura 2, e (ii) fluxo da aplicação, no qual são realizados o processamento de vídeo em tempo real e a síntese sonora, como ilustrado na Figura 3. Essas etapas foram organizadas para garantir uma abordagem sistemática e eficiente na implementação do sistema.

No fluxo de dados, a primeira etapa consiste na definição de um conjunto de gestos que será reconhecidos pelo modelo. Em seguida, ocorre a aquisição dos pontos de referência, que serão utilizadas para representar cada gesto. Na terceira etapa, realiza-se o processo de rotulagem (labeling) dos dados, associando cada conjunto de pontos de referência ao gesto correspondente. Por fim, o modelo passa pelas fases de treinamento, validação e salvamento, assegurando sua capacidade de generalizar o reconhecimento dos gestos em situações variadas

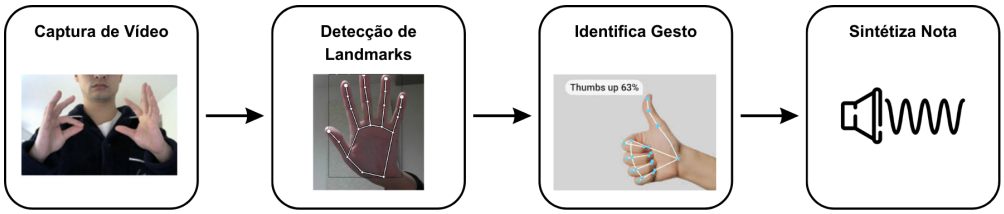
Figura 2 – Fluxo de Dados




**Fonte:** Próprio Autor

Já no fluxo da aplicação, o processo inicia-se com a captura de vídeo em tempo real. Em cada quadro capturado, ocorre a detecção de pontos de referência das mãos, permitindo a identificação precisa dos movimentos. Essas informações são então processadas para estimar qual gesto está sendo realizado e com base no gesto detectado, o sistema traduz a ação em comandos musicais, sintetizando a nota musical correspondente. Esse fluxo garante que a interação entre o usuário e o sistema ocorra de forma fluida e responsiva. Ambos os fluxos descritos acima são detalhados nas subseções seguintes, fornecendo uma visão abrangente de cada etapa do desenvolvimento e integração do sistema.

Figura 3 – Fluxo da Aplicação



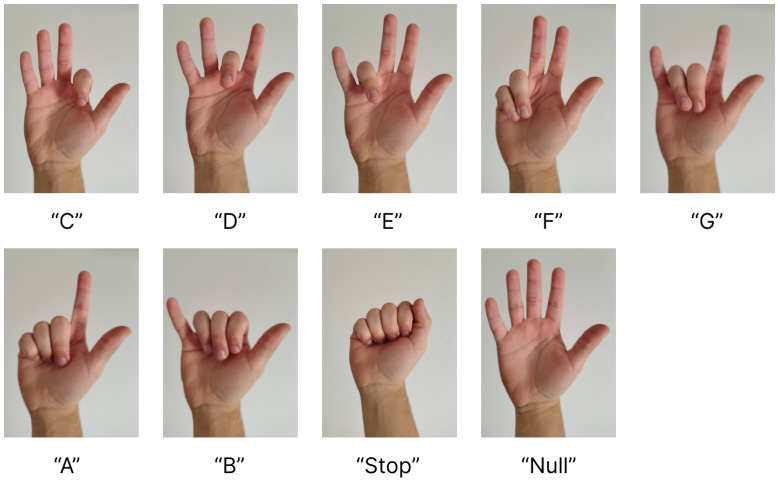

**Fonte:** Próprio Autor

## 4.1 Etapa 1: Fluxo de Dados

### 4.1.1 Definição do Conjunto de Gestos

O conjunto de gestos pode ser composto por no máximo 10 gestos estáticos, que serão treinados e posteriormente identificados pelo modelo. É necessário que os gestos tenham diferenças claras entre si, pois gestos com características muito similares podem comprometer a precisão e a responsividade do modelo, dificultando a distinção entre as classes. O conjunto de dados foi proposto pelo próprio autor pensando em facilitar a transição entre diferentes gestos, tendo inspirações no conjunto de gestos apresentado no trabalho de Shang e Wang (2022).

Figura 4 – Conjunto de Gestos Definido



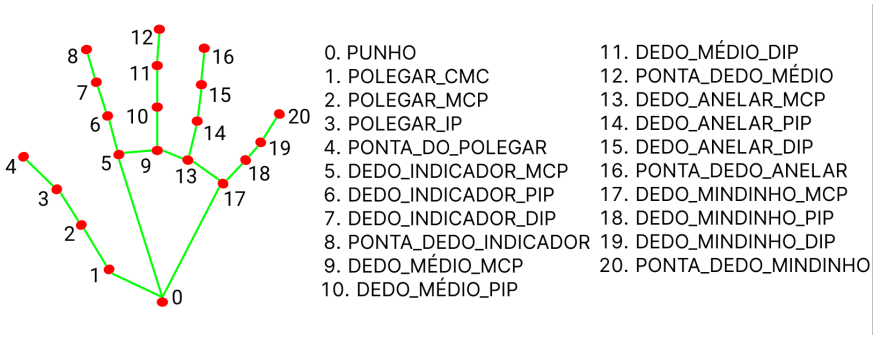

**Fonte:** Próprio Autor

### 4.1.2 Aquisição do Conjunto de Dados

Todas as amostras do conjunto de dados utilizado para a realização deste trabalho foram coletadas pelo próprio autor. Para a construção do conjunto de dados foi utilizado uma implementação baseada no uso do framework MediaPipe para a detecção de mãos. O processo foi adaptado do trabalho desenvolvido por Takahashi (2020) com a finalidade de capturar os pontos de referência das mãos garantindo que esses dados fossem representativos para o treinamento do modelo de reconhecimento de gestos.

Cada amostra no conjunto de dados é composta por um identificador numérico, variando de 0 a 8, que está associado ao gesto correspondente, além de 21 pares de coordenadas (x, y) que representam os pontos de referência das mãos, como ilustrado na Figura 5.

Figura 5 – Legenda dos pontos de referências




**Fonte:** Figura Adaptada de Gesture Recognition Task Guide

Para garantir a consistência dos dados em diferentes condições de câmera e evitar problemas como variação de resolução, foi realizado um pré-processamento das coordenadas dos pontos de referência da mãos antes de armazená-las. Esse processo incluiu a normalização das coordenadas e sua conversão para um vetor unidimensional, como descrito na Figura 6.

Figura 6 – Pré-Processamento dos Dados

① Coordenadas dos pontos de referência

| ID : 0 | ID : 1 | ID : 2 | ID : 3 | ··· | ID : 17 | ID : 18 | ID : 19 | ID : 20 |
|---|---|---|---|---|---|---|---|---|
| [551, 465] | [485, 428] | [439, 362] | [408, 307] | ... | [633, 315] | [668, 261] | [687, 225] | [702, 188] |

② Coordenadas relativas a partir do ID:0

| ID : 0 | ID : 1 | ID : 2 | ID : 3 | ··· | ID : 17 | ID : 18 | ID : 19 | ID : 20 |
|---|---|---|---|---|---|---|---|---|
| [0, 0] | [-66, -37] | [-112, -103] | [-143, -158] | ... | [82, -150] | [117, -204] | [136, -240] | [151, -277] |

③ Conversão para matriz unidimensional

| ID : 0 | | ID : 1 | | ID : 2 | | ID : 3 | | ··· | ID : 17 | | ID : 18 | | ID : 19 | | ID : 20 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | -66 | -37 | -112 | -103 | -143 | -158 | ... | 82 | -150 | 117 | -204 | 136 | -240 | 151 | -277 |

④ Coordenadas normalizadas para o valor máximo(valor absoluto)

| ID : 0 | | ID : 1 | | ID : 2 | | ID : 3 | | ··· | ID : 17 | | ID : 18 | | ID : 19 | | ID : 20 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | -0.24 | -0.13 | -0.4 | -0.37 | -0.52 | -0.57 | ... | 0.296 | -0.54 | 0.422 | -0.74 | 0.491 | -0.87 | 0.545 | -1 |

**Fonte:** Figura Adaptada de (TAKAHASHI, 2020)

### 4.1.3 Labeling de Dados

Para que o modelo funcione corretamente, cada identificador numérico foi associado a um nome específico que descreve o gesto correspondente. No conjunto de dados, foram definidos 9 gestos: o gesto neutro (mão aberta), o gesto de parada da síntese sonora (mão fechada) e 7 gestos associados às notas musicais (dó, ré, mi, fá, sol, lá, si). Essa rotulagem é essencial para que o modelo consiga mapear corretamente os gestos às ações musicais desejadas.

### 4.1.4 Definição da Arquitetura da Rede Neural

Antes da etapa de treinamento e validação é importante definir uma rede neural que seja adequada para a tarefa que será realizada. O sistema de reconhecimento de gestos desenvolvido neste trabalho utiliza uma rede neural com o objetivo de identificar gestos das mãos a partir dos pontos de referência obtidos através de uma abordagem de visão computacional. A rede neural implementada é um perceptron multicamadas, composto por várias camadas totalmente conectadas. Essa escolha deve-se à capacidade do MLP de modelar relações não lineares complexas entre os pontos de referência das mãos e os gestos associados (TAKAHASHI, 2020).

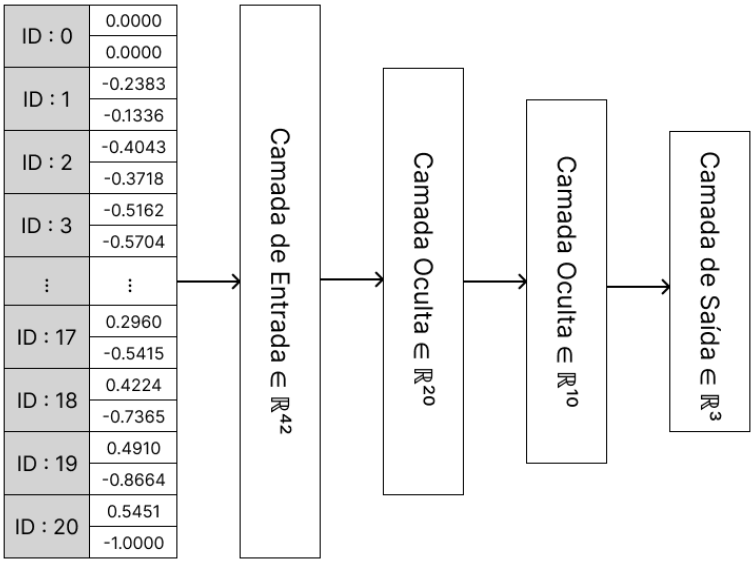
O modelo foi implementado utilizando a biblioteca TensorFlow com a API (Application Programming Interface) Keras, que oferece ferramentas de alto nível para a construção de redes neurais. Os principais elementos da arquitetura são detalhados abaixo:

- **Entrada:** A camada de entrada é responsável por receber as características que representam os dados de entrada da rede. Neste caso, cada amostra de entrada é um vetor de 42 elementos, correspondendo aos 21 pontos-chave das mãos, com coordenadas x e y obtidas pelo MediaPipe. Esse vetor representa a posição bidimensional de cada ponto-chave, essencial para identificar padrões espaciais associados a diferentes gestos.
- **Camadas Ocultas:**
  - Em redes neurais, as camadas ocultas são responsáveis por transformar os dados de entrada em representações intermediárias que permitem ao modelo aprender padrões mais complexos. No MLP utilizado neste trabalho, as camadas ocultas são compostas por duas camadas densas, ou totalmente conectadas, que aplicam uma combinação linear dos valores da entrada e passam esses valores por uma função de ativação não linear para introduzir complexidade ao modelo.
  - *Primeira camada oculta:* Composta por 20 unidades (neurônios), essa camada aplica uma função de ativação ReLU (Rectified Linear Unit), que é amplamente utilizada em redes neurais por sua simplicidade e eficácia na introdução de não linearidades. A ReLU transforma os valores negativos em zero (ABADI et al., 2015), permitindo que o modelo lide com problemas onde os pesos da rede são atualizados muito lentamente e ajude a treinar redes profundas. Para reduzir o risco de overfitting (explicação do termo na Seção 4.1.5), esta camada é seguida por uma camada de Dropout de 20%, que aleatoriamente desativa 20% dos neurônios em cada iteração de treinamento, aumentando a capacidade de generalização do modelo.
  - *Segunda camada oculta:* Essa camada contém 10 unidades e também utiliza a função de ativação ReLU. A redução no número de neurônios em comparação com a camada anterior ajuda a simplificar o modelo, limitando sua complexidade e evitando que ele aprenda detalhes excessivamente específicos do conjunto de treinamento. Aqui, é aplicado um Dropout de 40%, o que aumenta ainda mais a robustez do modelo contra o overfitting. As camadas de Dropout são particularmente úteis quando se trabalha com conjuntos de dados limitados ou quando o modelo possui um grande número de parâmetros, como no caso de redes com várias camadas densas.

- **Saída:** A camada de saída é composta por 9 neurônios, onde cada um corresponde a uma das nove classes de gestos que o modelo deve reconhecer. Para problemas de classificação multiclasse, como este, a função de ativação escolhida para a camada de saída é a *softmax*. Essa função transforma os valores de saída de cada neurônio em uma probabilidade, de modo que a soma das probabilidades seja igual a 1 (ABADI et al., 2015). A classe correspondente ao neurônio com o maior valor de probabilidade é então atribuída como a previsão final do modelo para cada entrada. Essa camada final permite que o modelo faça uma escolha entre as nove classes de gestos possíveis com base nas características aprendidas nas camadas ocultas.

Esses elementos da rede neural trabalham juntos para transformar as coordenadas dos pontos-chave da mão em previsões de gestos. A arquitetura multicamadas e a utilização de funções de ativação não lineares permitem ao modelo capturar e diferenciar padrões complexos entre os diferentes gestos, essencial para alcançar alta precisão. A presença de camadas de Dropout auxilia o modelo a generalizar melhor, reduzindo a dependência de exemplos específicos e aumentando sua capacidade de reconhecer corretamente gestos em condições variadas.

Figura 7 – Representação da estrutura do modelo sequencial desenvolvido




**Fonte:** Figura Adaptada de (TAKAHASHI, 2020)

### 4.1.5 Treinamento e Validação

A etapa de treinamento e validação do modelo é de suma importância para o funcionamento correto do sistema. No contexto de aprendizado de máquina, o processo de treinamento consiste em ensinar um modelo de aprendizado de máquina a realizar a tarefa de classificar os gestos das mãos. Durante o treinamento, o modelo ajusta seus parâmetros internos com o objetivo de reduzir a diferença entre as previsões feitas e os valores corretos presentes no conjunto de dados de treinamento. Esse processo é repetido por várias iterações, e a diferença entre a previsão do modelo e a resposta correta é medida por uma função de custo ou erro, no caso foi utilizado a função de perda entropia

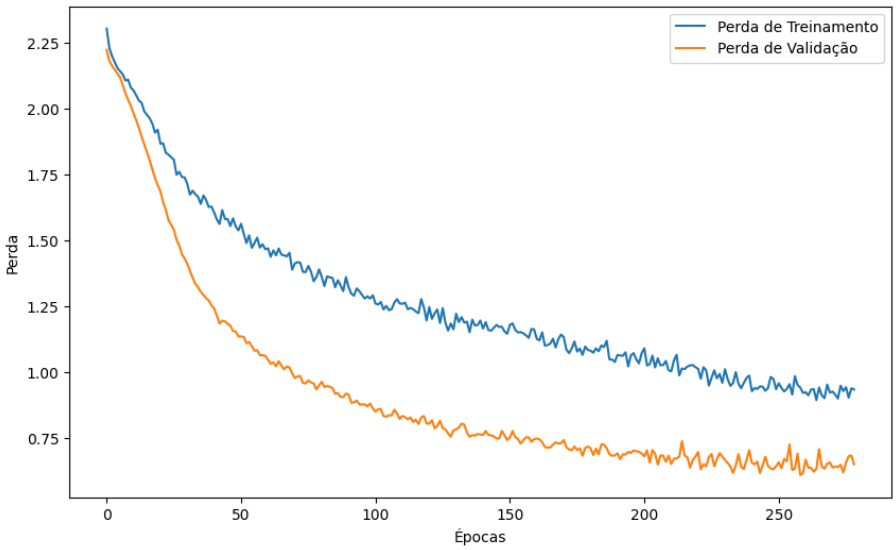
cruzada categórica, que é amplamente utilizada em problemas de classificação. Essa função calcula a diferença entre a distribuição de probabilidade das previsões do modelo e as respostas corretas, penalizando previsões incorretas (ABADI et al., 2015). Quanto maior essa diferença, maior será o valor da perda. O objetivo é minimizar essa perda para que o modelo aprenda a fazer previsões mais precisas ao longo do tempo.

A cada iteração, o modelo passa por uma série de ajustes usando um algoritmo de otimização de gradiente descendente, que ajusta os pesos e parâmetros do modelo para reduzir o erro. Dessa forma, ao final do treinamento, o modelo deve estar preparado para identificar corretamente os gestos a partir dos dados de entrada. O otimizador utilizado é o Adam, um método baseado em gradiente descendente que adapta a taxa de aprendizado automaticamente para cada parâmetro (CHOLLET et al., 2015).

No entanto, garantir a generalização do modelo é tão importante quanto otimizar sua precisão no treinamento. O overfitting — quando o modelo aprende padrões específicos do conjunto de treinamento, mas falha em generalizar para novos dados — é um problema recorrente em aprendizado de máquina (DIETTERICH, 1995). Para mitigar esse risco, foi utilizado o método de validação cruzada, em que uma fração do conjunto de dados (25% no caso deste trabalho) foi reservada exclusivamente para avaliar o desempenho do modelo em dados não vistos durante o treinamento.

Ao longo de 1000 épocas, foram monitoradas tanto a perda no conjunto de treinamento quanto no conjunto de validação A Figura 8 mostra a evolução dessas perdas ao longo das épocas. Observa-se que, conforme o treinamento avança, ambas as perdas apresentam uma tendência de diminuição, o que indica que o modelo está aprendendo a reconhecer padrões nos dados. No entanto, a perda de validação é particularmente importante, pois permite avaliar o comportamento do modelo em dados que não foram vistos durante o treinamento, sendo uma métrica essencial para identificar problemas de overfitting.

Figura 8 – Gráfico da Evolução da Função de Perda Durante o Treinamento

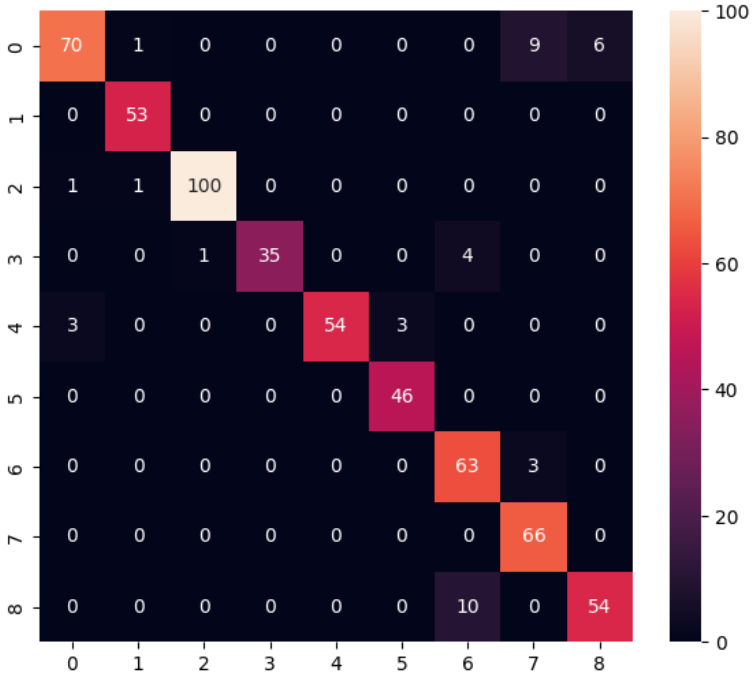



**Fonte:** Próprio Autor

Além da função de perda, o desempenho do modelo foi avaliado utilizando métricas complementares: matriz de confusão, precisão, recall e F1-score. Essas métricas oferecem uma visão detalhada sobre a capacidade do modelo em identificar corretamente as diferentes classes de gestos.

A matriz de confusão, apresentada na Figura 9, revela os padrões de acertos e erros do modelo. Observou-se maior confusão entre as classes 0 (Gesto Null) e 8 (Gesto Stop), o que é justificável devido às semelhanças visuais entre esses gestos (ver Figura 4). Por outro lado, classes como 1 (Dó) e 2 (Ré) exibiram maior consistência nas classificações, indicando que o modelo conseguiu distinguir gestos com características mais distintas.

Figura 9 – Matriz Confusão do Modelo de Classificação de Gestos




**Fonte:** Próprio Autor

A precisão indica a proporção de previsões corretas entre todas as instâncias previstas para uma classe, refletindo a confiabilidade das classificações positivas. No modelo, todas as classes apresentaram precisão próxima de 1.0, indicando alta confiabilidade. O recall mede a capacidade de identificar corretamente todas as instâncias de uma classe, sendo essencial para minimizar falsos negativos. Com valores médios de recall de 0,98, o modelo mostrou-se eficaz em detectar a maioria das instâncias de cada classe.

Por fim, o F1-score, uma média harmônica entre precisão e recall, alcançou um valor médio de 0,99, demonstrando um bom equilíbrio entre as métricas. A acurácia geral foi de 99%, sugerindo um desempenho robusto. A Tabela 2 resume as métricas de precisão, recall e F1-score, evidenciando a eficácia do modelo em classificar gestos de forma precisa e generalizável.

Tabela 2 – Relatório de Métricas de Avaliação

| Classe | Precisão | Recall | F1-Score | Suporte |
|---|---|---|---|---|
| 0 | 0.95 | 0.81 | 0.88 | 86 |
| 1 | 0.96 | 1.00 | 0.98 | 53 |
| 2 | 0.99 | 0.98 | 0.99 | 102 |
| 3 | 1.00 | 0.88 | 0.93 | 40 |
| 4 | 1.00 | 0.90 | 0.95 | 60 |
| 5 | 0.94 | 1.00 | 0.97 | 46 |
| 6 | 0.82 | 0.95 | 0.88 | 66 |
| 7 | 0.85 | 1.00 | 0.92 | 66 |
| 8 | 0.90 | 0.84 | 0.87 | 64 |
| Acurácia | | | 0.93 | 583 |
| Média Macro | 0.93 | 0.93 | 0.93 | 583 |
| Média Ponderada | 0.93 | 0.93 | 0.93 | 583 |

**Fonte:** Próprio Autor

## 4.2 Etapa 2: Fluxo da Aplicação

### 4.2.1 Captura de Vídeo

A captura de vídeo representa a etapa inicial no processo de reconhecimento de gestos, sendo essencial para a aquisição dos dados visuais necessários ao funcionamento do sistema. Neste projeto, utilizou-se a biblioteca OpenCV (Open Source Computer Vision Library), uma ferramenta amplamente reconhecida por sua eficiência e flexibilidade em tarefas de processamento de imagens. A OpenCV é responsável por acessar a câmera do dispositivo, configurando-a para capturar uma sequência contínua de quadros em tempo real, que serão processados pelo sistema.

Durante a captura, a qualidade dos quadros obtidos depende de diversos fatores, como a resolução da câmera, a taxa de quadros por segundo (FPS, do inglês Frames Per Second), e as condições de iluminação do ambiente. Esses fatores podem influenciar diretamente a precisão na detecção e classificação dos gestos, tornando necessário um equilíbrio entre desempenho computacional e qualidade de imagem.

### 4.2.2 Detecção dos Pontos de Referência

A detecção dos pontos de referência da mão é uma etapa crítica no funcionamento do sistema, sendo responsável por identificar e localizar com precisão os principais pontos anatômicos das mãos do usuário em tempo real. Essa tarefa é realizada com o auxílio da biblioteca MediaPipe, desenvolvida pelo Google, que fornece um modelo pré-treinado altamente eficiente e otimizado para a detecção dos pontos de referência das mãos. O modelo é capaz de identificar até 21 pontos em cada mão, abrangendo desde a ponta dos dedos até as articulações da palma, formando uma representação detalhada e robusta da estrutura da mão.

O funcionamento ocorre em cada quadro capturado pelo sistema, no qual o modelo processa a imagem e retorna as coordenadas dos pontos de referência para cada mão detectada. Essas coordenadas, originalmente em pixels, são normalizadas em um intervalo entre 0 e 1. Essa normalização permite que o sistema mantenha consistência no cálculo das posições dos pontos de referência, independentemente da resolução da câmera ou do tamanho do quadro capturado. Essa abordagem garante que o sistema funcione de

maneira uniforme em diferentes configurações de hardware, adaptando-se a dispositivos com especificações variadas.

Caso uma ou mais mãos sejam detectadas, as posições dos pontos de referência são armazenadas e usadas para calcular parâmetros musicais, bem como para a identificação de gestos. A detecção precisa e em tempo real dos pontos de referência é essencial para garantir que os gestos sejam capturados de forma confiável, proporcionando uma interação fluida entre o usuário e o sistema.

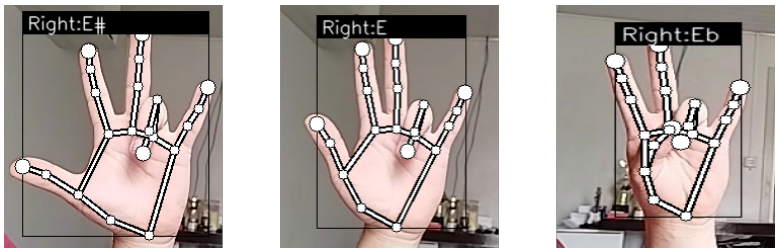
A detecção precisa e em tempo real dos pontos de referência é fundamental para garantir a responsividade e a confiabilidade do sistema. Qualquer atraso ou erro nessa etapa pode comprometer a interação do usuário com o sistema, prejudicando a experiência musical. Assim, a combinação da robustez do MediaPipe com a eficiência do pipeline de processamento garante uma captura de gestos fluida e confiável, proporcionando uma interação natural entre o usuário e o instrumento digital.

### 4.2.3 Identificação do Gesto

A identificação do gesto é uma etapa essencial do sistema, sendo responsável por interpretar a posição dos pontos de referência e convertê-la em comandos específicos que controlam o sintetizador musical. Para cada quadro capturado pela câmera, a imagem é inicialmente convertida do formato BGR para RGB, conforme exigido pelo modelo do MediaPipe, essa conversão é necessária para garantir a compatibilidade com o modelo pré-treinado, que realiza a detecção dos pontos de referência em cada mão.

Após a detecção dos pontos de referência, o sistema diferencia as mãos em esquerda e direita utilizando os classificadores do modelo do MediaPipe, essa distinção é fundamental para atribuir funções específicas a cada mão no contexto do sistema. A mão direita é encarregada de reconhecer os gestos previamente treinados pelo modelo de aprendizado de máquina, os quais são utilizados para tocar notas musicais. Além disso, a mão direita também determina se a nota tocada será natural, sustenido ou bemol, essa definição é feita com base na angulação da ponta do polegar (Ponto de número 4) em relação à a base do polegar (Ponto de número 2), conforme ilustrado na Figura 10.

Figura 10 – Diferença de angulação do Dedão entre as notas




**Fonte:** Próprio Autor

A mão esquerda desempenha a função de controlar a frequência das notas que serão tocadas, permitindo que estas sejam ajustadas para tons mais graves ou mais agudos. Essa modificação ocorre com base na quantidade de dedos levantados. Por exemplo:

- Um dedo levantado (independentemente de qual seja) indica que as notas serão

tocadas na primeira oitava.

- Dois dedos levantados correspondem às notas da segunda oitava. Assim por diante, até o limite das cinco disponíveis.
- Nenhum dedo levantado, em contrapartida, instrui o sistema a tocar as notas padrão, que são as notas da quarta oitava (conforme descrito na Tabela 1).

### 4.2.4 Síntese da Nota

A síntese de notas no sistema proposto é baseada na utilização de amostras de áudio pré-gravadas de notas de piano, armazenadas no formato WAV (Waveform Audio File Format). Cada arquivo de áudio representa uma nota específica do piano, abrangendo as notas das cinco primeiras oitavas, desde C1 até B4#, sendo esta última equivalente a C6. Essa abordagem foi escolhida devido à sua capacidade de capturar as características acústicas autênticas do piano, proporcionando um timbre natural e realista em comparação a métodos puramente sintéticos, como a geração por ondas senoidais.

Durante a inicialização do sistema, todas as amostras são carregadas na memória para garantir uma reprodução em tempo real sem atrasos, para otimizar a performance e uniformizar o processamento, as amostras originalmente em estéreo são convertidas para mono, reduzindo a complexidade computacional e o uso de memória. Além disso, os dados de cada arquivo são convertidos para vetores de valores em ponto flutuante, permitindo uma manipulação eficiente durante a reprodução. Quando uma nota é ativada pelo sistema, a reprodução do áudio correspondente é iniciada. O processo é gerenciado por uma função que processa pequenos blocos de áudio (audio buffers) de forma contínua, isso permite a síntese em tempo real e a manipulação de múltiplas notas simultaneamente, caso necessário.

Para garantir a responsividade do sistema e evitar interferências no fluxo principal de reconhecimento de gestos, todo o processamento de áudio ocorre em uma thread separada. Embora o Python utilize o Global Interpreter Lock (GIL), que impede a execução simultânea de múltiplas threads em um único núcleo, o uso de threads ainda é altamente eficaz para lidar com tarefas de entrada e saída (I/O), como leitura e reprodução de áudio (FOUNDATION, 2023). Essa separação permite que o sistema mantenha a fluidez na captura de vídeo e na interpretação de gestos, enquanto gerencia a reprodução das notas de forma independente.

## 5 Resultados Experimentais

O objetivo principal desta seção é demonstrar a funcionalidade prática do sistema proposto. Essa demonstração busca evidenciar a capacidade do sistema em traduzir gestos manuais em sons musicais em tempo real, avaliando qualitativamente parâmetros como responsividade, precisão na execução musical e usabilidade geral. A análise dos resultados também considera aspectos técnicos, como o impacto do hardware na performance do sistema.

Os experimentos foram conduzidos em dois ambientes computacionais distintos para observar a influência das configurações de hardware na performance do sistema. O primeiro ambiente utilizou um processador Intel® Core™ i5-2540M, com 2 núcleos e 4 threads, acompanhado por 8 GB de memória RAM. Já o segundo ambiente contou com uma CPU Intel® Core™ i7-3770K de 3.50 GHz, com 4 núcleos e 8 threads, além de uma placa de vídeo dedicada NVIDIA GeForce GTX 750 Ti e 16 GB de memória RAM.

Para validar o sistema, foram realizadas duas demonstrações práticas em um ambiente controlado, com condições de iluminação adequadas para garantir a precisão na detecção dos gestos. As composições escolhidas para os testes foram:

- **Asa Branca** – Luiz Gonzaga: Um clássico do baião, escolhido por sua melodia característica e simplicidade, facilitando a demonstração dos gestos e das transições entre notas.
- **Sinfonia n.º 9 (Ode à Alegria)** – Ludwig van Beethoven: Essa composição apresenta maior complexidade em relação à anterior, devido ao maior número de notas e padrões de repetição mais elaborados.

Essas músicas foram selecionadas estrategicamente, pois estão dentro do escopo de notas que o sistema é capaz de reproduzir e possuem um caráter didático com uma progressão de notas simples, o que facilita a avaliação das principais funcionalidades do projeto.

Antes de executar as composições, foi gravado um vídeo demonstrando as principais funcionalidades do sistema (youtu.be/kvk15qE4WdY). Esse vídeo apresenta as mudanças entre oitavas, reprodução de notas musicais e a transição entre notas naturais, sustenidos e bemóis.

Nos primeiros testes realizados no ambiente com o processador Intel® Core™ i5, foi observada uma queda significativa de quadros por segundo (FPS) ao detectar as mãos. Esse comportamento gerou atrasos perceptíveis entre a execução do gesto e a reprodução do som da nota, dificultando a execução precisa das melodias. Esses atrasos são particularmente visíveis nos vídeos de Asa Branca (youtu.be/Ln-3GnhC-CM) e Ode à Alegria (youtu.be/FcEVWKdNKxo), onde a responsividade limitada prejudicou a fluidez e a fidelidade das músicas às composições originais.

Os testes realizados no segundo ambiente, que possui um hardware mais robusto, demonstraram uma melhora significativa na performance do sistema. Nos vídeos de Asa Branca (youtu.be/dVb43Hgwk2s) e Ode à Alegria (youtu.be/mOh1aKtrUl0), o sistema apresentou maior fluidez, com atrasos minimizados entre a execução do gesto e a reprodução do som. Essa melhora permitiu que as melodias fossem executadas de forma mais precisa e com maior semelhança às composições originais. A melhora na performance é atribuída à presença de uma placa de vídeo dedicada, que alivia a carga do processamento gráfico, e a um processador mais potente com maior quantidade de núcleos. Além disso, um processador mais potente e a disponibilidade de mais memória RAM no segundo ambiente contribuíram para um processamento mais eficiente, reduzindo gargalos durante a execução.

## 6 Considerações Finais

A constante intersecção entre a tecnologia e a música abre muitas possibilidades para explorar a criatividade e criar novos meios de performance e expressão . Este trabalho apresentou o desenvolvimento de uma interface musical baseada no reconhecimento de gestos em tempo real, integrando tecnologias de visão computacional e aprendizado de máquina. O sistema demonstrou sua capacidade de traduzir gestos manuais em notas musicais, validando sua proposta como uma ferramenta interativa e inovadora para performances musicais.

A demonstração prática destacou as funcionalidades do sistema, evidenciando a fluidez e precisão na conversão de gestos em sons, além de ressaltar áreas com potencial para melhorias futuras. Entre os aspectos positivos, o uso de amostras reais de áudio garantiu um timbre autêntico e natural, enquanto a combinação de tecnologias permitiu uma experiência interativa promissora para músicos de diferentes níveis.

Embora os resultados sejam satisfatórios, algumas limitações foram identificadas durante o desenvolvimento. O sistema não foi submetido a testes sistemáticos para avaliação quantitativa de desempenho, o que limita uma análise mais detalhada sobre sua precisão e robustez em diferentes condições de uso. Além disso, o uso de gestos estáticos restringe a complexidade das interações possíveis, o que pode ser um ponto de melhoria em versões futuras. Integração com softwares de música consolidados no mercado como Ableton, FL Studio entre outros também pode ser uma ótima alternativa para expandir os horizontes da aplicação e utilizar de uma grande quantidade de funções que estes programas trazem consigo. Melhorias de performance também são essenciais, pois embora o sistema seja funcional mesmo em configurações de hardware mais modestas, a responsividade e a precisão são significativamente impactadas pelo desempenho do equipamento. Ambientes com maior capacidade de processamento e recursos gráficos dedicados permitem uma experiência muito mais fluida e próxima da execução musical em tempo real.

Como trabalhos futuros, propõe-se a realização de testes sistemáticos com diferentes tipos de usuários, incluindo músicos experientes e pessoas sem conhecimento musical, para avaliar a precisão, a responsividade e a usabilidade do sistema em cenários variados. Esse tipo de avaliação pode fornecer dados quantitativos mais robustos, contribuindo para uma análise mais detalhada do desempenho do sistema. Outra linha de evolução é o desenvolvimento de modelos de aprendizado de máquina capazes de reconhecer gestos dinâmicos, ampliando significativamente as possibilidades de interação e expressão musical além dos gestos estáticos atualmente suportados. Além disso, a integração com softwares de produção musical amplamente utilizados, como Ableton Live e FL Studio, pode expandir as funcionalidades do sistema, permitindo que ele seja utilizado em um ambiente profissional e aproveite os recursos avançados dessas plataformas. Outro aspecto a ser explorado é a adaptação do sistema para dispositivos móveis, como smartphones e tablets, o que exigirá otimizações específicas para garantir uma performance fluida mesmo em hardware com recursos limitados, essa expansão aumentaria consideravelmente a acessibilidade e o alcance do sistema.

## Referências


ABADI, M. et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. Software available from tensorflow.org. Disponível em: <https://www.tensorflow.org/>. Citado (3) vezes nas páginas [17, 18 e 19].

BEGUM, H. et al. Digital beethoven—an android based virtual piano. In: IEEE. *2017 13th International Conference on Emerging Technologies (ICET)*. [S.l.], 2017. p. 1–5. Citado na página [14].

BUSCHERT, J. Musician maker: Play expressive music without practice. In: *NIME*. [S.l.: s.n.], 2012. Citado na página [7].

CHANG, H.-H. et al. Real-time virtual instruments based on neural network system. In: IEEE. *2015 8th International Conference on Ubi-Media Computing (UMEDIA)*. [S.l.], 2015. p. 163–167. Citado na página [13].

CHOLLET, F. et al. *Adam Optimizer*. [S.l.], 2015. Acessado em: 2024-11-11. Disponível em: <https://keras.io/api/optimizers/adam/>. Citado na página [19].

DIETTERICH, T. Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, v. 27, n. 3, p. 326–327, 1995. Citado na página [19].

FEEZELL, M. Music theory fundamentals. *Music Theory Fundamentals*, p. 1–46, 2011. Citado (2) vezes nas páginas [8 e 9].

FOUNDATION, P. S. *Python Documentation*. [S.l.], 2023. Accessed: 2024-11-22. Disponível em: <https://docs.python.org/3/library/threading.html>. Citado na página [23].

GARDNER, M. W.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, Elsevier, v. 32, n. 14-15, p. 2627–2636, 1998. Citado na página [12].

GERON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd. ed. [S.l.]: O'Reilly Media, Inc., 2019. ISBN 1492032646. Citado na página [11].

GUZSVINECZ, T.; SZUCS, V.; SIK-LANYI, C. Suitability of the kinect sensor and leap motion controller—a literature review. *Sensors*, MDPI, v. 19, n. 5, p. 1072, 2019. Citado na página [10].

HONGWEI, W. et al. The optical flow method research of particle image velocimetry. *Procedia Engineering*, v. 99, p. 918–924, 2015. ISSN 1877-7058. 2014 Asia-Pacific International Symposium on Aerospace Technology, APISAT2014 September 24-26, 2014 Shanghai, China. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877705814037394>. Citado na página [11].

HUGHES, A.; FIGUEROA, P. B. Everything is musical: Creating new instruments for musical expression and interaction with accessible open-source technology—the laser room and other devices. In: *Innovation in Music*. [S.l.]: Routledge, 2019. p. 358–367. Citado na página [7].

Iazzetta, Fernando. *Tabela de Frequências e Notas Musicais*. 2024. <https://iazzetta.eca.usp.br/tutor/acustica/introducao/tabela1.html>. Acessado em: 2024-11-12. Citado na página [9].

LOKHANDE, P.; PRAJAPATI, R.; PANSARE, S. Data gloves for sign language recognition system. *International Journal of Computer Applications*, Citeseer, v. 975, p. 8887, 2015. Citado na página [10].

MITCHELL, T. *Machine Learning*. [S.l.]: McGraw-Hill Education, 1997. Citado na página [11].

NIETO, O.; SHASHA, D. Hand gesture recognition in mobile devices: Enhancing the musical experience. In: *Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*. [S.l.: s.n.], 2013. p. 15–18. Citado na página [7].

POLINER, G. E. et al. Melody transcription from music audio: Approaches and evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 15, n. 4, p. 1247–1256, 2007. Citado na página [10].

POPESCU, M.-C. et al. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, World Scientific and Engineering Academy and Society (WSEAS) Stevens Point ..., v. 8, n. 7, p. 579–588, 2009. Citado (2) vezes nas páginas [11 e 12].

RUCK, D. W.; ROGERS, S. K.; KABRISKY, M. Feature selection using a multilayer perceptron. *Journal of neural network computing*, Citeseer, v. 2, n. 2, p. 40–48, 1990. Citado (2) vezes nas páginas [11 e 12].

SARI, N. W.; UTOMO, W. H. Hand gesture recognition for tone classification according to kodaly handsign using cnn. In: *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*. [S.l.: s.n.], 2023. p. 44–49. Citado na página [13].

SHANG, K.; WANG, Z. A music performance method based on visual gesture recognition. In: *2022 China Automation Congress (CAC)*. [S.l.: s.n.], 2022. p. 2624–2631. Citado (3) vezes nas páginas [7, 13 e 15].

TAKAHASHI, S. *hand-gesture-recognition-using-mediapipe*. 2020. If you use 'hand-gesture-recognition-using-mediapipe' in your research, please cite it using these metadata. Disponível em: <https://github.com/Kazuhito00/hand-gesture-recognition-using-mediapipe>. Citado (4) vezes nas páginas [15, 16, 17 e 18].

WESSEL, D. L. Timbre space as a musical control structure. *Computer music journal*, JSTOR, p. 45–52, 1979. Citado na página [9].

XING, Y.; LEI, H. Analysis and design of intelligent music player system based on gesture recognition. In: *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*. [S.l.: s.n.], 2022. p. 1807–1811. Citado na página [13].

YASEN, M.; JUSOH, S. A systematic review on hand gesture recognition techniques, challenges and applications. *PeerJ Computer Science*, PeerJ Inc., v. 5, p. e218, 2019. Citado na página [7].

YOST, W. A. Pitch perception. *Attention, Perception, & Psychophysics*, Springer, v. 71, n. 8, p. 1701–1715, 2009. Citado na página [9].

ZHANG, F. et al. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. Citado na página [11].

ZHANG, M.; ZHANG, J. A gesturally controlled virtual musical instruments for chinese guqin. In: IEEE. *2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*. [S.l.], 2020. p. 95–100. Citado na página [12].