

UNIVERSIDADE NOVE DE JULHO – UNINOVE

**PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA E GESTÃO DO
CONHECIMENTO**

MATHEUS MARTINS DORETTO

**RECURSOS EDUCACIONAIS BASEADOS EM VISÃO COMPUTACIONAL PARA
ANALISAR A RELAÇÃO ENTRE ATENÇÃO E DESEMPENHO ESCOLAR**

São Paulo

2026

MATHEUS MARTINS DORETTO

**RECURSOS EDUCACIONAIS BASEADOS EM VISÃO COMPUTACIONAL PARA
ANALISAR A RELAÇÃO ENTRE ATENÇÃO E DESEMPENHO ESCOLAR**

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Informática e Gestão do Conhecimento (PPGI) da Universidade Nove de Julho UNINOVE, como requisito parcial para a obtenção do grau de Mestre em Informática e Gestão do Conhecimento.

Prof. Dr. Luiz Fernando Rodrigues Pinto – Orientador.
Prof. Dr. Sidnei Alves de Araújo – Coorientador.

**São Paulo
2026**

Doretto, Matheus Martins.

Recursos educacionais baseados em visão computacional para analisar a relação entre atenção e desempenho escolar. / Matheus Martins Doretto. 2026.

80 f.

Dissertação (Mestrado)- Universidade Nove de Julho - UNINOVE, São Paulo, 2026.

Orientador (a): Prof. Dr. Luiz Fernando Rodrigues Pinto.

1. Visão computacional. 2. Atenção escolar. 3. Inteligência artificial. 4. Desempenho escolar. 5. YOLO.

I. Pinto, Luiz Fernando Rodrigues. II. Título.

CDU 004

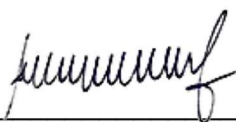
MATHEUS MARTINS DORETTO

**RECURSOS EDUCACIONAIS BASEADOS EM VISÃO COMPUTACIONAL PARA
ANALISAR A RELAÇÃO ENTRE ATENÇÃO E DESEMPENHO ESCOLAR**

Dissertação de mestrado
apresentada ao Programa de Pós-
graduação em Informática e Gestão do
Conhecimento (PPGI) da Universidade
Nove de Julho UNINOVE, como requisito
parcial para a obtenção do grau de Mestre
em Informática e Gestão do
Conhecimento, pela Banca Examinadora,
formada por:



Prof. Dr. Luiz Fernando Rodrigues Pinto



Prof. Dr. Sidnei Alves de Araújo



Prof. Dr. Peterson Adriano Belan



Prof. Dr. Emilcy Juliana Hernández Leal

São Paulo

2026

Dedico esta dissertação à minha família, pelo amor e apoio incondicionais, aos meus professores, que foram guias nessa caminhada do conhecimento, e aos colegas de classe, que tornaram essa jornada mais significativa.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, por ter me sustentado em todos os momentos, especialmente nos dias mais difíceis, me concedendo forças, sabedoria e serenidade para seguir em frente.

À minha família, com todo o meu amor e gratidão, em especial aos meus pais, Gislaine e João, que sempre foram meu alicerce. Obrigado por cada palavra de incentivo, por cada gesto de carinho, pelo apoio incondicional e por acreditarem em mim mesmo quando eu duvidava. Sem vocês, eu jamais teria chegado até aqui.

Aos professores do Programa de Pós-Graduação da UNINOVE, minha sincera gratidão. A atenção, o comprometimento e a generosidade com que compartilharam seus conhecimentos foram fundamentais para o amadurecimento do meu trabalho e para minha formação como pesquisador. Em especial, agradeço ao meu orientador Luiz, pela orientação, pelas críticas construtivas e por me impulsionar a ir além.

Aos amigos e a todas as pessoas que, de alguma forma, estiveram ao meu lado durante essa jornada: obrigado por me ouvirem, me motivarem e celebrarem comigo cada pequena conquista.

A todos vocês, o meu muito obrigado.

RESUMO

O monitoramento da atenção em sala de aula representa um desafio recorrente no contexto educacional, especialmente porque métodos tradicionais, como a observação direta, tendem a ser subjetivos e pouco adequados para acompanhar sistematicamente as variações atencionais dos alunos. Nesse contexto, a visão computacional apresenta-se como alternativa promissora para a extração automatizada de informações visuais associadas ao comportamento discente. A partir da revisão da literatura, identificou-se a escassez de estudos que relacionam diretamente a atenção mensurada por visão computacional e o desempenho escolar. Dessa forma, o objetivo desta pesquisa foi desenvolver e aplicar um método computacional para correlacionar indicadores de atenção identificados em imagens de sala de aula com o desempenho escolar dos alunos. Para isso, adotou-se uma abordagem mista, de natureza aplicada, exploratória, correlacional e documental, com base em dados cedidos por uma instituição privada de ensino da cidade de São Paulo. As imagens das aulas foram processadas com os modelos YOLOv11 e YOLO Face, possibilitando a detecção de pessoas e faces e a extração dos dados utilizados na heurística de classificação dos níveis de atenção. Posteriormente, as frequências desses níveis foram consolidadas e convertidas em notas de atenção de 0 a 10, organizadas com as notas de desempenho escolar, formando pares de dados para análises descritivas e correlacionais. Os resultados evidenciaram maior estabilidade nas notas de atenção em comparação às notas de desempenho, além de associação positiva entre as variáveis. O coeficiente de Pearson foi de $r = 0,285$, com $p = 0,127$, indicando correlação positiva fraca e não significativa. Já o coeficiente de Spearman foi de $\rho = 0,375$, com $p = 0,041$, indicando correlação positiva moderada e estatisticamente significativa. Conclui-se que a abordagem proposta demonstrou viabilidade para transformar registros de imagem em indicadores quantitativos de atenção e correlacioná-los com dados escolares. Como contribuição, o estudo apresenta um caminho metodológico aplicável à análise educacional, embora limitado pelo recorte da amostra e pela heurística adotada, indicando a necessidade de futuras pesquisas com bases ampliadas e novos indicadores visuais.

Palavras-chave: Visão computacional. Atenção escolar. Inteligência artificial. Desempenho escolar. YOLO.

ABSTRACT

Monitoring attention in the classroom represents a recurring challenge in the educational context, especially because traditional methods, such as direct observation, tend to be subjective and poorly suited to systematically tracking students' attentional variations. In this context, computer vision emerges as a promising alternative for the automated extraction of visual information associated with student behavior. Based on the literature review, a scarcity of studies was identified that directly relate attention measured through computer vision to school performance. Therefore, the objective of this research was to develop and apply a computational method to correlate attention indicators identified in classroom images with students' school performance. To this end, a mixed-method approach was adopted, with an applied, exploratory, correlational, and documentary nature, based on data provided by a private educational institution in the city of São Paulo. Classroom images were processed using the YOLOv11 and YOLO Face models, enabling the detection of people and faces and the extraction of data used in the heuristic classification of attention levels. Subsequently, the frequencies of these levels were consolidated and converted into attention scores on a scale from 0 to 10 and organized with school performance scores, forming data pairs for descriptive and correlational analyses. The results showed greater stability in attention scores compared to performance scores, as well as a positive association between the variables. The Pearson coefficient was $r = 0.285$, with $p = 0.127$, indicating a weak and non-significant positive correlation. The Spearman coefficient was $\rho = 0.375$, with $p = 0.041$, indicating a moderate and statistically significant positive correlation. It is concluded that the proposed approach demonstrated feasibility in transforming image records into quantitative attention indicators and correlating them with school data. As a contribution, the study presents a methodological path applicable to educational analysis, although limited by the sample scope and the adopted heuristic, indicating the need for future research with expanded datasets and new visual indicators.

Keywords: Computer vision. Classroom attention. Artificial intelligence. School performance. YOLO.

LISTA DE ILUSTRAÇÕES

FIGURAS

Figura 1 - Representação de uma imagem digital (A) como matriz de pixels (B) e função bidimensional (C).....	22
Figura 2 - Etapas de um sistema de visão computacional	23
Figura 3 - Modelo lógico de um neurônio artificial	26
Figura 4 - Estrutura básica de uma rede neural artificial alimentada adiante	27
Figura 5 - Estrutura simplificada de uma rede neural convolucional aplicada à interpretação visual	30
Figura 6 - Estrutura simplificada do processamento realizado pelo modelo YOLO aplicado a imagens de sala de aula	32
Figura 7 - Elementos que compõem a saída de uma detecção do modelo YOLO....	33
Figura 8 - Fluxograma do projeto de pesquisa	42
Figura 9 - Fluxograma das etapas de aplicação do PRISMA	46
Figura 10 - Fluxo do método proposto para estruturação dos dados	48
Figura 11 - Exemplos de detecção de pessoas e faces pelos modelos YOLOv11 e YOLO Face, com caixas delimitadoras e respectivos valores de confiança.....	50
Figura 12 - Exemplos de classificação dos níveis de atenção	51
Figura 13 - Exemplo de detecção de pessoas e faces em imagem de sala de aula ..	60
Figura 14 - Exemplo de atribuição dos níveis de atenção em imagem processada ..	60
Figura 15 - Variação das notas de atenção e de desempenho escolar.....	64

QUADROS

Quadro 1 - Síntese dos resultados da RSL	39
Quadro 2 - Termos utilizados para compor os constructos	44
Quadro 3 - Sentenças de busca para as consultas	45
Quadro 4 - Critérios de inclusão e exclusão de trabalhos na pesquisa.....	45
Quadro 5 - Critérios atribuídos aos níveis de atenção	51
Quadro 6 - Classificação da intensidade das correlações.....	55

LISTA DE TABELAS

Tabela 1 - Artigos selecionados	35
Tabela 2 - Caracterização da base de dados analisada	59
Tabela 3 - Amostra dos indicadores extraídos por imagem para cálculo da nota de atenção	61
Tabela 4 - Registros consolidados de desempenho escolar e atenção	62
Tabela 5 - Estatísticas descritivas das notas de atenção e do desempenho escolar	64
Tabela 6 - Resultados da correlação entre atenção e desempenho escolar.....	65

SUMÁRIO

1	INTRODUÇÃO	13
1.1	CONTEXTUALIZAÇÃO DO TEMA.....	13
1.2	TRABALHOS CORRELATOS E QUESTÃO DE PESQUISA	14
1.3	OBJETIVO GERAL E OBJETIVOS ESPECÍFICOS	16
1.4	JUSTIFICATIVA DE PESQUISA	16
1.5	DELIMITAÇÃO DA PESQUISA	17
2	FUNDAMENTAÇÃO TEÓRICA.....	19
2.1	EDUCAÇÃO E O PROCESSO DE APRENDIZAGEM	19
2.1.1	Estrutura do Ensino Fundamental no Brasil.....	19
2.1.2	A Atenção como Elemento Central do Processo de Aprendizagem	20
2.2	PROCESSAMENTO DE IMAGENS DIGITAIS E VISÃO COMPUTACIONAL.....	21
2.3	INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA.....	24
2.3.1	Inteligência Artificial Aplicada à Interpretação Visual.....	29
2.3.2	Modelo YOLO para Detecção Visual	31
2.4	REVISÃO DA LITERATURA SOBRE VISÃO COMPUTACIONAL EM ANÁLISE DE ATENÇÃO EM SALA DE AULA	35
3	MATERIAIS E MÉTODOS	42
3.1	TIPOLOGIA DA PESQUISA	43
3.2	MÉTODO PARA A REVISÃO DA LITERATURA.....	44
3.3	DADOS EMPREGADOS NA PESQUISA.....	46
3.4	TRATAMENTO DE DADOS	48
3.4.1	Estruturação e Processamento dos Dados.....	49
3.4.2	Abordagens Analíticas	53
3.4.3	Outras Abordagens Avaliadas	56

4	RESULTADOS	58
4.1	CARACTERIZAÇÃO DOS DADOS ANALISADOS	58
4.2	CONSOLIDAÇÃO DOS REGISTROS DE ATENÇÃO E DESEMPENHO	59
4.3	ANÁLISE DESCRITIVA DAS NOTAS DE ATENÇÃO E DE DESEMPENHO ESCOLAR	63
4.4	ANÁLISE CORRELACIONAL ENTRE AS NOTAS DE ATENÇÃO E DE DESEMPENHO ESCOLAR.....	64
4.5	DISCUSSÃO	66
5	CONCLUSÃO	69
	REFERÊNCIAS.....	71
	APÊNDICE A – CÓDIGO USADO NO PROCESSAMENTO DAS IMAGENS	75
	APÊNDICE B – TERMO DE CONFIDENCIALIDADE ASSINADO	78

1 INTRODUÇÃO

Este capítulo apresenta a contextualização do tema, os trabalhos correlatos, a identificação das lacunas e da questão de pesquisa, o objetivo geral e os objetivos específicos, a justificativa da pesquisa e a delimitação da pesquisa.

1.1 CONTEXTUALIZAÇÃO DO TEMA

O ambiente de sala de aula é tradicionalmente caracterizado pela interação direta entre alunos e professores, e uma das tarefas rotineiras mais comuns é o monitoramento da atenção dos alunos. Esse processo, embora necessário, consome tempo valioso que poderia ser dedicado ao ensino, além de não garantir que os alunos estejam realmente engajados ou atentos às atividades (Liu, Jiang e Jiang, 2025).

O monitoramento da atenção comumente é responsabilidade dos professores, os quais determinam os níveis de atenção dos alunos com base no comportamento em sala de aula. No entanto, esse método impõe altas exigências aos docentes e é limitado por suas percepções subjetivas, o que pode levar a problemas como sobrecarga de trabalho, avaliações imprecisas e ineficiências (Schwartz, Schmitt e Lose, 2012). Portanto, métodos tradicionais de observação direta demandam tempo e não oferecem dados precisos sobre a atenção dos estudantes (Gottipati, Shankararaman e Jie, 2020).

Para superar os problemas enfrentados pelos métodos tradicionais de avaliação da atenção, alguns pesquisadores começaram a utilizar técnicas de inteligência artificial (IA) na área de visão computacional (VC) para monitorar a atenção em sala de aula presencial ou online. Nos últimos anos, com a rápida evolução da IA, o reconhecimento do comportamento dos alunos em sala de aula baseado em redes neurais profundas tornou-se um tema de pesquisa em destaque. Esse processo geralmente envolve duas etapas: a detecção individual dos alunos no vídeo e a classificação dos diferentes estados desses alunos (Zhang, Q. et al., 2024). A pesquisa de Wang et al. (2024), por exemplo, apresentou um modelo eficiente de detecção de alvos em tempo real, utilizando a velocidade e a precisão do modelo You Only Look Once (YOLO) para identificar os alunos em um vídeo, estabelecendo a base para análises comportamentais.

Outro contexto a ser considerado é o período pós-pandemia, no qual a expansão das aulas on-line e das transmissões educacionais evidenciou a necessidade de monitoramento da concentração dos alunos em ambientes virtuais. Em resposta a essa demanda, foram desenvolvidos projetos voltados à análise de comportamentos não verbais por meio de técnicas de processamento de imagens (Uçar e Özdemir, 2022). Nesse contexto, Wang et al. (2023) apresentam um sistema que utiliza estruturas piramidais para processar mapas de características em múltiplas escalas e discutem os desafios dos métodos baseados em redes neurais convolucionais em ambientes com múltiplos alvos e escalas variadas.

Diante desse cenário, observa-se que a literatura aponta o uso de técnicas computacionais para o monitoramento da atenção e do comportamento de alunos em ambientes educacionais como um tema recente e em crescimento. Esses estudos mostram que a aplicação dessas tecnologias vem se consolidando de forma viável, o que reflete o avanço das soluções digitais voltadas à análise e à melhoria dos processos de ensino e aprendizagem.

1.2 TRABALHOS CORRELATOS E QUESTÃO DE PESQUISA

Diversos estudos recentes têm explorado o uso de tecnologias baseadas em visão computacional para a análise de comportamentos e atenção em ambientes educacionais. Nesta seção são apresentados alguns estudos destacando os objetivos e tecnologias envolvidas em cada um deles.

O trabalho de Huang, Cheng e Wu (2022) propõe um sistema para reconhecer posturas humanas em ambientes de educação *maker*, utilizando CNN com resultados promissores em termos de acurácia, o que valida a viabilidade técnica da detecção postural em ambientes educacionais dinâmicos. De modo semelhante, Lin et al. (2021) utilizam a extração de esqueletos corporais com o *OpenPose* e técnicas de correção para reconhecer múltiplos comportamentos estudantis, demonstrando robustez mesmo em ambientes com grande número de pessoas.

Redes neurais profundas são um tópico comum em diversas pesquisas. Os trabalhos apresentam a aplicação da tecnologia fazendo o uso de representações vetoriais das faces baseadas em atenção e emoção para prever o engajamento de alunos a partir de vídeos, explorando percepções de papéis dos alunos com inferência visual, adaptando técnicas de detecção para ambientes tradicionais com foco em

estados de atenção e algoritmos de detecção de pose de cabeça em alguns casos em tempo real (Sümer et al., 2021; Wei e Ma, 2024; Orosoo et al., 2024; Zhou et al., 2024; Shen e Sun, 2022).

Os desafios técnicos comuns enfrentados nesses estudos incluem oclusão, variações de escala e desequilíbrio de amostras. Para amenizar essas dificuldades, alguns trabalhos propõem abordagens baseadas em módulos de fusão de múltiplos níveis, que combinam informações extraídas de diferentes camadas da rede neural. Essa estratégia permite considerar tanto detalhes visuais simples, como bordas e contornos, quanto características mais complexas, como postura e comportamento dos estudantes, contribuindo para a detecção de comportamentos em salas inteligentes com alta precisão e baixo custo computacional. Além disso, estudos recentes exploram o uso de modelos como Deformable DETR, Swin Transformer e redes piramidais leves, destacando os modelos Transformer como alternativa promissora às CNNs convencionais (Wang et al., 2023; Wang et al., 2024).

Além disso, há exemplos do uso de detecção comportamental fora de disciplinas tradicionais, como por exemplo a identificação de comportamentos anormais em aulas de atletismo, ou no reconhecimento de comportamentos em campus universitários com destaque de ações incomuns (Zhang J. et al., 2024; Liu et al., 2023).

Uma revisão sistemática da literatura acerca do uso de visão computacional no contexto de reconhecimento do comportamento dos alunos em sala de aula aponta que os estudos existentes priorizam o reconhecimento de ações físicas e engajamento, mas carecem de conexões diretas com eventos instrucionais e conteúdos pedagógicos, além de não explorarem suficientemente a relação entre atenção visual e desempenho escolar (Liu, Jiang e Jiang, 2025).

Conforme o levantamento realizado, observa-se que, embora diversos estudos abordem o uso de visão computacional para reconhecer comportamentos, posturas, expressões e estados de engajamento em sala de aula, ainda são limitadas as pesquisas que relacionam diretamente indicadores visuais de atenção com dados de desempenho escolar. Essa lacuna justifica a presente investigação, que busca propor e aplicar um método computacional baseado em visão computacional para estimar indicadores de atenção em sala de aula e correlacioná-los com o desempenho escolar dos alunos. Neste sentido, este estudo pretende responder à seguinte questão: Como correlacionar a atenção em sala de aula, estimada por meio de um recurso

educacional baseado em visão computacional, com o desempenho escolar dos alunos?

1.3 OBJETIVO GERAL E OBJETIVOS ESPECÍFICOS

Nesse contexto, o objetivo geral desta dissertação é avaliar como um recurso educacional baseado em visão computacional pode ser utilizado para estimar a atenção em sala de aula e correlacioná-la com o desempenho escolar dos alunos.

Para alcançar este objetivo geral, foram definidos três objetivos específicos:

- Definir um conjunto de indicadores visuais de atenção viáveis de serem extraídos a partir de imagens de sala de aula por meio de modelos de detecção.
- Desenvolver e aplicar um método computacional baseado em visão computacional para detectar pessoas e faces, classificando os níveis de atenção para se obter notas de forma automatizada e normalizada.
- Consolidar os indicadores de atenção obtidos com os dados de desempenho escolar dos estudantes, de modo a analisar a associação entre as variáveis por meio de testes estatísticos correlacionais.

1.4 JUSTIFICATIVA DE PESQUISA

O cenário educacional brasileiro tem enfrentado desafios significativos no período pós-pandemia, sobretudo no que diz respeito ao desempenho escolar dos estudantes. Diversos estudos têm apontado um aumento nos índices de defasagem de aprendizagem, dificuldade de concentração e desengajamento dos alunos, especialmente no ensino fundamental, etapa crítica para o desenvolvimento das habilidades cognitivas básicas (Bartholo et al., 2022). Nesse contexto, torna-se interessante repensar práticas pedagógicas que contribuam para a melhoria do processo de ensino e aprendizagem.

Entre os desafios enfrentados, destaca-se a dificuldade de monitorar a atenção dos alunos em sala de aula, especialmente em turmas com muitos estudantes ou ambientes com recursos limitados. Os métodos tradicionais de avaliação do engajamento, como observação direta ou aplicação de questionários, embora úteis, apresentam limitações importantes: são subjetivos, dependem da percepção do

professor, não permitem o acompanhamento em tempo real e são de difícil aplicação sistemática (Sümer et al., 2021; Fredricks et al., 2004). Nesse sentido, o uso de tecnologias baseadas em visão computacional surge como uma alternativa promissora, permitindo a coleta automatizada de indicadores visuais de atenção, como orientação do olhar, postura corporal e expressões faciais, de forma objetiva e contínua (Bozkir et al., 2025; Adeika et al., 2024).

A implementação de sistemas inteligentes de monitoramento da atenção pode gerar impactos significativos na prática docente. Esses dados podem, por exemplo, indicar momentos em que os alunos apresentam queda de atenção, podendo futuramente apoiar ajustes na estratégia didática em tempo real. Entretanto, existem lacunas na literatura quanto à relação entre indicadores de atenção estimados por meio de visão computacional e o desempenho escolar dos alunos (Liu, Jiang e Jiang, 2025).

Do ponto de vista científico, esta pesquisa pode servir de base para o avanço do conhecimento na interface entre educação e ciência da computação ao propor e aplicar um método computacional baseado em visão computacional para estimar níveis de atenção e investigar sua associação com o desempenho escolar. Essa contribuição é relevante porque a literatura ainda apresenta número reduzido de estudos que relacionam indicadores visuais de atenção com dados escolares concretos. Também se destacam as contribuições sociais e práticas desta investigação, pois em um país marcado por desigualdades educacionais, tecnologias acessíveis e escaláveis como as baseadas em visão computacional podem democratizar o acesso a ferramentas de avaliação mais precisas, beneficiando tanto redes públicas quanto privadas (Bulathwela et al., 2021).

1.5 DELIMITAÇÃO DA PESQUISA

Esta pesquisa foi conduzida em uma escola de Ensino Fundamental II, com foco em turmas do 9º ano, abrangendo estudantes com faixa etária entre 14 e 15 anos, localizados em um contexto urbano no Brasil. A escolha por esse segmento da educação básica justifica-se por sua relevância na consolidação das competências cognitivas fundamentais previstas na Base Nacional Comum Curricular (BNCC), como leitura, escrita e raciocínio lógico, que servem de alicerce para os ciclos escolares subsequentes. Além disso, trata-se de uma etapa marcada por desafios crescentes

no que se refere ao engajamento e à atenção dos alunos, sobretudo em ambientes com turmas numerosas e heterogêneas (Brasil, 2017).

Do ponto de vista pedagógico, o Ensino Fundamental representa um período decisivo para o desenvolvimento da autonomia intelectual dos estudantes. Nessa fase, a atenção em sala de aula desempenha um papel central no processo de aprendizagem e formação pessoal, influenciando diretamente a capacidade dos alunos de absorver conteúdos e de participar ativamente das atividades propostas. Estudos apontam que a atenção, enquanto dimensão do engajamento medida da forma tradicional, apresenta forte correlação com o desempenho escolar e com a persistência escolar (Fredricks et al., 2004; Fonsêca et al., 2020). No entanto, professores frequentemente relatam dificuldades em monitorar de forma contínua e objetiva o nível de atenção dos alunos, o que evidencia a necessidade de soluções tecnológicas que viabilizem esse acompanhamento de maneira mais eficaz.

Nesse contexto, o uso de técnicas de visão computacional surge como uma alternativa promissora. Diferentemente dos métodos tradicionais, essa abordagem oferece maior escalabilidade, precisão e consistência na coleta de dados relacionados ao comportamento atencional (Liu, Jiang e Jiang, 2025; Adeika et al., 2024). Tais sistemas podem ser integrados ao cotidiano escolar sob a forma de recursos educacionais baseados em visão computacional, fornecendo apoio automatizado aos professores na identificação de padrões de atenção em sala de aula.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os fundamentos teóricos que sustentam a pesquisa, reunindo conceitos relacionados à atenção em sala de aula, ao desempenho escolar, ao processamento de imagens digitais, à visão computacional, à inteligência artificial e aos modelos de detecção visual. Inicialmente, discute-se a atenção no contexto educacional e sua relação com os processos de aprendizagem, destacando a importância de indicadores que auxiliem na compreensão do comportamento discente. Em seguida, são abordados os fundamentos computacionais das imagens digitais, incluindo sua representação matemática e as etapas gerais de um sistema de visão computacional. Na sequência, apresentam-se os conceitos de inteligência artificial, aprendizado de máquina, redes neurais artificiais e redes neurais convolucionais, enfatizando sua aplicação na extração automática de características visuais, com destaque para o modelo YOLO como técnica de detecção visual. Por fim, apresenta-se uma revisão da literatura sobre o uso da visão computacional no monitoramento da atenção em contexto escolar, com o objetivo de identificar abordagens, contribuições e lacunas que fundamentam esta dissertação.

2.1 EDUCAÇÃO E O PROCESSO DE APRENDIZAGEM

O entendimento do comportamento atencional em sala de aula exige, primeiramente, uma análise abrangente do contexto educacional em que o aluno está inserido, bem como dos fatores que influenciam a aprendizagem. Este tópico apresenta uma base conceitual sobre a estrutura do Ensino Fundamental no Brasil e seus desafios pedagógicos, abordando os mecanismos cognitivos da atenção e seu papel central no processo de aprendizagem.

2.1.1 Estrutura do Ensino Fundamental no Brasil

A educação básica no Brasil está estruturada em três etapas: a Educação Infantil, o Ensino Fundamental e o Ensino Médio. O Ensino Fundamental, por sua vez, possui duração de nove anos, do 1º ao 9º Ano, conforme estabelecido pela Lei de Diretrizes e Bases da Educação Nacional, Lei nº 9.394/1996 (Brasil, 1996). O foco desta pesquisa está especificamente no 9º ano do Ensino Fundamental, etapa que

marca a finalização da educação básica obrigatória antes do ingresso no Ensino Médio.

Segundo a BNCC, é no 9º ano que se consolida o domínio das competências fundamentais em áreas como leitura, escrita, raciocínio lógico e resolução de problemas. A BNCC reforça também a importância do desenvolvimento de competências socioemocionais, como persistência, responsabilidade e autorregulação, que estão diretamente ligadas à capacidade de manter a atenção durante o processo de aprendizagem (Brasil, 2017).

2.1.2 A Atenção como Elemento Central do Processo de Aprendizagem

A atenção dos alunos é um fator significativo para o sucesso estudantil e o desempenho acadêmico (Fredricks et al., 2004; Uçar e Özdemir, 2022). A participação em sala de aula é vista como evidência de aprendizado, promovendo habilidades técnicas como: pensamento crítico, escrita, valorização de diferenças culturais e gestão de tempo, além de habilidades interpessoais como a comunicação, escuta e fala (Gottipati, Shankararaman e Jie, 2020).

Embora a atenção seja um fenômeno complexo e influenciado por fatores cognitivos, emocionais, sociais e ambientais, sua manifestação em sala de aula pode ser parcialmente observada por meio de comportamentos visíveis, como postura corporal, orientação do olhar, participação nas atividades e direcionamento do aluno em relação à tarefa proposta.

A responsabilidade de controlar e manter a disciplina e o comportamento dos alunos em sala de aula recai em grande parte sobre os professores, o que aumenta consideravelmente sua carga de trabalho, além das tarefas normais de ensino (Emmer e Stough, 2003; Chen e Guan, 2022). A detecção dos estados de aprendizado dos alunos de maneira automatizada pode fornecer informações relevantes ao professor, permitindo que ele identifique alunos com dificuldades. Assim, o grau de atenção dos alunos pode apoiar os docentes a entregarem recursos de aprendizagem personalizados (Riad, Qbadou e Aoula, 2024).

Dessa forma, é possível considerar a atenção em sala de aula como uma dimensão relevante para a compreensão do desempenho escolar, embora esse desempenho seja influenciado por múltiplos fatores pedagógicos, cognitivos e contextuais.

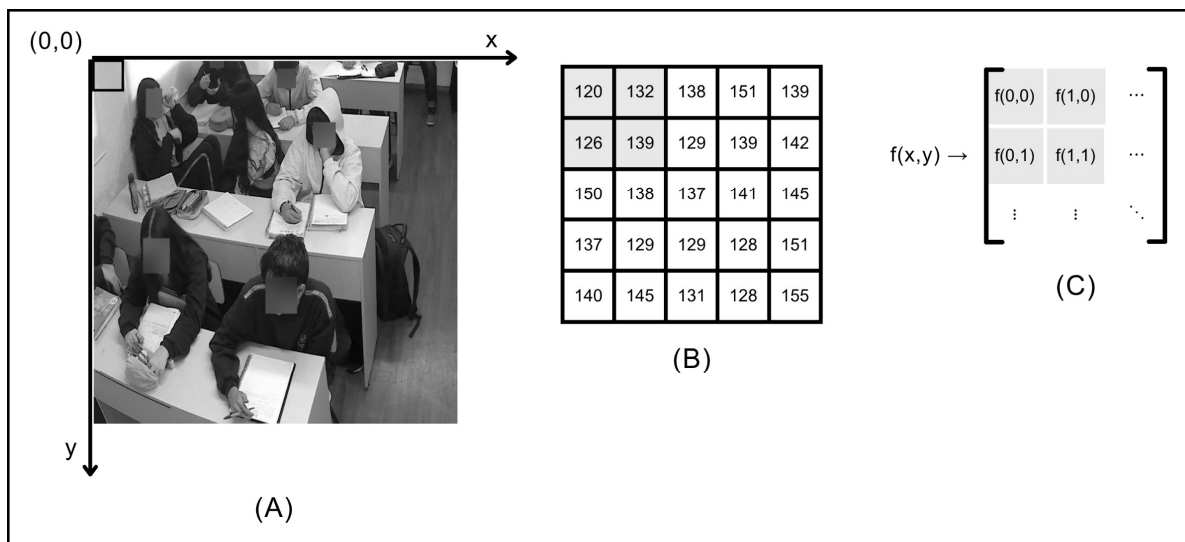
2.2 PROCESSAMENTO DE IMAGENS DIGITAIS E VISÃO COMPUTACIONAL

O processamento de imagens digitais é um campo da ciência da computação que se dedica à representação, manipulação e análise de imagens em formato digital por meio de modelos computacionais. Para que uma imagem possa ser processada por um sistema computacional, é necessário compreendê-la como uma estrutura numérica, formada por pixels organizados espacialmente. De forma geral, uma imagem pode ser representada como uma função bidimensional $f(x, y)$, em que (x) e (y) indicam as coordenadas espaciais de um ponto da imagem, enquanto a função representa a intensidade luminosa ou o nível de cinza associado àquele ponto.

No contexto digital, essa função é convertida em uma representação discreta por meio da digitalização, que envolve os processos de amostragem e quantização. A amostragem corresponde à discretização das coordenadas espaciais da imagem, enquanto a quantização corresponde à atribuição de valores discretos de intensidade a cada ponto amostrado. Dessa forma, uma imagem digital pode ser representada como uma matriz $(M \times N)$, na qual cada elemento corresponde a um pixel. Em imagens em escala de cinza, cada pixel armazena um valor de intensidade, em imagens coloridas, essa representação pode envolver múltiplos canais, como no padrão RGB, em que cada pixel é descrito por componentes de vermelho, verde e azul (Pedrini e Schwartz, 2007).

A Figura 1 ilustra essa representação a partir de uma imagem de sala de aula. Na subfigura (A), a imagem é apresentada em um sistema de coordenadas espaciais, no qual cada ponto da cena pode ser localizado por um par de coordenadas (x, y) . A região destacada nessa imagem representa uma pequena área composta por pixels. Na subfigura (B), essa região é ampliada e representada por uma matriz de valores numéricos, correspondentes aos níveis de intensidade em escala de cinza de cada pixel. Já a subfigura (C) apresenta a notação matemática, indicando que cada elemento da matriz pode ser descrito por uma função $f(x, y)$.

Figura 1 - Representação de uma imagem digital (A) como matriz de pixels (B) e função bidimensional (C)



Fonte: Elaborado pelo autor, com base em Pedrini e Schwartz (2007).

Essa representação é importante porque permite que imagens sejam utilizadas como entrada para algoritmos computacionais. No caso desta pesquisa, as imagens de sala de aula não são tratadas apenas como registros visuais, mas como dados digitais passíveis de processamento automatizado. Assim, a cena capturada pode ser convertida em informações numéricas que permitem a detecção de elementos visuais.

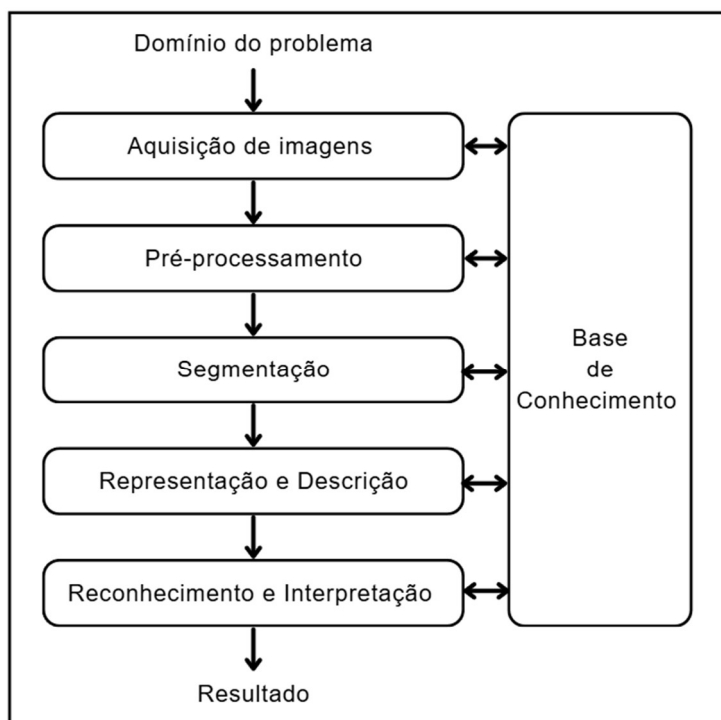
O desenvolvimento do processamento de imagens digitais está diretamente relacionado à evolução da computação. Seu surgimento remonta à década de 1920, com aplicações iniciais em procedimentos de impressão de imagens para melhora da qualidade. No entanto, foi a partir da década de 1960, com o avanço dos processadores, que os computadores se tornaram poderosos o suficiente para realizar operações relevantes no processamento de imagens, permitindo que esse campo se tornasse independente, com aplicações cada vez mais amplas, incluindo vigilância, medicina e indústria (Gonzalez e Woods, 2009; Jain, 1989).

Nesse contexto, o processamento de imagens digitais pode ser compreendido como uma base técnica importante para sistemas mais amplos de interpretação visual. Enquanto o processamento atua principalmente sobre a melhoria, transformação e extração inicial de informações presentes nos pixels, a visão computacional busca interpretar o conteúdo das imagens de forma automatizada, permitindo que sistemas computacionais identifiquem objetos, pessoas, padrões e relações presentes em uma cena.

A visão computacional é um campo que busca desenvolver sistemas capazes de interpretar estímulos visuais de maneira automatizada e confiável, com o mínimo de intervenção humana. Uma de suas tarefas centrais é a análise de imagens, cujo objetivo é descrever o conteúdo visual e distinguir características de interesse em uma cena (Szeliski, 2022). Dessa forma, a visão computacional amplia o processamento de imagens ao incorporar etapas de segmentação, reconhecimento de objetos, classificação e interpretação do conteúdo visual.

A Figura 2 sintetiza, de maneira simplificada, algumas etapas típicas de um sistema de visão computacional. O domínio do problema orienta a aquisição das imagens e influencia as etapas posteriores de pré-processamento, segmentação, representação e descrição, reconhecimento e interpretação. A base de conhecimento, por sua vez, relaciona-se com as diferentes fases do processo, pois reúne informações, parâmetros e critérios utilizados para orientar a análise da imagem e a interpretação dos resultados.

Figura 2 - Etapas de um sistema de visão computacional



Fonte: Pedrini e Schwartz (2007).

Na etapa de aquisição, a imagem é obtida a partir de um dispositivo de captura, como câmera, sensor ou scanner. Em seguida, o pré-processamento pode envolver

operações destinadas a melhorar a qualidade da imagem ou prepará-la para etapas posteriores, como ajustes de contraste, redução de ruído ou padronização de dimensões. A segmentação consiste na separação de regiões ou objetos de interesse presentes na cena. A etapa de representação e descrição busca organizar as informações extraídas da imagem, permitindo que características visuais sejam utilizadas em processos de reconhecimento.

A capacidade humana de interpretar grandes volumes de informação visual, com base em inferências e experiências prévias, motiva o desenvolvimento de técnicas computacionais que possam simular essa habilidade. A visão computacional, portanto, procura auxiliar na tomada de decisão baseada em imagens, replicando processos de percepção visual por meio de algoritmos. Esse campo pode ser dividido em dois níveis principais: o processamento de imagens, que compreende operações diretas sobre os pixels, como remoção de ruído, ajuste de contraste e detecção de bordas, e a análise de imagens, que inclui tarefas como segmentação, reconhecimento de objetos e interpretação do conteúdo visual. A análise computacional, nesse sentido, permite a representação de estruturas e padrões visuais que viabilizam a automatização da percepção, sendo um componente central na construção de sistemas capazes de interpretar elementos presentes em ambientes reais (Gonzalez e Woods, 2009; Marr, 2010).

Essa distinção é importante, pois o recurso educacional baseado em visão computacional utilizado nesta pesquisa depende tanto da representação digital quanto do processamento e interpretação. No entanto, a realização dessas tarefas em ambientes reais, marcados por variações de iluminação, posicionamento dos alunos, oclusões e diferentes escalas, exige modelos capazes de aprender padrões a partir dos dados. Por esse motivo, o avanço da visão computacional está fortemente associado ao desenvolvimento da inteligência artificial e do aprendizado de máquina, os quais são abordados na seção seguinte.

2.3 INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA

A inteligência artificial é um campo da ciência da computação voltado ao desenvolvimento de sistemas capazes de executar tarefas que, quando realizadas por seres humanos, exigiriam algum nível de inteligência, como raciocínio lógico, reconhecimento de padrões, aprendizado, tomada de decisão e resolução de

problemas. Segundo Russell e Norvig (2022), a IA pode ser compreendida como o estudo de agentes racionais, isto é, sistemas capazes de agir de forma a maximizar suas chances de alcançar determinados objetivos. O campo teve origem oficialmente em 1956, durante a Conferência de Dartmouth, considerada o marco inicial da área, a partir da proposta de que aspectos da inteligência humana poderiam ser simulados computacionalmente.

Com o avanço tecnológico e o aumento da capacidade computacional, a IA passou a abranger diversas subáreas, como o processamento de linguagem natural, a robótica, os sistemas especialistas, o aprendizado de máquina e a visão computacional. Entre essas abordagens, o aprendizado de máquina destaca-se por permitir que modelos computacionais identifiquem padrões em dados e ajustem seu desempenho sem depender exclusivamente de regras previamente programadas para cada situação (Mitchell, 1997).

Essa característica diferencia os algoritmos de aprendizado de máquina dos modelos tradicionais baseados em instruções fixas, pois o comportamento do sistema passa a ser ajustado a partir de exemplos. Assim, em tarefas de visão computacional, o modelo pode aprender relações e regularidades presentes nas imagens, mesmo diante de variações de iluminação, ângulo de captura, escala, posicionamento dos objetos e composição da cena, fatores que tornam difícil a definição manual de regras gerais.

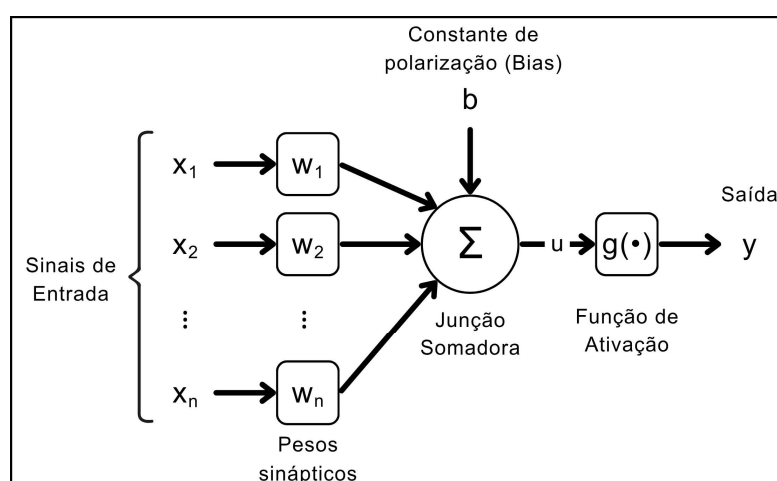
O aprendizado de máquina pode ocorrer de diferentes formas, sendo o aprendizado supervisionado uma das abordagens mais utilizadas em aplicações relacionadas à interpretação visual. Nesse tipo de aprendizado, o modelo é treinado a partir de exemplos previamente rotulados, nos quais as entradas e as saídas esperadas são conhecidas. Assim, o algoritmo aprende a relacionar padrões presentes nos dados de entrada com respostas desejadas, ajustando seus parâmetros internos para reduzir o erro entre a saída prevista e a saída correta (Mitchell, 1997). Em tarefas de análise de imagens, por exemplo, as entradas podem ser imagens digitais ou regiões dessas imagens, enquanto as saídas podem corresponder a classes de objetos, posições de elementos na cena ou índices de confiança associados às detecções.

De modo geral, o processo de aprendizagem envolve duas etapas principais: treinamento e inferência. Durante o treinamento, o modelo recebe exemplos, compara suas previsões com as respostas esperadas e ajusta seus parâmetros internos com o

objetivo de melhorar o desempenho. Após essa etapa, o modelo treinado pode ser aplicado a novos dados, em um processo denominado inferência, no qual passa a produzir respostas para entradas que não fizeram parte do treinamento. Essa distinção é importante para compreender os modelos de detecção visual adotados nesta pesquisa, uma vez que a análise das imagens não depende da definição manual de regras para cada cena, mas da aplicação de parâmetros previamente aprendidos por modelos treinados.

Os primeiros modelos formais de neurônios artificiais remontam ao trabalho de McCulloch e Pitts (1943), que propuseram uma representação lógico-matemática simplificada da atividade neural. A partir dessa base, as redes neurais artificiais passaram a ser desenvolvidas como modelos computacionais inspirados, de forma simplificada, no funcionamento do sistema nervoso biológico. Essas redes são compostas por unidades de processamento denominadas neurônios artificiais, organizadas em camadas e conectadas por pesos ajustáveis. Cada neurônio recebe sinais de entrada, associa um peso a cada sinal, realiza uma combinação ponderada desses sinais e aplica uma função de ativação para produzir uma saída (Haykin, 2009). A Figura 3 apresenta o modelo lógico simplificado de um neurônio artificial, evidenciando os sinais de entrada, os pesos sinápticos, o bias, a função de ativação e a saída produzida pelo neurônio.

Figura 3 - Modelo lógico de um neurônio artificial



Fonte: elaborado pelo autor, com base em Haykin (2009).

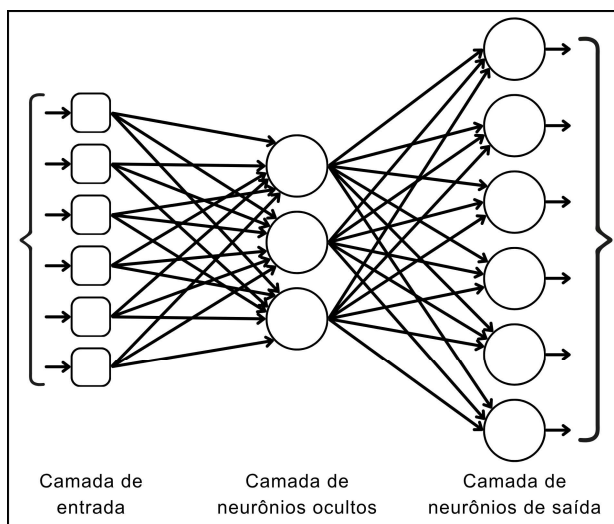
Na Figura 3, os valores $x_1, x_2, \dots, x_n$ representam os sinais de entrada do neurônio artificial, enquanto $w_1, w_2, \dots, w_n$ correspondem aos pesos sinápticos

associados a cada entrada e indicam a relevância de cada sinal no processo de combinação realizado pelo neurônio. A junção somadora combina as entradas ponderadas, considerando também a constante de polarização b , ou bias. O resultado desse processo é indicado por u , sendo posteriormente submetido à função de ativação $g(\cdot)$, responsável por produzir a saída y do neurônio.

A função de ativação desempenha um papel importante nesse processo, pois define como o sinal combinado será transformado em uma resposta. Em redes neurais artificiais, essa etapa permite que o modelo represente relações mais complexas entre entradas e saídas, ampliando sua capacidade de aprendizagem. Assim, o neurônio artificial pode ser compreendido como uma unidade básica de processamento que recebe informações, pondera sua relevância e produz uma resposta a ser utilizada pelas etapas seguintes da rede.

Quando diversos neurônios artificiais são organizados em conjunto, forma-se uma rede neural artificial. Em geral, essas redes possuem uma camada de entrada, uma ou mais camadas intermediárias, também chamadas de camadas ocultas, e uma camada de saída. A camada de entrada recebe os dados a serem analisados, as camadas ocultas realizam transformações sucessivas sobre esses dados, e a camada de saída apresenta a resposta do modelo, como uma classificação, uma probabilidade ou outro valor de interesse. A Figura 4 apresenta uma estrutura básica de uma rede neural artificial alimentada adiante, na qual os dados de entrada são propagados por uma camada de neurônios ocultos até a camada de saída.

Figura 4 - Estrutura básica de uma rede neural artificial alimentada adiante



Fonte: elaborado pelo autor, com base em Haykin (2009).

Na Figura 4, observa-se que o vetor de entrada é composto por valores que representam os dados fornecidos ao modelo. Esses dados são transmitidos aos neurônios da camada oculta por meio de conexões ponderadas, permitindo que a rede combine os sinais recebidos e produza novas representações internas. A camada de saída, por sua vez, organiza a resposta final do modelo, que pode assumir diferentes formas conforme o tipo de problema analisado. Em tarefas de classificação, por exemplo, a saída pode representar a probabilidade de determinada entrada pertencer a uma classe específica.

Durante o treinamento, os pesos das conexões entre os neurônios são ajustados de modo a reduzir a diferença entre a saída produzida pelo modelo e a saída esperada. Esse processo permite que a rede aprenda padrões presentes nos dados, melhorando gradualmente sua capacidade de generalização. Assim, a aprendizagem não ocorre pela programação manual de todas as regras possíveis, mas pelo ajuste progressivo dos parâmetros internos da rede a partir dos exemplos apresentados.

O avanço das redes neurais artificiais e o aumento da capacidade computacional possibilitaram o desenvolvimento de arquiteturas mais profundas, compostas por múltiplas camadas de processamento. Essas arquiteturas deram origem ao aprendizado profundo, ou *deep learning*, que permite a construção de modelos capazes de aprender progressivamente representações de diferentes níveis de abstração (Goodfellow, Bengio e Courville, 2016).

No caso das imagens digitais, essa abordagem é relevante porque, embora uma imagem possa ser representada como uma matriz de intensidades, ela não se resume a valores isolados. Há relações espaciais entre seus elementos, e essas relações formam bordas, texturas, regiões, faces, corpos e objetos. Por esse motivo, a interpretação visual exige modelos capazes de considerar não apenas os valores individuais da imagem, mas também os padrões formados por sua organização espacial.

Entre os modelos de aprendizado profundo aplicados a esse tipo de problema, destacam-se as redes neurais convolucionais, desenvolvidas especialmente para o processamento de dados com estrutura espacial, como imagens e vídeos. Essas arquiteturas constituem a base técnica de diversos modelos de detecção visual, incluindo o YOLO, discutido nas seções seguintes.

2.3.1 Inteligência Artificial Aplicada à Interpretação Visual

A aplicação da inteligência artificial em tarefas de interpretação visual tornou-se uma das áreas mais relevantes da visão computacional. A partir do avanço das redes neurais profundas, sistemas computacionais passaram a apresentar elevada capacidade de reconhecimento automático de padrões em imagens e vídeos, permitindo a identificação de objetos, pessoas, expressões faciais, movimentos e comportamentos de forma automatizada (Szeliski, 2022; Goodfellow, Bengio e Courville, 2016).

Nesse contexto, as CNNs, ou *Convolutional Neural Networks*, são especialmente relevantes por explorarem a organização espacial das imagens. Diferentemente de redes totalmente conectadas, nas quais todos os neurônios de uma camada podem se conectar aos neurônios da camada seguinte, as CNNs analisam regiões locais da imagem por meio de operações de convolução. Nessas operações, pequenos filtros, também chamados de *kernels*, percorrem a imagem ou os mapas de características, permitindo a extração de padrões visuais locais, como bordas, contrastes, texturas e formas elementares (Goodfellow, Bengio e Courville, 2016; Catarino, 2025).

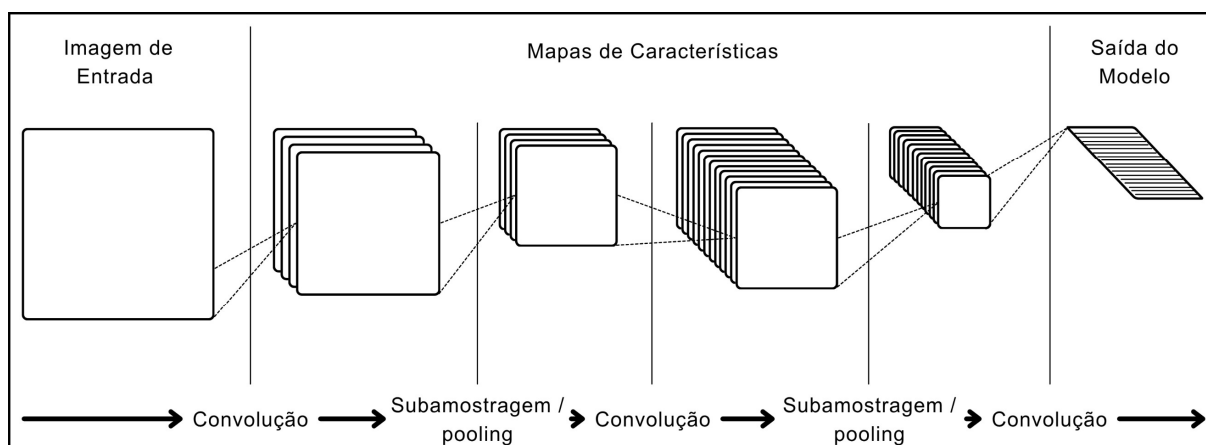
Em uma CNN, as camadas iniciais tendem a aprender características mais simples, como variações de intensidade, contornos e bordas. À medida que a informação avança pelas camadas seguintes, essas características são combinadas em representações mais complexas, como texturas, formas, partes de objetos, faces e corpos. Esse processo hierárquico permite que o modelo construa representações visuais progressivamente mais abstratas, o que torna possível realizar tarefas como classificação, detecção, segmentação, reconhecimento e rastreamento.

Além das camadas convolucionais, as CNNs podem utilizar etapas de subamostragem, também chamadas de *pooling*, que reduzem a dimensão espacial dos mapas de características. Essa redução contribui para diminuir o custo computacional e preservar informações relevantes extraídas pelas camadas anteriores. Assim, ao longo da rede, a imagem de entrada é transformada em sucessivos mapas de características, que passam a representar diferentes níveis de informação visual.

A Figura 5 apresenta, de forma simplificada, a estrutura de uma CNN aplicada ao processamento de imagens. A imagem de entrada passa inicialmente por

operações de convolução, responsáveis pela geração dos primeiros mapas de características. Em seguida, etapas de subamostragem reduzem a dimensão dessas representações, enquanto novas convoluções extraem características mais abstratas. Ao final, as informações extraídas ao longo da rede são condensadas na saída do modelo, que representa a resposta produzida a partir das características identificadas.

Figura 5 - Estrutura simplificada de uma rede neural convolucional aplicada à interpretação visual



Fonte: Elaborado pelo autor, com base em Haykin (2009).

Na Figura 5, os mapas de características representam as transformações sucessivas realizadas pela rede sobre a imagem de entrada. As linhas entre os blocos indicam que as saídas de uma etapa são utilizadas como entrada para a etapa seguinte, por meio de operações de convolução ou subamostragem. O bloco final, indicado como saída do modelo, representa uma forma compacta de resposta construída a partir das características extraídas ao longo da rede. Dependendo da tarefa, essa saída pode estar associada a uma classificação, a uma detecção ou a outro tipo de informação estruturada.

Nesse contexto, os modelos baseados em inteligência artificial são capazes de transformar informações presentes em imagens em dados estruturados, possibilitando análises quantitativas e reconhecimento automático de padrões. Essa capacidade tornou possível o desenvolvimento de aplicações relacionadas à segurança, medicina, direção autônoma, monitoramento industrial e reconhecimento comportamental (Marr, 2010).

Em aplicações educacionais, a inteligência artificial vem sendo utilizada para analisar comportamentos estudantis, monitorar níveis de engajamento e identificar

padrões associados à atenção dos alunos em sala de aula. Diversos estudos recentes utilizam técnicas de visão computacional para detectar posturas corporais, direção do olhar, expressões faciais e movimentação dos estudantes como forma de inferir estados de atenção e participação em atividades pedagógicas (Lin et al., 2021; Huang, Cheng e Wu, 2022; Wang et al., 2024; Liu, Jiang e Jiang, 2025).

A utilização de modelos de aprendizado profundo permitiu avanços significativos nesse tipo de análise, especialmente em cenários com múltiplos indivíduos e ambientes dinâmicos. Técnicas modernas de detecção visual, fundamentadas em redes neurais convolucionais, passaram a possibilitar o reconhecimento simultâneo de diversos objetos em uma mesma cena, ampliando a aplicabilidade da inteligência artificial em contextos complexos, como salas de aula (Redmon et al., 2016; Bochkovskiy, Wang e Liao, 2020).

Assim, a integração entre inteligência artificial, redes neurais convolucionais e visão computacional permite transformar elementos visuais observáveis em indicadores quantitativos sobre o comportamento dos alunos. Essa base fundamenta os modelos de detecção visual utilizados nesta pesquisa, especialmente o YOLO, apresentado na seção seguinte.

2.3.2 Modelo YOLO para Detecção Visual

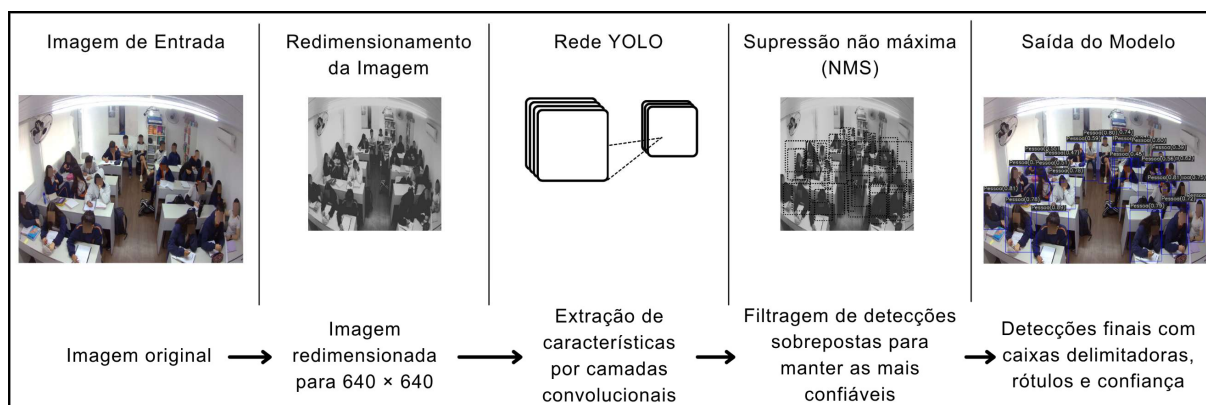
O YOLO, sigla de *You Only Look Once*, proposto originalmente por Joseph Redmon, é um modelo de detecção de objetos amplamente utilizado em aplicações de visão computacional que demandam rapidez e precisão. Diferentemente de abordagens tradicionais, que executam a detecção em múltiplas etapas, o YOLO trata a identificação de objetos como um único problema de regressão, permitindo que a localização e a classificação sejam realizadas simultaneamente em apenas uma passagem pela rede neural convolucional, característica que proporciona elevado desempenho e capacidade de processamento em tempo real. Essa abordagem tornou o modelo amplamente utilizado em aplicações que exigem respostas rápidas, como sistemas de vigilância, direção autônoma, análise médica por imagens, monitoramento industrial e reconhecimento comportamental (Redmon et al., 2016).

Em sua formulação original, o funcionamento desse modelo baseia-se na divisão da imagem em uma grade composta por múltiplas células. Cada célula é responsável por prever caixas delimitadoras, níveis de confiança e probabilidades de

pertencimento a determinadas classes. As caixas delimitadoras representam a localização espacial dos objetos detectados na imagem, enquanto o valor de confiança indica a probabilidade de que determinado objeto esteja realmente presente naquela região (Redmon et al., 2016).

De forma simplificada, o processamento realizado por esse modelo pode ser compreendido como uma sequência que envolve a entrada da imagem, seu redimensionamento para um formato compatível com o modelo, o processamento por uma rede convolucional e a geração das detecções finais. Após a predição inicial das caixas delimitadoras, pode ser aplicada a supressão não máxima, ou *Non-Maximum Suppression* (NMS), procedimento utilizado para reduzir detecções sobrepostas e manter aquelas com maior confiança. Esse fluxo é apresentado de maneira simplificada na Figura 6.

Figura 6 - Estrutura simplificada do processamento realizado pelo modelo YOLO aplicado a imagens de sala de aula



Fonte: elaborado pelo autor, com base em Redmon et al. (2016).

Na Figura 6, observa-se que a imagem original de sala de aula é inicialmente redimensionada para um formato padronizado, como 640×640 pixels. Em seguida, a rede realiza a extração de características visuais por meio de camadas convolucionais e subamostragem. A etapa de supressão não máxima atua sobre as detecções sobrepostas, preservando aquelas consideradas mais confiáveis. Por fim, a saída do modelo apresenta as detecções finais, compostas por caixas delimitadoras, rótulos de classe e valores de confiança.

Em versões recentes da família YOLO, essa arquitetura costuma ser compreendida em três componentes gerais: *backbone*, *neck* e *head*. O *backbone* é

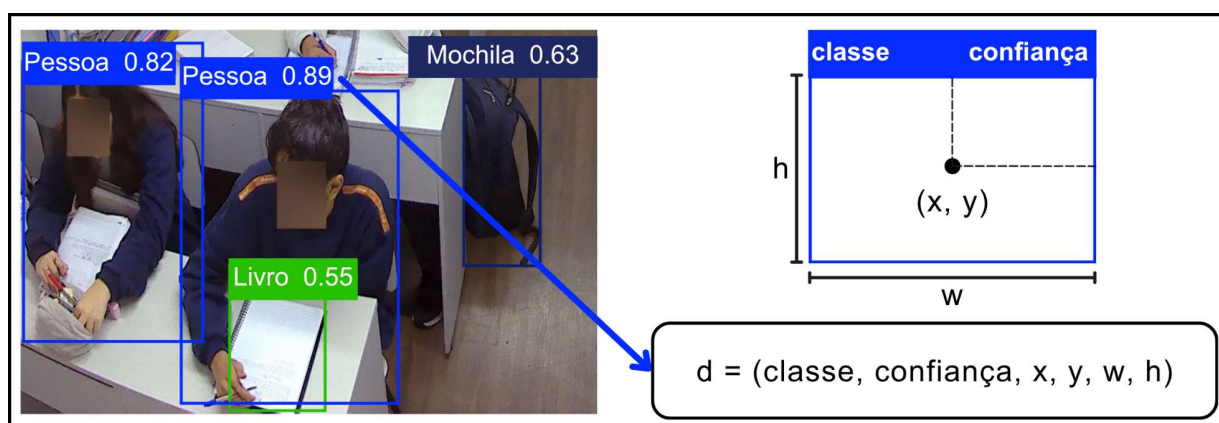
responsável pela extração inicial de características da imagem, o *neck* combina e refina características em diferentes escalas, e a *head* realiza a predição final, gerando caixas delimitadoras, classes e valores de confiança. Essa divisão é útil para compreender a organização geral desses modelos, incluindo versões recentes utilizadas em bibliotecas de detecção visual (Bochkovskiy, Wang e Liao, 2020; Ultralytics, 2024).

Desde sua proposta inicial, o YOLO passou por diversas evoluções em sua arquitetura, incluindo versões especializadas e treinadas para a detecção de classes específicas. Cada nova versão incorporou melhorias relacionadas à velocidade de processamento, precisão de detecção e capacidade de generalização em cenários complexos, incluindo ambientes com oclusões, múltiplos objetos e diferentes escalas (Bochkovskiy, Wang e Liao, 2020).

É importante destacar que a saída desse modelo não corresponde diretamente a uma interpretação complexa da cena, como atenção, engajamento ou participação. O modelo retorna informações visuais estruturadas sobre os objetos detectados. Cada detecção pode ser descrita por uma classe, um valor de confiança e uma caixa delimitadora. A classe indica o tipo de objeto identificado, como pessoa, mochila ou livro. A confiança representa o grau de segurança atribuído pelo modelo à detecção. Já a caixa delimitadora indica a localização espacial do objeto na imagem.

A Figura 7 ilustra a saída da detecção feita pelo modelo. Nela, cada objeto identificado na imagem é associado a uma caixa delimitadora e a um rótulo contendo a classe prevista e o valor de confiança. A caixa delimitadora pode ser descrita pelas coordenadas (x, y) , que indicam a posição central, e por suas dimensões w e h .

Figura 7 - Elementos que compõem a saída de uma detecção do modelo YOLO



Fonte: elaborado pelo autor, com base em Redmon et al. (2016) e Vanin (2021).

De forma simplificada, uma detecção pode ser representada como $d = (classe, confiança, x, y, w, h)$. Nessa representação, a classe corresponde ao objeto identificado, a confiança indica o grau de segurança atribuído pelo modelo à detecção, x e y representam a posição central da caixa delimitadora na imagem, e w e h indicam sua largura e altura. Essa estrutura permite que a informação visual extraída da imagem seja organizada em dados quantitativos e posteriormente utilizada em análises computacionais.

Diversos estudos recentes demonstram a aplicação do YOLO em problemas relacionados ao comportamento humano e ao contexto educacional. Wang et al. (2024), por exemplo, utilizaram o YOLOv5 para detectar alunos em vídeos de sala de aula, permitindo análises comportamentais automatizadas. Lin et al. (2021) utilizaram técnicas de detecção corporal associadas à estimativa de pose para reconhecimento de comportamentos estudantis, enquanto Wang et al. (2023) integraram outros mecanismos para melhorar a detecção de comportamentos em salas inteligentes. Esse modelo também vem sendo empregado em áreas como reconhecimento facial, monitoramento de tráfego, segurança eletrônica e análise médica por imagens.

Além disso, a revisão sistemática conduzida por Liu, Jiang e Jiang (2025) reforça a relevância do YOLO no contexto educacional ao demonstrar que a maior parte dos estudos recentes sobre reconhecimento de comportamento em sala de aula utiliza esse modelo como principal técnica de detecção visual. Segundo os autores, essa predominância está relacionada principalmente à elevada velocidade de processamento, à capacidade de detecção em tempo real e à eficiência do modelo em ambientes com múltiplos indivíduos (Liu, Jiang e Jiang, 2025).

Nesse contexto, o modelo YOLO mostra-se relevante para o método proposto nesta dissertação por apresentar equilíbrio entre velocidade de processamento, precisão e viabilidade em ambientes computacionais acessíveis. Assim, foi empregado no recurso educacional baseado em visão computacional utilizado nesta pesquisa para a detecção de pessoas e faces, permitindo extrair informações como quantidade de indivíduos detectados, quantidade de faces identificadas e valores de confiança associados às detecções.

2.4 REVISÃO DA LITERATURA SOBRE VISÃO COMPUTACIONAL EM ANÁLISE DE ATENÇÃO EM SALA DE AULA

Esta seção apresenta uma análise com base em 14 artigos relacionados à aplicação de visão computacional no monitoramento dos alunos em âmbito escolar, dando ênfase nos objetivos, técnicas e conclusões dos estudos. Esse tipo de avaliação permite identificar padrões, contribuições e limitações, oferecendo subsídio para esta pesquisa.

Os artigos selecionados para a pesquisa estão listados na Tabela 1 em ordem decrescente de citações, com o intuito de evidenciar o impacto das publicações na comunidade científica. Porém, é importante ressaltar que esse critério pode favorecer artigos antigos, uma vez que possuem mais tempo de exposição.

Tabela 1 - Artigos selecionados

Ano	Título de artigo	Autores	Citações
2021	Student Behavior Recognition System for the Classroom Environment Based on Skeleton Pose Estimation and Person Detection	Lin et al.	81
2021	Multimodal Engagement Analysis from Facial Videos in the Classroom	Sümer et al.	48
2024	Detecting non-verbal speech and gaze behaviors with multimodal data and computer vision to interpret effective collaborative learning interactions	Zhou, Suraworachet e Cukurova.	15
2023	Students' Classroom Behavior Detection System Incorporating Deformable DETR with Swin Transformer and Light-Weight Feature Pyramid Network	Wang et al.	14
2022	Analysis of Learning Behavior of Human Posture Recognition in Maker Education	Huang, Cheng e Wu.	7
2023	Campus Abnormal Behavior Recognition with Temporal Segment Transformers	Liu et al.	4
2024	SBD-Net: Incorporating Multi-Level Features for an Efficient Detection Network of Student Behavior in Smart Classrooms	Wang et al.	4
2022	Identifying and Monitoring Students' Classroom Learning Behavior Based on Multisource Information	Yin Albert et al.	3
2022	Design of International Chinese Education Promotion Platform Based on Artificial Intelligence and Facial Recognition Technology	Shen e Sun.	1
2023	AMBCON: virtual learning environment connected to smart technologies for student monitoring	Missaglia et al.	0
2024	Application of Video Abnormal Behavior Detection Algorithm in Evaluation of Track and Field Teaching and Training Effect	Zhang, J. et al	0

2024	Enhancing English Learning Environments Through Real-Time Emotion Detection and Sentiment Analysis	Orosoo et al.	0
2024	Research on the Informatization Strategy and Practical Effect of Chinese Traditional Culture Education	Wei e Ma.	0
2025	Classroom Behavior Recognition Using Computer Vision: A Systematic Review	Liu, Jiang e Jiang.	0

Fonte: Elaborado pelo autor.

O estudo de Lin et al. (2021) propõe um sistema para reconhecer o comportamento de alunos em sala de aula usando estimativa de pose com base na detecção da posição do esqueleto. Utilizando imagens capturadas em tempo real e um modelo de rede neural treinado e a normalização das posições das juntas, o sistema conseguiu classificar diferentes posturas estudantis com maior precisão em comparação a abordagens anteriores. Dessa forma, os resultados indicaram que essa tecnologia pode ser útil para monitorar automaticamente o comportamento dos alunos no ambiente escolar.

Sümer et al. (2021) analisam de forma automatizada o engajamento de estudantes do ensino médio em ambientes de aprendizagem tradicionais como salas de aula, utilizando visão computacional para classificar o nível de engajamento a partir das expressões faciais de alunos do ensino médio. Foram aplicadas redes neurais treinadas e personalizadas com 60 segundos de vídeo devidamente rotulado por especialistas, para a estimativa de pose da cabeça e reconhecimento de expressões faciais. De forma complementar, Missaglia et al. (2023) propõem a arquitetura AMBCON, voltada à integração entre ambientes virtuais de aprendizagem e tecnologias inteligentes capazes de monitorar a concentração do aprendiz durante a interação com objetos de aprendizagem. A proposta considera o uso de recursos como reconhecimento facial, análise de expressões e rastreamento ocular, utilizando a câmera do equipamento do estudante como meio de captação dos dados. Embora o estudo esteja direcionado ao contexto da educação a distância, sua contribuição é relevante para esta pesquisa por demonstrar a viabilidade de integrar tecnologias de monitoramento visual a ambientes educacionais, com o objetivo de acompanhar estados de atenção e apoiar a avaliação do processo de aprendizagem.

Huang, Cheng e Wu (2022) investigam a aplicação da inteligência artificial e do processamento de imagens para reconhecer comportamentos de aprendizagem em ambientes de *Maker Education*, caracterizados por práticas pedagógicas baseadas

na experimentação, criação e resolução prática de problemas. Os autores desenvolveram um sistema baseado em redes neurais convolucionais para identificar posturas e ações dos alunos durante essas atividades práticas, permitindo avaliar o desempenho de forma automatizada e alcançando precisão de 0,83 no reconhecimento de comportamentos. De forma similar, Zhou, Suraworachet e Cukurova (2024) mostram como interações em aprendizagem colaborativa podem ser analisadas usando ferramentas de visão computacional. A pesquisa, conduzida com estudantes de pós-graduação ao longo de 10 semanas, identificou padrões de comportamento não verbal que diferenciam grupos com melhores resultados colaborativos. Utilizando técnicas de mineração, os autores observaram que grupos mais eficazes apresentam ciclos de comportamento semelhantes entre os membros. Ambos os estudos sugerem que comportamentos não verbais podem ser automaticamente detectados, oferecendo apoio à análise da aprendizagem.

Wang et al. (2023) propõem uma abordagem inovadora para a detecção de comportamentos em sala de aula, superando as limitações dos métodos baseados em CNN. Para isso, os autores integram técnicas para aprimorar a detecção mesmo em imagens com deformações, permitindo o processamento eficaz de múltiplas escalas. Para validar o método proposto, foram realizados experimentos em um conjunto de dados real de alunos em sala de aula, os quais demonstraram uma melhoria de 6,1% na acurácia em relação aos métodos de referência.

Liu et al. (2023) criam um sistema de vigilância inteligente para campus escolar, focado na detecção de comportamentos anormais. A pesquisa também desenvolve uma nova estrutura que utiliza modelagem temporal de longo alcance de vídeos e uma estratégia de amostragem uniforme para extrair características. O modelo proposto também usou o conjunto de dados CABR50, composto por 50 classes de ações anormais, e demonstrou um desempenho competitivo, alcançando acurácias de até 97,16%, ressaltando seu potencial para aprimorar a segurança em ambientes escolares.

Wang et al. (2024) desenvolveram a rede Student Behavior Detection Network (SBD-Net), uma abordagem voltada ao monitoramento e à análise do comportamento dos alunos em salas de aula inteligentes em tempo real. O estudo utiliza recursos como ES Loss, Focal Modulation e DyHead para melhorar a fusão de características e lidar com desafios como a variedade de comportamentos, oclusões entre alunos e distribuição desigual das classes analisadas. Os resultados experimentais com o

conjunto de dados SCBehavior indicaram aumento de 4,3% na precisão, mesmo com baixa complexidade computacional. De forma complementar, Wei e Ma (2024) apresentam a Retina Student Network, desenvolvida para detectar e codificar comportamentos dos alunos em sala de aula. Para isso, os autores criaram um sistema de codificação específico para estados de olhar e não olhar. Após a análise, os autores apontam que o método tradicional de “explicação-questionamento-resposta-feedback” permanece predominante.

Yin Albert et al. (2022) usam uma CNN de três dimensões para o reconhecimento de comportamentos dos estudantes em tempo real em sala de aula. A pesquisa aborda as limitações dos modelos anteriores, que não consideram a informação de movimento entre quadros consecutivos, e destaca como a convolução 3D pode melhorar a análise de comportamento humano ao estender a convolução 2D para o domínio espaço-temporal. Já o estudo de Orosoo et al. (2024) também utiliza CNN tridimensional, para investigar a melhoria da aprendizagem de idiomas, integrando redes LSTM para análise multimodal de emoções. A abordagem combina a análise de imagens 3D e o processamento de linguagem natural para extrair padrões emocionais de alunos, oferecendo feedback abrangente que pode aprimorar as estratégias de ensino. Ambos os estudos apresentam resultados experimentais que demonstram a melhora significativa na análise de sala de aula, oferecendo suporte para professores e alunos.

Shen e Sun (2022) exploram a aplicação da tecnologia de processamento de imagens digitais para o reconhecimento facial e verificação de identidade, no contexto de ensino de língua chinesa. A pesquisa foca no uso dessa tecnologia em modelos de educação online, onde a observação e análise das emoções faciais dos alunos se tornam uma ferramenta eficaz para avaliar o engajamento e o desempenho dos estudantes, bem como para identificar conflitos e dificuldades para uma avaliação contínua.

Zhang, J. et al. (2024) estudam tecnologias de detecção de comportamentos anormais em vídeos para avaliar a qualidade do ensino em aulas de atletismo. A pesquisa foca em três tecnologias-chave para detectar comportamentos anormais: detecção de alvos em movimento, rastreamento de alvos em movimento e detecção de comportamentos anormais usando teorias de visão computacional, processamento de imagens e tecnologias de análise de vídeo.

Liu, Jiang e Jiang (2025) fazem uma revisão sistemática sobre o reconhecimento de comportamento em sala de aula com base em visão computacional, analisando 80 artigos científicos segundo as diretrizes PRISMA. A pesquisa identificou que os estudos se concentram principalmente na detecção de ações físicas e engajamento dos alunos, utilizando majoritariamente a série YOLO como técnica de reconhecimento. No entanto, destaca-se uma lacuna na correlação entre a caracterização do comportamento estudantil e os conteúdos instrucionais, assim como observado na revisão feita nessa pesquisa, evidenciando a necessidade de análises que integrem melhor os aspectos pedagógicos. Além disso, são discutidos desafios no design experimental, na aplicabilidade dos métodos e na implementação prática dessas tecnologias no contexto educacional.

As informações extraídas dos artigos estão organizadas no

Quadro 1, que apresenta os principais aspectos de cada artigo: os autores, o propósito central da pesquisa, as técnicas computacionais empregadas e as limitações apontadas pelos próprios estudos. Essa abordagem permite identificar os padrões metodológicos mais recorrentes e os desafios ainda enfrentados.

Quadro 1 - Síntese dos resultados da RSL

Autor	Objetivo	Técnica	Limitações
Lin et al. (2021)	Reconhecer posturas de estudantes em sala de aula.	OpenPose, DNN e análise de pose corporal.	Desempenho afetado por oclusões e salas cheias.
Sümer et al. (2021)	Detectar engajamento a partir de vídeos faciais.	Attention-Net, Affect-Net e análise temporal.	Exige vídeos faciais e rotulação especializada.
Huang, Cheng e Wu (2022)	Reconhecer comportamentos em ambientes de educação maker.	CNN e reconhecimento de postura.	Aplicação restrita a atividades maker.
Shen e Sun (2022)	Aplicar reconhecimento facial em plataforma educacional.	Reconhecimento facial e análise de emoções.	Foco restrito ao ensino de língua chinesa.
Yin Albert et al. (2022)	Monitorar comportamentos em sala de aula.	CNN 3D e análise espaço-temporal.	Depende de vídeos de boa qualidade e rotulação.
Missaglia et al. (2023)	Integrar AVA a tecnologias inteligentes de monitoramento.	AMBCON, reconhecimento facial e rastreamento ocular.	Validado por prova de conceito em contexto de EaD.

Liu et al. (2023)	Reconhecer comportamentos anormais em campus.	Temporal Segment Transformers.	Foco em segurança, não em aprendizagem.
Wang et al. (2023)	Melhorar a detecção de comportamentos em sala.	Deformable DETR, Swin Transformer, FPN e CARAFE.	Maior demanda computacional e curadoria de dados.
Zhou, Suraworachet e Cukurova (2024)	Analisar interações colaborativas por dados multimodais.	Visão computacional e mineração de processos.	Foco em grupos de pós-graduação.
Zhang, J. et al. (2024)	Avaliar aulas de atletismo por detecção comportamental.	Detecção de movimento e PSO-TSVM.	Restrito a aulas de educação física.
Orosoo et al. (2024)	Detectar emoções no ensino de inglês em tempo real.	CNN 3D, LSTM e PLN.	Foco específico no ensino de idiomas.
Wei e Ma (2024)	Detectar estados de atenção em aulas de cultura tradicional.	Retina Student Network, Inception-v3 e Vision Transformer.	Restrito ao contexto da cultura chinesa.
Wang et al. (2024)	Detectar múltiplos comportamentos em salas inteligentes.	SBD-Net, ES Loss, Focal Modulation e DyHead.	Necessita de base própria e adaptação.
Liu, Jiang e Jiang (2025)	Revisar estudos sobre comportamento em sala com visão computacional.	Revisão sistemática.	Publicações de 2016 a 2024.

Fonte: Elaborado pelo autor.

A análise dos trabalhos sintetizados no Quadro 1 evidencia o predomínio do uso de redes neurais e técnicas de aprendizado profundo para detectar posturas, faces, expressões e comportamentos em ambientes educacionais. Embora essas abordagens demonstrem avanços na extração automatizada de informações visuais, grande parte dos estudos concentra-se nas etapas de detecção ou classificação. Observa-se menor detalhamento sobre como as saídas produzidas pelos modelos podem ser transformadas em indicadores quantitativos de atenção, consolidadas em unidades comparáveis e associadas temporalmente a dados de desempenho escolar.

Assim, a lacuna identificada não se restringe à quantidade limitada de pesquisas que relacionam atenção e desempenho escolar, mas envolve também a necessidade de um percurso metodológico que permita tornar os dados extraídos das imagens correlacionáveis com informações escolares. Diante disso, esta pesquisa propõe e aplica um recurso educacional baseado em visão computacional que utiliza as detecções de pessoas e faces para operacionalizar indicadores visuais de atenção,

convertê-los em notas e integrá-los aos dados de desempenho para a aplicação de análises estatísticas. Esse percurso busca responder como correlacionar a atenção em sala de aula, estimada por meio de visão computacional, com o desempenho escolar dos alunos.

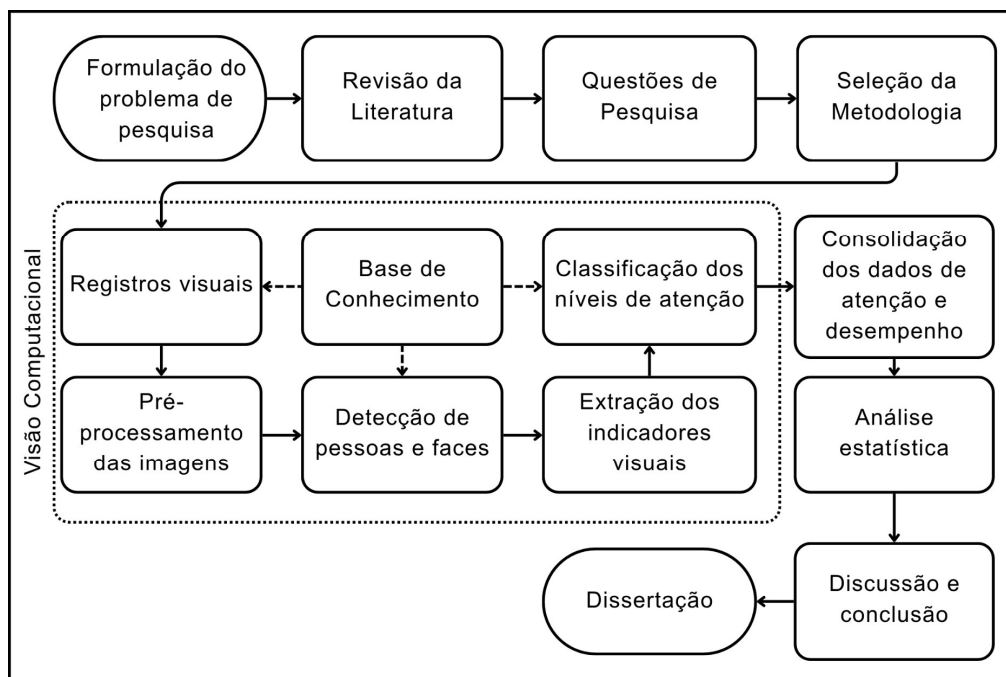
3 MATERIAIS E MÉTODOS

Este capítulo apresenta os procedimentos metodológicos adotados para o desenvolvimento da pesquisa, contemplando a caracterização do estudo, o método utilizado para a revisão da literatura, a descrição dos dados empregados e as etapas de tratamento e análise dos dados.

Em alinhamento à questão de pesquisa, são descritos os procedimentos utilizados para transformar os registros visuais de sala de aula em indicadores quantitativos de atenção e associá-los aos dados de desempenho escolar. Esse percurso compreende a organização dos dados de entrada, o processamento computacional das imagens, a extração e consolidação dos indicadores visuais e a posterior aplicação das análises estatísticas correlacionais.

A Figura 8 apresenta o fluxo geral da pesquisa, desde a formulação do problema e a revisão da literatura até a seleção da metodologia, o processamento dos registros visuais, a consolidação dos dados, a análise estatística e a elaboração da discussão e da conclusão. O eixo relacionado à visão computacional sintetiza as etapas de pré-processamento das imagens, detecção de pessoas e faces, extração dos indicadores visuais e classificação dos níveis de atenção.

Figura 8 - Fluxograma do projeto de pesquisa



Fonte: Elaborado pelo autor

Como indicado na Figura 8, a pesquisa articula etapas teóricas, metodológicas e computacionais. As etapas iniciais orientam a definição do problema e dos procedimentos adotados. Na sequência, os registros visuais são processados para a geração dos indicadores de atenção, posteriormente consolidados com os dados de desempenho escolar. O detalhamento das entradas, das etapas computacionais, dos indicadores produzidos e das análises estatísticas é apresentado nas subseções seguintes.

3.1 TIPOLOGIA DA PESQUISA

Esta dissertação caracteriza-se como uma pesquisa aplicada, de natureza mista e objetivos correlacionais, desenvolvida por meio de um estudo de caso, utilizando técnicas de coleta de dados documentais e bibliográficos.

De acordo com Gil (2008), a pesquisa aplicada tem como propósito gerar conhecimentos voltados à solução de problemas concretos, atendendo a demandas específicas de determinado contexto. Nesse sentido, o presente estudo propõe a aplicação de um recurso educacional baseado em visão computacional como ferramenta de monitoramento da atenção em sala de aula.

A abordagem quantitativa se evidencia pela análise de dados numéricos e objetivos processados por meio de modelos computacionais e técnicas estatísticas de correlação. Por outro lado, a dimensão qualitativa complementa a investigação ao permitir a interpretação dos resultados e os aspectos que envolvem o fenômeno da atenção no ambiente escolar (Marconi et al., 2002).

Quanto aos objetivos, trata-se de uma pesquisa correlacional, pois busca identificar relações estatísticas entre as notas médias de atenção dos períodos analisados e as notas de desempenho escolar correspondentes, sem estabelecer causalidade direta entre as variáveis (Cervo e Bervian, 1996).

Em relação ao método, a pesquisa configura-se como um estudo de caso, uma vez que investiga em profundidade um grupo específico de alunos do 9º ano do Ensino Fundamental de uma escola privada do estado de São Paulo. Esse método é apropriado para a análise de fenômenos em seu contexto real, permitindo examinar múltiplas variáveis e suas inter-relações (Yin, 2015).

As técnicas de coleta de dados adotadas foram a pesquisa documental e a pesquisa bibliográfica. A pesquisa documental fundamenta-se na utilização de registros institucionais previamente coletados, os quais, segundo Gil (2008), permitem novas interpretações sobre materiais que ainda não receberam tratamento analítico sistemático. Nesta dissertação, essa técnica foi empregada na análise dos registros visuais das aulas e dos dados de desempenho escolar disponibilizados pela instituição, mediante as condições apresentadas no Apêndice B. Já a pesquisa bibliográfica baseou-se na revisão de publicações científicas relacionadas à atenção, aprendizagem e visão computacional, fornecendo suporte teórico para a análise e interpretação dos dados empíricos.

Assim, a combinação metodológica adotada demonstra-se adequada aos propósitos desta dissertação, ao integrar fundamentos teóricos e procedimentos capazes de investigar a relação entre atenção e desempenho escolar.

3.2 MÉTODO PARA A REVISÃO DA LITERATURA

Para construir a base teórica desta pesquisa, foi realizada uma revisão da literatura estruturada com base nas diretrizes do *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA), propostas por Moher et al. (2009). Inicialmente, foram selecionados os constructos apresentados no Quadro 2, utilizados para compor as buscas realizadas nas bases Scopus, IEEE Xplore e Web of Science.

Quadro 2 - Termos utilizados para compor os constructos

Constructo	Termo em inglês	Termo em português
C1	"Classroom"	"Sala de aula"
C2	"Students"	"Alunos"
C3	"School"	"Escola"
C4	"Biomechanics"	"Biomecânica"
C5	"Facial"	"Facial"
C6	"Eye tracking"	"Rastreamento ocular"
C7	"Behavior"	"Comportamento"
C8	"Image analysis"	"Análise de imagem"
C9	"Image processing"	"Processamento de imagem"
C10	"Computer vision"	"Visão computacional"
C11	"Attention"	"Atenção"
C12	"Interactions"	"Interações"
C13	"Learn"	"Aprender"

Fonte: Elaborado pelo autor.

A seleção dos constructos a serem aplicados nas buscas foi aprimorada de acordo com o Tesouro Brasileiro de Ciência da Informação, que permitiu a montagem de algumas sentenças de consulta a serem utilizadas, as quais estão apresentadas no Quadro 3.

Quadro 3 - Sentenças de busca para as consultas

Consulta	Sentença	Resultados
Consulta 1	(C1 or C2 or C3) and (C4 or C5 or C6 or C7) and (C8 or C9 or C10) and (C11 or C12 or C13)	237
Consulta 2	(C1 or C2 or C3) and (C4 or C5 or C6 or C7) and (C8 or C9) and (C10)	36
Total		273

Fonte: Elaborado pelo autor.

Cada busca realizada gerou resultados nas três bases de dados, e o número total de resultados encontrados está apresentado no Quadro 3. Devido à quantidade expressiva de resultados, foram aplicados os métodos de exclusão descritos por Liao et al. (2017) adaptados para conduzir o processo de inclusão e exclusão dos documentos. Os critérios utilizados nesse processo estão detalhadamente descritos no Quadro 4.

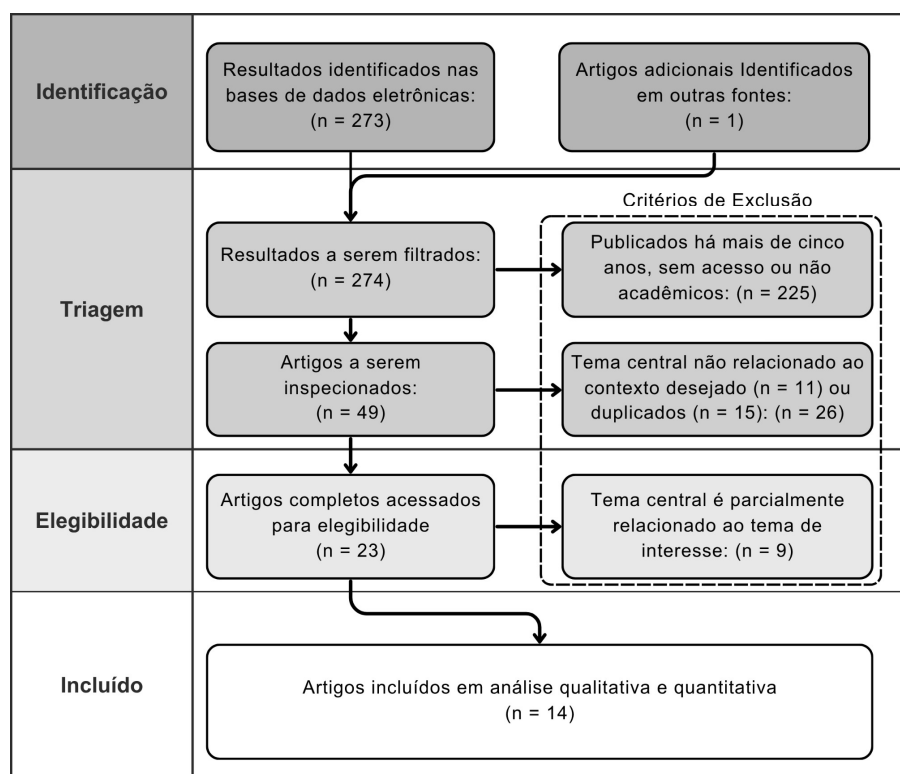
Quadro 4 - Critérios de inclusão e exclusão de trabalhos na pesquisa

Critério		Descrição
Inclusão	CI1	A relação entre atenção, desempenho escolar e visão computacional é um dos aspectos analisados
Exclusão	CE1	Trabalho publicado há mais de cinco anos, sem acesso ou de natureza não acadêmica.
	CE2	A visão computacional está sendo aplicada em um contexto distinto ao de interesse
	CE3	Uma pesquisa que aborda visão computacional sem foco direto em educação

Fonte: Adaptado de Liao et al. (2017).

A busca por artigos se restringiu aos últimos cinco anos e nela foi possível verificar a evolução dos trabalhos a respeito do tema. Os critérios de exclusão aplicados reduziram gradualmente o número de artigos incluídos na pesquisa, conforme ilustrado na Figura 9.

Figura 9 - Fluxograma das etapas de aplicação do PRISMA



Fonte: Adaptado de Moher et al. (2009)

Como apresentado na Figura 9, a aplicação das etapas do PRISMA permitiu organizar o processo de identificação, triagem, elegibilidade e inclusão dos estudos analisados. Inicialmente, foram identificados 274 trabalhos, considerando os resultados das bases eletrônicas e o artigo adicional selecionado em outras fontes. Após a aplicação dos critérios de exclusão e a leitura dos trabalhos elegíveis, permaneceram 14 artigos, os quais compuseram a base final da revisão da literatura desta pesquisa.

3.3 DADOS EMPREGADOS NA PESQUISA

Os dados utilizados nesta pesquisa foram cedidos por uma escola da rede privada localizada no estado de São Paulo, Brasil. A amostra foi composta por 65 estudantes do 9º ano do Ensino Fundamental, com idades entre 14 e 15 anos e distribuição equilibrada entre os gêneros. O período analisado compreendeu os meses de setembro a novembro de 2025, correspondentes ao último trimestre letivo.

O contexto da instituição caracteriza-se pelo acesso a recursos tecnológicos e infraestrutura adequada para o registro e o armazenamento das aulas, possibilitando

a geração sistemática dos dados visuais utilizados na pesquisa. Todo o material foi disponibilizado mediante autorização institucional e conforme as condições estabelecidas no Termo de Confidencialidade apresentado no APÊNDICE B – TERMO DE CONFIDENCIALIDADE ASSINADO.

A amostragem adotada foi não probabilística, por conveniência, uma vez que o pesquisador não interferiu no processo de seleção dos participantes ou na geração dos dados institucionais analisados (Gil, 2008). O conjunto de dados foi organizado em dois grupos principais: os registros visuais das aulas e os dados de desempenho escolar dos estudantes.

Os registros visuais foram constituídos por imagens extraídas das aulas em intervalos de 10 segundos. Cada imagem preservava informações de data e horário, permitindo manter sua organização temporal e relacioná-la posteriormente ao período das atividades avaliativas. Foram consideradas apenas imagens obtidas durante aulas expositivas em formato tradicional, nas quais os estudantes permaneciam em uma configuração compatível com o contexto analisado.

Foram excluídos os registros relacionados à aplicação de provas, atividades em grupo, dinâmicas coletivas, ensaios e outras situações com características distintas das aulas expositivas. Esse critério teve como objetivo reduzir variações decorrentes da natureza da atividade e manter maior uniformidade nas condições em que os indicadores visuais de atenção foram extraídos.

O segundo grupo de dados foi composto pelas notas das avaliações contínuas aplicadas pela própria instituição ao longo do período analisado. Essas notas, expressas em escala de 0 a 10, foram utilizadas como indicadores de desempenho escolar. As atividades avaliativas também possuíam referência temporal, o que permitiu estabelecer correspondência entre os dados escolares e os registros visuais.

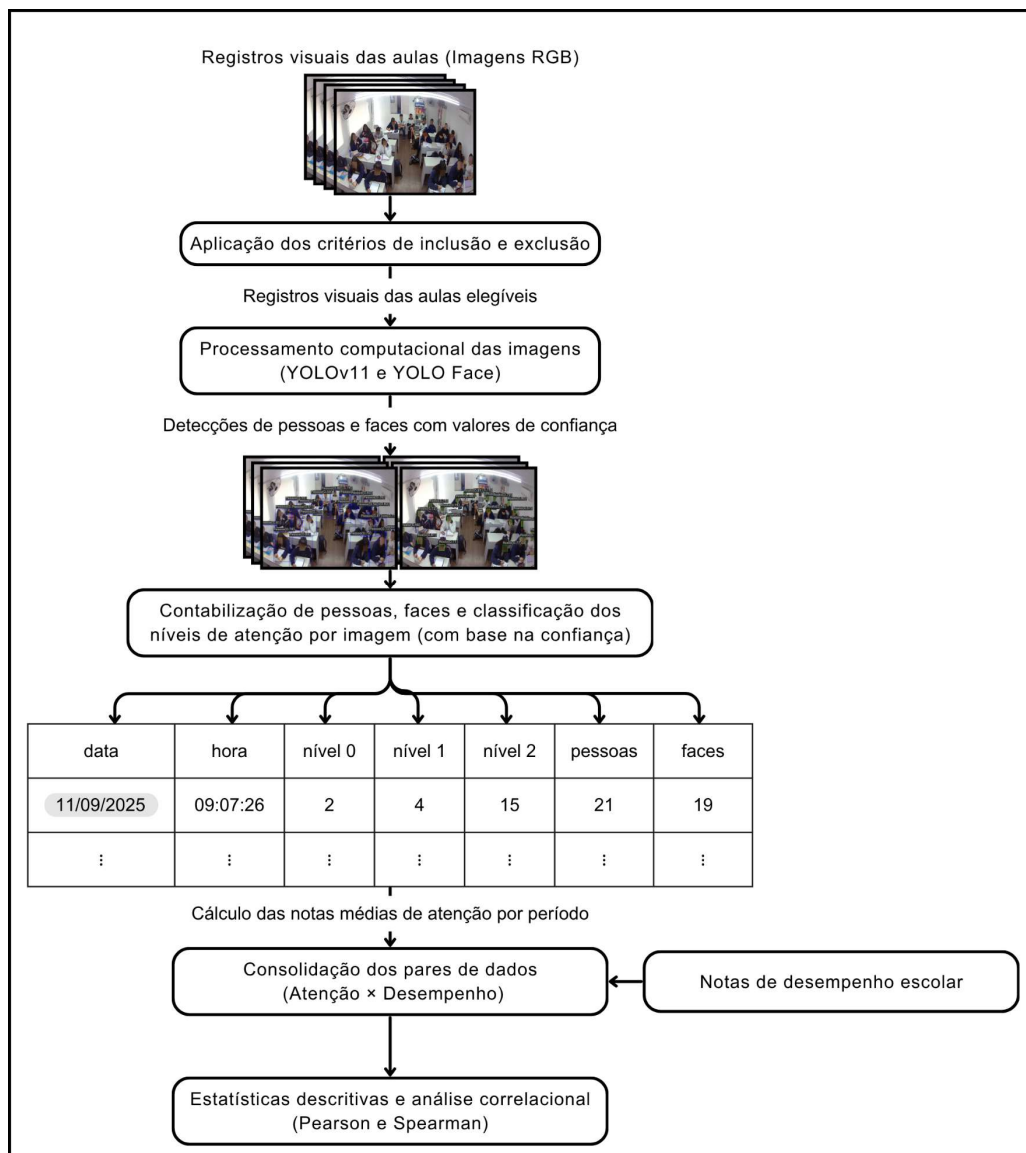
Para realizar essa associação, cada atividade avaliativa considerada na pesquisa foi utilizada como um marco temporal de referência. Os registros visuais correspondentes às aulas que antecederam cada avaliação foram posteriormente agrupados para a obtenção de uma nota média de atenção. Essa nota foi então associada à respectiva nota de desempenho escolar, formando os pares de dados utilizados nas análises correlacionais. Os procedimentos computacionais empregados para extrair, calcular e consolidar esses indicadores são detalhados na seção seguinte.

3.4 TRATAMENTO DE DADOS

Este tópico descreve os procedimentos adotados para o tratamento e a análise dos dados, abrangendo desde o processamento dos registros visuais até a formação da base utilizada nas análises estatísticas. O objetivo foi transformar as informações extraídas das imagens em indicadores quantitativos de atenção e associá-los aos dados de desempenho escolar.

A Figura 10 sintetiza o fluxo adotado para esse tratamento, incluindo a entrada das imagens e das notas escolares, o processamento computacional dos registros visuais e as demais etapas até a aplicação das análises descritivas e correlacionais.

Figura 10 - Fluxo do método proposto para estruturação dos dados



Fonte: Elaborado pelo autor

Como apresentado na Figura 10, o método foi organizado em etapas sucessivas. Os detalhes da estruturação e do processamento dos registros visuais, bem como os procedimentos empregados na análise dos dados, são apresentados nas subseções seguintes.

3.4.1 Estruturação e Processamento dos Dados

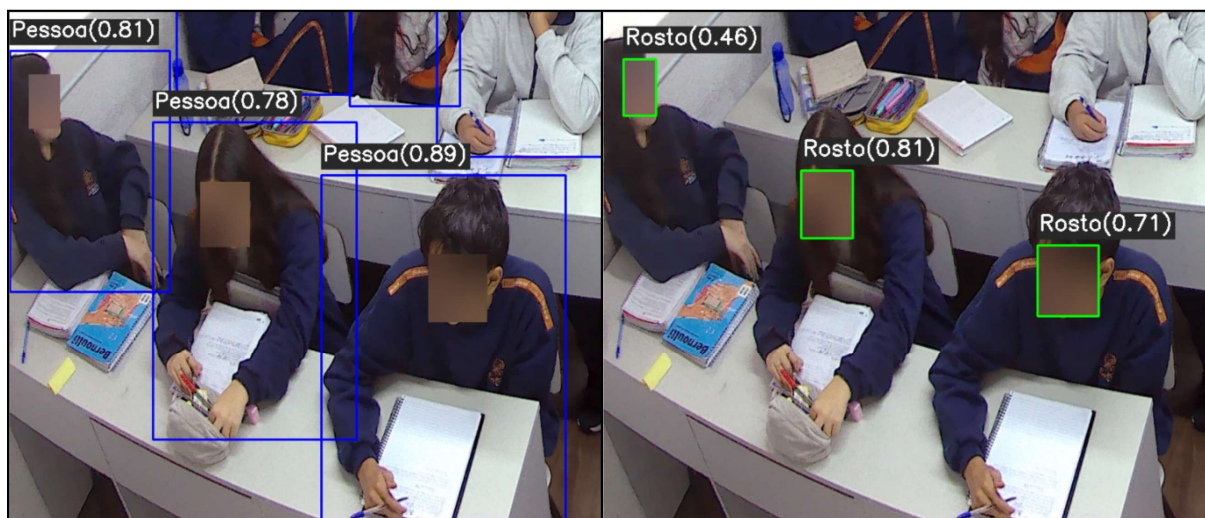
A etapa de estruturação dos dados teve como objetivo transformar os registros visuais em informações quantitativas que pudessem ser posteriormente relacionadas aos dados de desempenho escolar. Inicialmente, as imagens foram organizadas em diretórios digitais conforme a turma, a data e o horário do registro. Essa organização preservou a sequência temporal das imagens e permitiu seu agrupamento de acordo com os períodos relacionados às atividades avaliativas.

Os registros visuais disponibilizados foram extraídos em intervalos de 10 segundos, estratégia adotada para reduzir o volume total de dados a ser analisado e viabilizar o processamento computacional das imagens. Essa definição foi importante porque, caso fossem utilizados vídeos em sua taxa original de quadros por segundo, o volume de imagens gerado seria substancialmente maior, dificultando a organização, o armazenamento e o processamento dos dados no escopo desta pesquisa.

Após a aplicação dos critérios de inclusão e exclusão, as imagens foram processadas no ambiente Google Colab, por meio do código apresentado no Apêndice A desenvolvido em Python.

As imagens foram utilizadas como entrada em dois modelos. O YOLOv11 foi configurado para detectar a classe pessoa e retornou, para cada detecção, uma caixa delimitadora, suas coordenadas e o respectivo valor de confiança. O YOLO Face recebeu as mesmas imagens como entrada e retornou as caixas delimitadoras correspondentes às faces, suas coordenadas e os valores de confiança associados. Embora os modelos tenham capacidade de extrair todas essas informações, os dados utilizados nas etapas seguintes foram a quantidade de pessoas, a quantidade de faces e os valores de confiança das detecções faciais. A Figura 11 ilustra as saídas produzidas pelos modelos durante o processamento das imagens.

Figura 11 - Exemplos de detecção de pessoas e faces pelos modelos YOLOv11 e YOLO Face, com caixas delimitadoras e respectivos valores de confiança



Fonte: Elaborado pelo autor.

A identificação dos níveis de atenção foi realizada por meio de uma heurística baseada na detecção facial. Nesta pesquisa, a atenção não foi tratada como uma medida direta do estado cognitivo dos estudantes, mas como uma aproximação operacional. No contexto das aulas expositivas, considerou-se que a presença da face e o valor de confiança fornecido pelo YOLO Face poderiam constituir indícios do direcionamento visual do estudante para a região frontal da sala. A literatura indica que técnicas de visão computacional podem reconhecer padrões relacionados ao comportamento dos estudantes, como postura corporal, orientação facial, expressões e direção do olhar, utilizados como indícios de engajamento ou atenção em ambientes educacionais (Sümer et al., 2021; Lin et al., 2021; Uçar e Özdemir, 2022; Liu, Jiang e Jiang, 2025).

Modelos baseados em YOLO retornam a localização dos objetos detectados e valores de confiança que expressam a segurança do modelo em relação às detecções realizadas (Redmon et al., 2016; Bochkovskiy, Wang e Liao, 2020). No método proposto, a diferença entre o número de pessoas e faces detectadas em cada imagem foi atribuída ao nível 0. As faces detectadas com confiança igual ou inferior a 0,60 foram classificadas no nível 1, enquanto aquelas com confiança superior a 0,60 foram classificadas no nível 2. Os critérios adotados são apresentados no Quadro 5.

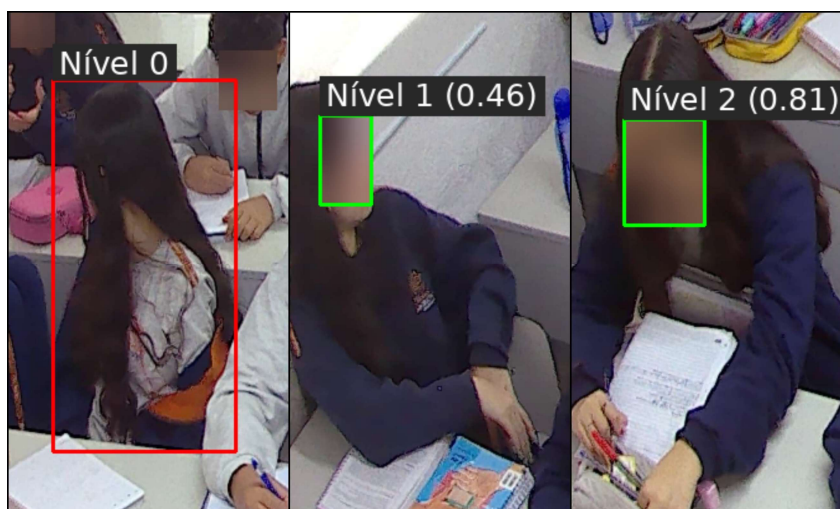
Quadro 5 - Critérios atribuídos aos níveis de atenção

Nível	Descrição
0	Atribuída aos indivíduos cuja face não foi detectada na imagem.
1	Associada a indivíduos com baixa confiança na detecção facial (confiança $\leq 0,60$).
2	Correspondente a indivíduos com alta confiança na detecção facial, interpretada como indicativo de maior direcionamento da atenção para a atividade proposta (confiança $> 0,60$).

Fonte: Elaborado pelo autor.

Com base nos critérios apresentados no Quadro 5, a Figura 12 apresenta de forma ilustrativa as três situações consideradas na pesquisa: pessoa sem face detectada, face detectada com confiança igual ou inferior a 0,60 e face detectada com confiança superior a 0,60.

Figura 12 - Exemplos de classificação dos níveis de atenção



Fonte: Elaborado pelo autor.

Após a contabilização dos níveis de atenção em cada imagem, os dados foram convertidos em uma nota de atenção da imagem, expressa em uma escala de 0 a 10. Aos níveis 0, 1 e 2 foram atribuídos, respectivamente, os pesos 0, 0,5 e 1. A nota de atenção correspondente a cada imagem foi calculada conforme a Equação 1:

$$A_i = 10 \times \frac{(0 \times n_{0i}) + (0,5 \times n_{1i}) + (1 \times n_{2i})}{n_{0i} + n_{1i} + n_{2i}} \quad (1)$$

em que:

- A_i representa a nota de atenção na imagem i ;
- n_{0i} representa a quantidade de indivíduos classificados no nível 0;
- n_{1i} representa a quantidade de indivíduos classificados no nível 1;
- n_{2i} representa a quantidade de indivíduos classificados no nível 2.

A multiplicação por 10 converte a média ponderada, originalmente situada entre 0 e 1, para a escala adotada na pesquisa. Assim, valores mais próximos de 10 indicam maior predominância de registros classificados no nível 2, enquanto valores mais próximos de 0 indicam maior predominância do nível 0.

Para cada um dos trinta períodos de referência, as notas obtidas nas imagens correspondentes foram consolidadas por meio da média aritmética, conforme a Equação 2:

$$A_p = \frac{\sum_{i \in I_p} A_i}{m_p} \quad (2)$$

em que:

- A_p representa a nota média de atenção do período p ;
- A_i representa a nota de atenção da imagem i ;
- I_p representa o conjunto de imagens pertencentes ao período p ;
- i representa uma imagem pertencente a esse conjunto;
- m_p representa a quantidade de imagens pertencentes ao período p .

Paralelamente, os dados de desempenho escolar foram organizados em planilhas eletrônicas de acordo com as trinta atividades avaliativas realizadas durante o período analisado. As notas, também expressas em uma escala de 0 a 10, foram utilizadas como indicadores de desempenho escolar.

Na etapa final da estruturação dos dados, a nota média de atenção de cada período foi associada à nota de desempenho escolar correspondente ao mesmo intervalo temporal. Esse procedimento resultou na formação dos pares de dados de atenção e desempenho utilizados nas análises descritivas e correlacionais apresentadas na subseção seguinte.

3.4.2 Abordagens Analíticas

Após a estruturação dos dados, com as variáveis organizadas de modo a permitir sua comparação, foram aplicadas duas abordagens complementares: a análise descritiva e a análise correlacional. A análise descritiva teve como objetivo caracterizar o comportamento das notas de atenção e de desempenho escolar, enquanto a análise correlacional foi utilizada para verificar a direção, a intensidade e a significância estatística da associação entre as variáveis.

Na etapa descritiva, foram calculadas medidas de tendência central e dispersão, incluindo média, mediana, valores mínimo e máximo e desvio-padrão. Também foram elaboradas tabelas e representações gráficas para observar a distribuição, a variabilidade e o comportamento das notas ao longo do período analisado. Essa etapa permitiu uma compreensão inicial dos dados antes da aplicação dos coeficientes de correlação (Gil, 2008; Field, 2018).

Já na etapa correlacional, foram aplicados os coeficientes de Pearson e Spearman aos pares formados pela nota média de atenção de cada período e pela respectiva nota de desempenho escolar. O coeficiente de correlação de Pearson foi utilizado para mensurar a intensidade e a direção da relação linear entre as duas variáveis quantitativas. Assim, cada par de notas de atenção e desempenho foi representado por (A_p, D_p) . O cálculo do coeficiente de Pearson é apresentado na Equação 3.

$$r = \frac{\sum_{p \in P} (A_p - \bar{A})(D_p - \bar{D})}{\sqrt{\left[\sum_{p \in P} (A_p - \bar{A})^2 \right] \left[\sum_{p \in P} (D_p - \bar{D})^2 \right]}} \quad (3)$$

em que:

- r representa o coeficiente de correlação de Pearson;
- A_p representa a nota média de atenção do período p ;
- D_p representa a nota de desempenho escolar no período p ;
- $\bar{A}$ representa a média das notas de atenção;
- $\bar{D}$ representa a média das notas de desempenho escolar;
- P representa o conjunto dos períodos analisados.

O valor desse coeficiente varia entre -1 e 1 . Valores positivos indicam associação direta entre as variáveis, valores negativos indicam associação inversa e valores próximos de zero indicam ausência ou baixa intensidade de relação linear.

Além do coeficiente de Pearson, foi aplicado o coeficiente de correlação de Spearman, indicado pela letra grega ρ . Esse coeficiente foi utilizado por ser uma medida não paramétrica de associação, adequada para avaliar relações monotônicas entre variáveis, mesmo quando a relação entre elas não é estritamente linear. Diferentemente de Pearson, que utiliza os valores originais das variáveis, Spearman considera os postos ou posições ocupadas pelos valores em cada conjunto de dados. Dessa forma, cada par formado pelas notas de atenção e desempenho é convertido em um par de posições relativas (R_{A_p}, R_{D_p}) , permitindo verificar se aumentos em uma variável tendem a ser acompanhados por aumentos ou reduções na outra variável. O cálculo do coeficiente de Spearman é apresentado na Equação 4.

$$\rho = \frac{\sum_{p \in P} (R_{A_p} - \bar{R}_A)(R_{D_p} - \bar{R}_D)}{\sqrt{\left[\sum_{p \in P} (R_{A_p} - \bar{R}_A)^2 \right] \left[\sum_{p \in P} (R_{D_p} - \bar{R}_D)^2 \right]}} \quad (4)$$

em que:

- ρ representa o coeficiente de correlação de Spearman;
- R_{A_p} representa o posto da nota média de atenção do período p ;
- R_{D_p} representa o posto da nota de desempenho escolar do período p ;
- $\bar{R}_A$ representa a média dos postos das notas de atenção;
- $\bar{R}_D$ representa a média dos postos das notas de desempenho escolar;
- P representa o conjunto dos períodos analisados;
- p representa um período pertencente ao conjunto P .

Como se observa nas Equações 3 e 4, os coeficientes de Pearson e Spearman apresentam formulações semelhantes. Entretanto, enquanto Pearson é calculado diretamente a partir dos valores das notas, Spearman é aplicado aos postos atribuídos a esses valores. Dessa forma, o coeficiente ρ também varia entre -1 e 1 , sendo que valores positivos indicam que postos mais elevados em uma variável tendem a estar

associados a postos mais elevados na outra, enquanto valores negativos indicam uma associação inversa e valores próximos de zero indicam baixa associação monotônica.

A interpretação da intensidade dos coeficientes de correlação seguiu os critérios propostos por Cohen (1988), apresentados no Quadro 6. Essa classificação foi utilizada como referência para caracterizar a associação entre atenção e desempenho escolar como fraca, moderada ou forte. Contudo, os resultados foram interpretados com cautela, considerando o caráter exploratório da pesquisa, o número de pares analisados e a natureza multifatorial do desempenho escolar.

Quadro 6 - Classificação da intensidade das correlações

Intensidade da Correlação	Valor
Fraco	$0,10 \leq c < 0,30$
Moderado	$0,30 \leq c < 0,50$
Forte	$ c \geq 0,50$

Fonte: Adaptado de Cohen (1988)

Além dos coeficientes de correlação, foram considerados os respectivos p-valores, os quais indicam a probabilidade de se observar um resultado igual ou mais extremo do que o encontrado, assumindo como verdadeira a hipótese nula de ausência de associação entre as variáveis analisadas, isso é, a chance de se obter resultados similares caso não houvesse associação real entre as variáveis.

Para essa análise, adotou-se o nível de significância de 5% e teste bicaudal, uma vez que o objetivo foi verificar a existência de associação entre atenção e desempenho escolar, independentemente da direção dessa associação. Assim, p-valores inferiores a 0,05 foram interpretados como indicativos de correlação estatisticamente significativa, enquanto valores iguais ou superiores a esse limite indicaram ausência de evidência estatística suficiente para rejeitar a hipótese nula (Field, 2018).

O cálculo do p-valor foi realizado a partir da transformação do coeficiente de correlação em uma estatística de teste, considerando o número de pares de dados analisados, conforme apresentado na Equação 5.

$$t = c \sqrt{\frac{n-2}{1-c^2}} \quad (5)$$

em que:

- t representa a estatística de teste;
- c representa o coeficiente de correlação analisado, r ou ρ ;
- n representa o número total de pares de dados analisados.

Como a base final foi composta por 30 pares de dados ($n = 30$), foram considerados 28 graus de liberdade ($n - 2$). Após a obtenção da estatística de teste, os respectivos p-valores foram calculados com apoio do Google Planilhas, a partir da distribuição t de Student bicaudal. Embora seja possível consultar manualmente tabelas estatísticas dessa distribuição para verificar faixas de significância, esse procedimento geralmente permite apenas identificar intervalos aproximados de probabilidade. Por esse motivo, optou-se pelo uso do cálculo computacional, que possibilitou obter os p-valores utilizados na interpretação dos resultados.

3.4.3 Outras Abordagens Avaliadas

Durante o desenvolvimento da pesquisa, foram realizadas tentativas adicionais com o objetivo de aprimorar a qualidade das detecções e avaliar alternativas metodológicas para o processamento das imagens. Essas abordagens, embora não tenham sido incorporadas à versão final da análise, contribuíram para o refinamento das decisões técnicas adotadas no estudo.

Uma das primeiras alternativas testadas consistiu no uso de um modelo treinado especificamente para a pesquisa, a partir de imagens rotuladas em ambiente Roboflow. A proposta dessa abordagem era construir um detector mais ajustado às características do contexto analisado, buscando um reconhecimento mais específico dos padrões presentes nas imagens da sala de aula. Entretanto, os resultados obtidos com esse modelo não foram satisfatórios em comparação aos resultados dos modelos pré-treinados adotados na versão final da pesquisa. Uma possível explicação para esse desempenho está relacionada ao volume e à diversidade dos dados utilizados no treinamento. Enquanto o modelo desenvolvido para este estudo foi treinado com um conjunto limitado de imagens, os modelos pré-treinados utilizados já haviam sido ajustados com bases de dados substancialmente maiores e mais variadas, o que

tende a proporcionar maior capacidade de generalização e robustez diante de diferentes condições de imagem. Dessa forma, optou-se por manter o uso dos modelos pré-treinados, por apresentarem desempenho mais consistente no contexto analisado.

Outra abordagem investigada envolveu a correção da distorção óptica das imagens antes do processamento principal. Considerando que parte dos registros foi obtida com uma câmera de grande angular, foi testada uma etapa prévia de calibração e remoção de distorção por meio de recursos da biblioteca OpenCV. Para isso, foram utilizadas imagens de um padrão do tipo tabuleiro de xadrez, que permitiram estimar os parâmetros intrínsecos da câmera e aplicar a correção geométrica às imagens. O objetivo era reduzir as deformações mais acentuadas, especialmente nas bordas, que poderiam interferir na detecção de pessoas e faces. Embora os resultados tenham apresentado melhora visual em parte das imagens, os testes comparativos indicaram redução na quantidade de detecções após a aplicação da correção. Diante desse resultado, a abordagem não foi incorporada ao fluxo principal de processamento, sendo mantida apenas como uma avaliação exploratória.

Essas tentativas adicionais foram importantes para orientar as decisões metodológicas da pesquisa, uma vez que permitiram comparar alternativas de processamento e verificar, na prática, quais estratégias se mostravam mais adequadas ao conjunto de dados disponível.

4 RESULTADOS

Este capítulo apresenta os resultados obtidos com a aplicação do método descrito no Capítulo 3. Inicialmente, são caracterizados os dados analisados e apresentados os registros consolidados de atenção e desempenho escolar. Em seguida, são examinados o comportamento e a variabilidade das notas ao longo do período analisado, bem como os resultados das correlações de Pearson e Spearman. Por fim, os achados são discutidos com base nos estudos revisados e nas limitações da pesquisa.

4.1 CARACTERIZAÇÃO DOS DADOS ANALISADOS

Esta seção descreve a composição da base utilizada nas análises, destacando o volume de imagens processadas, o recorte temporal e a forma como os registros foram organizados. Essa caracterização permite compreender a estrutura da base antes da apresentação dos valores consolidados de atenção e desempenho escolar, realizada na seção seguinte.

No total, foram 8.341 imagens processadas. A partir delas, foram extraídas informações relacionadas à quantidade de pessoas detectadas na cena, à detecção de faces e à classificação dos níveis de atenção estimados pelo método proposto. Entretanto, as imagens não foram correlacionadas individualmente com as notas escolares, pois os registros visuais eram produzidos continuamente durante as aulas, enquanto os dados de desempenho estavam associados a atividades avaliativas realizadas em momentos específicos.

Dessa forma, a base final foi organizada em 30 pares de dados. Esses registros correspondem às atividades avaliativas consideradas na pesquisa e, para cada atividade, a nota média de atenção foi calculada a partir da média das notas de atenção das imagens das aulas que antecederam aquela atividade, de acordo com a organização temporal dos dados. A amostra foi composta por 65 estudantes do 9º ano do Ensino Fundamental, acompanhados ao longo dos meses de setembro a novembro de 2025. Esse recorte corresponde ao período em que havia disponibilidade simultânea de registros visuais e dados escolares compatíveis com os objetivos da pesquisa. A Tabela 2 apresenta a caracterização geral da base de dados analisada, sintetizando os principais elementos utilizados na pesquisa.

Tabela 2 - Caracterização da base de dados analisada

Descrição	Valor
Total de imagens processadas	8.341
Total de registros temporais de referência	30
Total de estudantes considerados	65
Período analisado	Setembro a novembro de 2025
Total de pares de dados utilizados na análise	30
Escala das notas de atenção	0 a 10
Escala das notas de desempenho	0 a 10

Fonte: Elaborado pelo autor

Conforme apresentado na Tabela 2, as 8.341 imagens correspondem aos registros visuais empregados na geração dos indicadores de atenção, enquanto os 65 estudantes delimitam a amostra do estudo de caso. Os 30 pares consolidados, por sua vez, constituem as unidades efetivamente utilizadas nas análises descritivas e correlacionais.

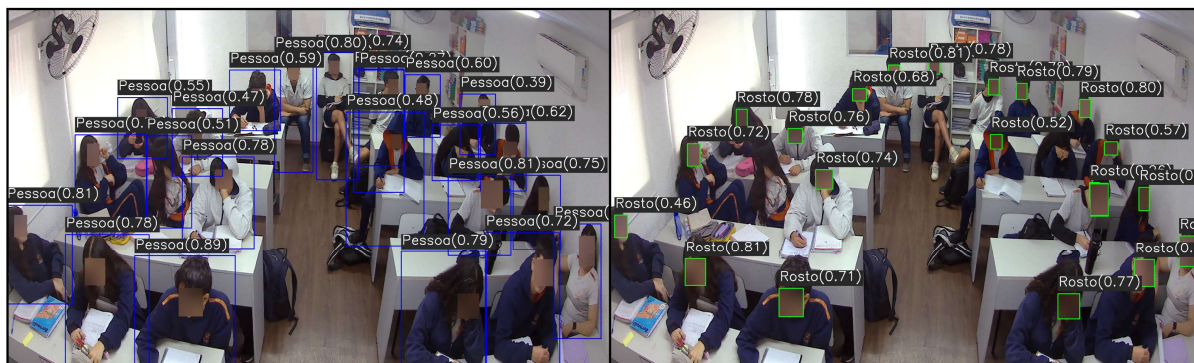
Essa distinção é importante para a interpretação dos resultados, pois demonstra que o volume de imagens representa a fonte dos dados visuais, enquanto os pares de atenção e desempenho correspondem à base final empregada para investigar a associação entre as variáveis.

4.2 CONSOLIDAÇÃO DOS REGISTROS DE ATENÇÃO E DESEMPENHO

Esta seção apresenta os resultados intermediários obtidos entre o processamento das imagens e a formação da base utilizada nas análises estatísticas. São apresentados exemplos das detecções realizadas pelos modelos de visão computacional, da atribuição dos níveis de atenção, dos indicadores gerados por imagem e, por fim, dos pares consolidados de atenção e desempenho escolar.

A Figura 13 apresenta um exemplo do processamento visual realizado em uma imagem real de sala de aula, previamente anonimizada por meio do desfoque das faces dos estudantes. A imagem ilustra a detecção automatizada de pessoas e faces realizada pelos modelos de visão computacional.

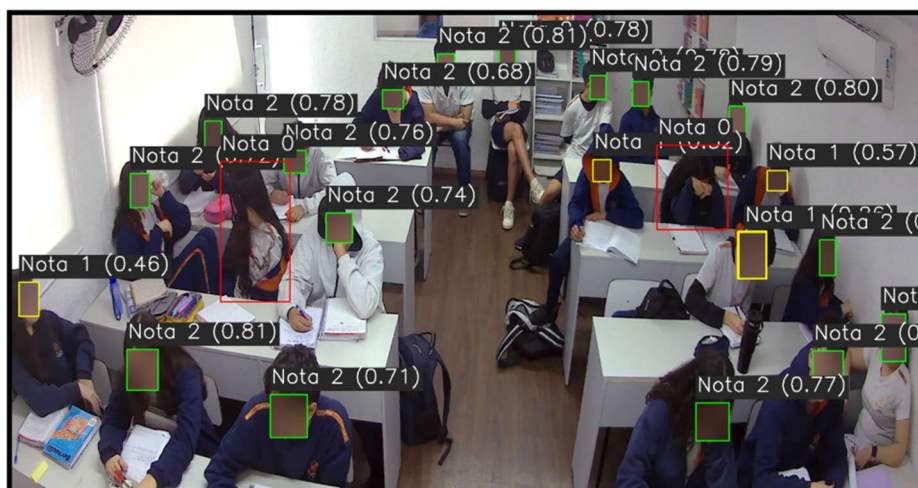
Figura 13 - Exemplo de detecção de pessoas e faces em imagem de sala de aula



Fonte: Elaborado pelo autor.

A partir das detecções realizadas, foi possível atribuir os níveis de atenção conforme os critérios definidos no Quadro 5. A Figura 14 apresenta um exemplo da classificação visual aplicada à imagem processada, evidenciando como os estudantes identificados foram associados aos níveis 0, 1 e 2, de acordo com a presença da face e o valor de confiança da detecção facial.

Figura 14 - Exemplo de atribuição dos níveis de atenção em imagem processada



Fonte: Elaborado pelo autor.

Após o processamento computacional, os resultados obtidos para cada imagem foram organizados em uma planilha eletrônica contendo informações como data, horário, quantidade de estudantes classificados nos três níveis, total de pessoas detectadas e total de faces identificadas. A partir desses dados, foi calculada uma nota de atenção para cada imagem, considerando os pesos atribuídos aos níveis de atenção.

A

Tabela 3 apresenta uma amostra dos indicadores extraídos em diferentes momentos da base processada. Como o conjunto completo é composto por 8.341 registros visuais, optou-se por apresentar apenas um recorte ilustrativo, com o objetivo de demonstrar a estrutura dos dados gerados e a forma como as notas de atenção por imagem foram organizadas antes da consolidação temporal. Ressalta-se que os registros apresentados não correspondem a uma sequência contínua de imagens, mas a exemplos de diferentes momentos do período analisado.

Tabela 3 - Amostra dos indicadores extraídos por imagem para cálculo da nota de atenção

Imagem	Data	Hora	Nível 0	Nível 1	Nível 2	Pessoas Detectadas	Faces Detectadas	Nota de atenção por imagem
1	11/09/2025	09:09:56	2	4	16	22	20	8,18
2	18/09/2025	10:09:37	1	1	15	17	16	9,12
3	26/09/2025	08:44:53	1	3	14	18	17	8,61
4	02/10/2025	10:05:04	1	3	12	16	15	8,44
5	09/10/2025	09:23:11	3	1	17	21	18	8,33
6	16/10/2025	09:24:05	1	2	12	15	14	8,67
7	23/10/2025	09:25:23	2	5	10	17	15	7,35
8	03/11/2025	12:14:17	3	9	11	23	20	6,74
9	13/11/2025	09:41:11	3	1	16	20	17	8,25
10	24/11/2025	09:43:15	1	4	11	16	15	8,12

Fonte: Elaborado pelo autor

O processamento das imagens e a estruturação dos indicadores foram realizados conforme os procedimentos descritos na Seção 3.4.1. Para cada registro visual, a soma dos estudantes classificados nos níveis 0, 1 e 2 corresponde ao total de pessoas detectadas na imagem. A nota de atenção da imagem foi calculada a partir da ponderação desses níveis, conforme a Equação 1.

A variação no número de pessoas detectadas não representa, necessariamente, alteração real na quantidade de estudantes presentes na sala, mas reflete as condições visuais de cada registro, como oclusões, movimentação dos alunos, sobreposição de corpos e limites próprios do modelo. Por esse motivo, a análise final não foi baseada em imagens isoladas, mas na média das notas de atenção de um conjunto de registros visuais associados ao período anterior a cada atividade avaliativa.

Após a consolidação das notas por imagem por meio da média aritmética apresentada na Equação 2, foram obtidas 30 notas médias de atenção. Cada uma dessas notas foi associada à respectiva nota de desempenho escolar, formando os pares utilizados nas análises descritivas e correlacionais.

A Tabela 4 apresenta esses 30 valores, os quais serviram de base para a elaboração do gráfico de evolução temporal das variáveis e, posteriormente, para a aplicação dos coeficientes de correlação de Pearson e Spearman.

Tabela 4 - Registros consolidados de desempenho escolar e atenção

Registro	Nota Desempenho	Nota Atenção
1	6,67	7,80
2	6,09	7,75
3	4,36	7,41
4	9,20	8,14
5	8,00	8,14
6	7,37	7,33
7	7,90	7,33
8	8,33	7,61
9	8,07	7,39
10	8,17	7,38
11	8,42	7,88
12	6,76	7,33
13	6,11	7,81
14	7,50	8,00
15	7,63	8,00
16	8,36	8,29
17	8,24	7,84
18	8,41	7,84
19	7,92	7,19
20	8,14	7,41
21	6,11	7,49
22	3,93	7,59
23	5,28	7,59
24	7,09	7,15
25	4,17	7,15
26	7,29	7,19
27	7,25	6,84
28	9,03	7,37
29	6,67	7,16
30	6,67	6,90

Fonte: Elaborado pelo autor

Os registros apresentados na Tabela 4 são utilizados nas análises descritivas e correlacionais. Na seção seguinte, esses dados são examinados quanto à distribuição e à variabilidade das notas de atenção e de desempenho escolar. Posteriormente, são apresentados os resultados das correlações de Pearson e Spearman.

4.3 ANÁLISE DESCRITIVA DAS NOTAS DE ATENÇÃO E DE DESEMPENHO ESCOLAR

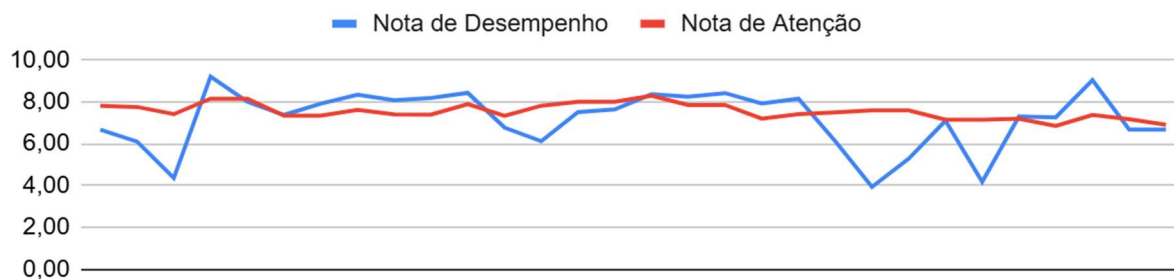
A análise descritiva teve como objetivo examinar o comportamento geral das variáveis investigadas, permitindo uma leitura inicial da distribuição dos dados antes da aplicação dos testes correlacionais.

De forma geral, verificou-se que as notas de atenção apresentaram menor amplitude de variação quando comparadas às notas de desempenho escolar. Enquanto os valores de atenção permaneceram concentrados, em sua maior parte, em uma faixa relativamente estável, as notas de desempenho escolar oscilaram de maneira mais acentuada entre os diferentes registros. Esse comportamento sugere que a atenção, da forma abordada nesta pesquisa, apresentou maior regularidade ao longo do período, quando comparada ao desempenho nas atividades avaliativas.

A Figura 15 apresenta a variação das notas médias de atenção e das notas de desempenho escolar entre os 30 registros consolidados. A representação gráfica permite observar que ambas as variáveis sofreram oscilações, embora não de forma idêntica. Em alguns momentos, nota-se que elevações nas notas de atenção foram acompanhadas por resultados melhores de desempenho, enquanto em outros, a correspondência entre as duas variáveis se mostrou menos evidente. Essa leitura inicial indica que a relação entre atenção e desempenho não se manifesta de forma uniforme em todos os registros, reforçando a necessidade de uma análise estatística complementar.

A Figura 15 apresenta a variação das notas médias de atenção e das notas de desempenho escolar entre os 30 registros consolidados.

Figura 15 - Variação das notas de atenção e de desempenho escolar



Com o objetivo de complementar essa leitura, a Tabela 5 apresenta as estatísticas descritivas das variáveis analisadas, incluindo medidas de tendência central e dispersão. A partir desses indicadores, foi possível verificar o comportamento médio das notas de atenção e das notas de desempenho, bem como sua variabilidade ao longo do período estudado.

Tabela 5 - Estatísticas descritivas das notas de atenção e do desempenho escolar

Variável	Média	Mediana	Desvio-padrão	Mínimo	Máximo
Nota de Atenção	7,54	7,45	0,374	6,84	8,29
Nota de Desempenho	7,17	7,43	1,375	3,93	9,20

Fonte: Elaborado pelo autor

Os valores apresentados na Tabela 5 confirmam a tendência observada graficamente. A nota de atenção apresentou comportamento mais estável ao longo do período, com menor dispersão em relação às notas de desempenho escolar. Por sua vez, o desempenho escolar demonstrou maior oscilação entre os registros, o que pode estar relacionado às diferenças entre as atividades avaliativas realizadas no período. Esses resultados descritivos fundamentam a análise correlacional apresentada na seção seguinte.

4.4 ANÁLISE CORRELACIONAL ENTRE AS NOTAS DE ATENÇÃO E DE DESEMPENHO ESCOLAR

Após a análise descritiva das variáveis, foi realizada a análise correlacional, com o objetivo de verificar a direção e a intensidade da associação entre as notas de

atenção e de desempenho escolar correspondentes aos 30 registros analisados. Para isso, foram calculados por meio das equações 3 e 4 os coeficientes de correlação de Pearson e de Spearman, e posteriormente analisados conforme os procedimentos descritos na Seção 3.4.2.

Os resultados indicaram coeficientes positivos entre as variáveis analisadas. O coeficiente de Pearson foi de $r = 0,285$, com p-valor de 0,127, indicando uma correlação positiva fraca, porém sem significância estatística ao nível de 5%. Já o coeficiente de Spearman foi de $\rho = 0,375$, com p-valor de 0,041, indicando uma correlação positiva moderada e estatisticamente significativa. A Tabela 6 apresenta os resultados obtidos para os dois coeficientes.

Tabela 6 - Resultados da correlação entre atenção e desempenho escolar

Variáveis	Coeficiente	Valor	p-valor	Interpretação
Nota de atenção × Nota de desempenho escolar	Pearson (r)	0,285	0,127	Correlação positiva fraca, não significativa
Nota de atenção × Nota de desempenho escolar	Spearman (ρ)	0,375	0,041	Correlação positiva moderada, significativa

Fonte: Elaborado pelo autor

Os resultados sugerem que maiores notas de atenção tendem a estar associadas a melhores resultados de desempenho escolar, especialmente quando considerados os postos ocupados pelas notas. O coeficiente de Pearson não apresentou evidência estatística suficiente de associação linear entre as variáveis. Por outro lado, o coeficiente de Spearman indicou uma associação monotônica positiva e estatisticamente significativa.

Essa diferença mostra que a associação entre atenção e desempenho foi mais evidente quando analisada a partir da posição relativa das notas do que por meio de uma relação linear entre seus valores originais. Esse resultado é compatível com a representação gráfica apresentada na seção anterior, na qual foram observadas aproximações entre as variações das duas notas em alguns registros, mas sem uma correspondência uniforme ao longo de todo o período.

Dessa forma, os resultados indicam uma associação positiva entre a nota de atenção estimada por visão computacional e o desempenho escolar, com significância estatística observada no coeficiente de Spearman. Ainda assim, essa associação

deve ser interpretada com cautela, considerando o caráter exploratório da pesquisa, o número de pares analisados e o fato de que o desempenho escolar é influenciado por diferentes fatores além da atenção.

4.5 DISCUSSÃO

Os resultados desta pesquisa demonstraram que foi possível transformar registros visuais de sala de aula em indicadores quantitativos de atenção, demonstrando a viabilidade técnica e operacional do método computacional proposto. Esse achado dialoga com a literatura analisada, que aponta a visão computacional como uma abordagem promissora para identificar padrões comportamentais relacionados ao engajamento, à participação e à atenção em ambientes educacionais. Estudos como os de Lin et al. (2021), Sümer et al. (2021), Huang, Cheng e Wu (2022) e Wang et al. (2024) também utilizaram técnicas de detecção visual, reconhecimento facial, estimativa de pose ou análise comportamental para monitorar ações e estados dos estudantes. Nesse sentido, a presente pesquisa corrobora a viabilidade técnica apontada por esses trabalhos ao demonstrar que imagens reais de sala de aula podem ser processadas e convertidas em indicadores utilizáveis em análises educacionais. Como implicação prática, esse resultado sugere que tecnologias baseadas em visão computacional podem apoiar escolas e professores na produção de dados complementares sobre a dinâmica atencional das turmas.

Outro resultado relevante foi a construção de uma heurística baseada na detecção facial para estimar níveis de atenção. Diferentemente de estudos que utilizam múltiplas fontes de informação, como expressões faciais, orientação da cabeça, postura corporal, direção do olhar ou dados multimodais, esta pesquisa adotou uma estratégia mais simples, baseada na presença da face e na confiança da detecção facial. Essa escolha aproxima-se de estudos que utilizam características visuais como indícios de engajamento e atenção, como observado em Sümer et al. (2021) e Wei e Ma (2024), mas também evidencia uma diferença metodológica importante: nesta pesquisa, a atenção não foi medida diretamente, mas estimada por meio de um indicador visual operacional.

A presença e a qualidade da detecção facial podem ser influenciadas por fatores como oclusões, movimentação dos estudantes, posição em relação à câmera, iluminação e enquadramento da imagem. Dessa forma, uma nota elevada representa

maior ocorrência de detecções faciais compatíveis com os critérios definidos pelo método, mas não confirma, por si só, o estado cognitivo de atenção do estudante. Do mesmo modo, a ausência ou a baixa confiança da detecção facial não significa necessariamente que o aluno estivesse desatento. Assim, os indicadores produzidos devem ser compreendidos como aproximações do comportamento visual observado e utilizados como informações complementares, e não como substitutos da avaliação realizada pelo professor.

A análise descritiva indicou que as notas de atenção apresentaram menor variação em comparação às notas de desempenho escolar. Esse resultado é coerente com a compreensão de que o desempenho escolar depende de múltiplos fatores, enquanto o indicador de atenção utilizado nesta pesquisa representa apenas uma dimensão visual e aproximada do comportamento dos estudantes. A literatura sobre comportamento e engajamento em sala de aula aponta que atenção, participação, postura e expressões podem contribuir para a compreensão do processo de aprendizagem, mas não explicam isoladamente os resultados acadêmicos. Nesse sentido, os achados desta pesquisa complementam estudos como os de Liu, Jiang e Jiang (2025), que destacam a necessidade de integrar os dados de comportamento estudantil a elementos pedagógicos e instrucionais mais amplos. Contudo, esse resultado sugere que indicadores de atenção devem ser analisados em conjunto com outras informações escolares, como tipo de atividade, frequência, histórico acadêmico e características das avaliações.

A interpretação dos coeficientes também deve considerar as etapas anteriores de organização dos dados. A consolidação temporal permitiu compatibilizar os registros visuais, produzidos continuamente durante as aulas, com as notas de desempenho vinculadas a atividades avaliativas específicas. Entretanto, o uso de médias pode reduzir variações pontuais da atenção e contribuir para a menor dispersão observada nessa variável. A conversão das notas de atenção para a escala de 0 a 10 facilitou sua interpretação em conjunto com as notas escolares, embora a utilização da mesma escala não seja uma exigência para o cálculo da correlação. A distribuição final das notas depende principalmente dos critérios da heurística, dos pesos atribuídos aos níveis e da forma de consolidação temporal adotada.

No que se refere à correlação entre atenção e desempenho escolar, os resultados indicaram associação positiva entre as variáveis analisadas. O coeficiente de Pearson apontou correlação positiva fraca e sem significância estatística, enquanto

o coeficiente de Spearman indicou correlação positiva moderada e estatisticamente significativa. Esse achado sugere que maiores notas de atenção estimadas por visão computacional tendem a estar associadas a melhores resultados de desempenho escolar, especialmente quando os registros são analisados por sua ordenação relativa. Tal resultado converge com estudos que apontam a atenção e o engajamento como dimensões relevantes para o processo de aprendizagem e para o desempenho dos estudantes (Fredricks, Blumenfeld e Paris, 2004; Fonsêca et al., 2020). Entretanto, os coeficientes obtidos devem ser interpretados considerando o caráter exploratório da pesquisa, o número de pares analisados e as decisões adotadas na construção do indicador de atenção. Na prática, os resultados indicam que o acompanhamento automatizado da atenção pode auxiliar na identificação de tendências gerais da turma, desde que interpretado de forma contextualizada e em conjunto com outros indicadores pedagógicos.

Dessa forma, a principal contribuição desta pesquisa é metodológica e prática, apresentando um caminho computacional para transformar imagens de sala de aula em indicadores visuais de atenção, organizar esses indicadores de acordo com os períodos analisados, associá-los aos dados de desempenho escolar e avaliar a relação entre as variáveis por meio de técnicas estatísticas. A proposta não pretende afirmar que a atenção causa alterações no desempenho, mas apresentar uma forma viável de integrar dados de visão computacional e informações educacionais.

5 CONCLUSÃO

Esta dissertação teve como objetivo avaliar como um recurso educacional baseado em visão computacional pode ser utilizado para estimar a atenção em sala de aula e correlacioná-la com o desempenho escolar dos alunos. Considera-se que esse objetivo foi alcançado, uma vez que foi estruturado e aplicado um procedimento metodológico capaz de transformar registros visuais de sala de aula em indicadores quantitativos de atenção e associá-los aos dados de desempenho escolar disponibilizados pela instituição.

A pesquisa demonstrou a viabilidade da aplicação dos modelos YOLOv11 e YOLO Face no processamento de imagens reais de sala de aula, permitindo a detecção de pessoas e faces e a extração dos valores de confiança utilizados na heurística de classificação dos níveis de atenção. As frequências desses níveis foram convertidas em notas de atenção de 0 a 10 e organizadas em conjunto com as notas de desempenho escolar, resultando em uma base consolidada de 30 pares de dados para a realização das análises descritivas e correlacionais.

Os resultados indicaram associação positiva entre as variáveis analisadas. O coeficiente de Pearson apontou correlação positiva fraca, sem significância estatística, enquanto o coeficiente de Spearman indicou correlação positiva moderada e estatisticamente significativa. Esses achados sugerem que maiores notas de atenção estimadas por visão computacional tendem a estar associadas a melhores resultados de desempenho escolar. Ainda assim, a relação identificada deve ser interpretada com cautela, pois não representa uma relação de causalidade, e o desempenho escolar é um fenômeno multifatorial que não pode ser explicado exclusivamente pela atenção estimada nas imagens.

Como contribuição para o avanço do conhecimento, esta pesquisa apresenta uma proposta aplicada de integração entre visão computacional e análise educacional, contribuindo para preencher uma lacuna identificada na literatura: a necessidade de transformar indicadores visuais de atenção extraídos por recursos computacionais em dados comparáveis ao desempenho escolar. O fluxo metodológico desenvolvido, que envolve o processamento das imagens, a classificação dos níveis de atenção, a conversão desses níveis em notas, a consolidação dos dados e a análise correlacional, constitui uma contribuição

metodológica da pesquisa e demonstra a viabilidade de correlacionar indicadores de atenção estimados por visão computacional com dados escolares.

Do ponto de vista prático e social, os resultados indicam que esses recursos podem apoiar instituições de ensino na produção de indicadores complementares sobre o comportamento atencional dos estudantes. Em contextos de gestão escolar pública ou privada, esses indicadores podem contribuir para o acompanhamento pedagógico e para a identificação de padrões de atenção em sala de aula. Para os professores, podem funcionar como apoio à tomada de decisão, sem substituir a observação docente, mas oferecendo informações adicionais sobre a dinâmica da turma ao longo das atividades. Nesse sentido, a pesquisa reforça o potencial da integração entre tecnologia e educação para auxiliar escolas e professores em situações nas quais a observação da atenção é dificultada pelo número de alunos, pela rotina escolar e pela complexidade das interações em sala de aula.

Apesar das contribuições apresentadas, a pesquisa possui limitações. O estudo foi realizado em uma única instituição privada de ensino, com estudantes do 9º ano do Ensino Fundamental e durante um período específico do ano letivo, o que restringe a generalização dos resultados. Além disso, a atenção foi estimada por meio de uma heurística baseada na detecção de pessoas e faces e nos valores de confiança fornecidos pelos modelos, representando uma aproximação visual do fenômeno, e não uma medida direta do estado cognitivo dos estudantes. As detecções também podem ser influenciadas por fatores como iluminação, oclusões, posicionamento da câmera, enquadramento dos alunos e qualidade das imagens. Por fim, os dados de desempenho escolar foram obtidos a partir das avaliações realizadas pela própria instituição, estando sujeitos às características das atividades, dos conteúdos trabalhados e do contexto pedagógico analisado.

Como sugestão para estudos futuros, recomenda-se ampliar a base de dados, incluindo diferentes turmas, instituições, faixas etárias e contextos escolares. Também seria relevante investigar outros indicadores visuais, como orientação da cabeça, direção do olhar, postura corporal e movimentação dos estudantes, bem como comparar diferentes modelos computacionais quanto à precisão, robustez e aplicabilidade em ambientes reais de sala de aula. Outra possibilidade consiste em integrar os dados de atenção a outras variáveis escolares, como frequência, participação, tipo de atividade, dificuldade da avaliação e histórico acadêmico dos estudantes.

REFERÊNCIAS

ADEIKA, Blessing Isoyiza; ABIODUN, Pelumi Olaitan; OWOLABI, Oludare Adegbola. Transforming Pedagogical Assessment: AI and Computer Vision-Enhanced Classroom Observations for Experiment-Centric Learning Environments. In: **2024 ASEE Annual Conference & Exposition**. 2024.

BARTHOLO, Tiago Lisboa et al. Learning loss and learning inequality during the Covid-19 pandemic. **Ensaio: Avaliação e Políticas Públicas em Educação**, v. 31, n. 119, p. e0223776, 2022.

BOCHKOVSKIY, Alexey; WANG, Chien-Yao; LIAO, Hong-Yuan Mark. Yolov4: Optimal speed and accuracy of object detection. **arXiv preprint arXiv:2004.10934**, 2020.

BOZKIR, Efe et al. Automated Visual Attention Detection using Mobile Eye Tracking in Behavioral Classroom Studies. **arXiv preprint arXiv:2505.07552**, 2025.

BRASIL. Lei Nº 9.394, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional. Brasília, DF: Diário Oficial da União, 1996.

BRASIL. **Ministério da Educação**. Base Nacional Comum Curricular. Brasília: MEC, 2017. Disponível em: <https://basenacionalcomum.mec.gov.br>. Acesso em: 21/05/25.

BULATHWELA, Sahan et al. Could AI democratise education? Socio-technical imaginaries of an edtech revolution. **arXiv preprint arXiv:2112.02034**, 2021.

CATARINO, Marino H. **Redes Neurais**. Freitas Bastos, 2025.

CERVO, Amado Luiz; BERVIAN, Pedro Alcino. Metodologia científica. In: **Metodologia científica**. 1996. p. xiv, 209-xiv, 209.

CHEN, Henghuai; GUAN, Jiansheng. Teacher–student behavior recognition in classroom teaching based on improved YOLO-v4 and Internet of Things technology. **Electronics**, v. 11, n. 23, p. 3998, 2022.

COHEN, Jacob. Statistical Power Analysis for the Behavioral Sciences. 2. ed. New York: **Routledge**, 1988.

EMMER, Edmund T.; STOUGH, Laura M. Classroom management: A critical part of educational psychology, with implications for teacher education. In: **Educational psychology**. Routledge, 2003. p. 103-112.

FIELD, Andy. Discovering Statistics Using IBM SPSS Statistics. 5. ed. Los Angeles: **SAGE**, 2018.

FONSÊCA, Patrícia Nunes da et al. Engajamento escolar e sua relação com as forças de caráter dos adolescentes. **Psicología, Conocimiento y Sociedad**, v. 10, n. 1, p. 160-179, 2020.

FREDRICKS, Jennifer A.; BLUMENFELD, Phyllis C.; PARIS, Alison H. School engagement: Potential of the concept, state of the evidence. **Review of educational research**, v. 74, n. 1, p. 59-109, 2004.

GIL, Antonio Carlos. Métodos e técnicas de pesquisa social. 6. ed. **Editora Atlas SA**, 2008.

GONZALEZ, R. C.; WOODS, R. E. *Processamento Digital de Imagens*. 3. ed. São Paulo: **Pearson Education do Brasil**, 2009.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. Deep learning. Cambridge: **MIT Press**, 2016.

GOTTIPATI, Swapna; SHANKARARAMAN, Venky; JIE, Mark NG Wei. Using Student Perceptions to Design Smart Class Participation Tools: A Technology Framework. In: **2020 IEEE Frontiers in Education Conference (FIE)**. IEEE, 2020. p. 1-9.

HAYKIN, Simon. **Neural networks and learning machines**, 3. ed. Pearson education india, 2009.

HUANG, Yueh-Min; CHENG, An-Yen; WU, Ting-Ting. Analysis of learning behavior of human posture recognition in Maker Education. **Frontiers in Psychology**, v. 13, p. 868487, 2022.

JAIN, Anil K. **Fundamentals of digital image processing**. Prentice-Hall, Inc., 1989.

LIAO, Yongxin et al. Past, present and future of Industry 4.0-a systematic literature review and research agenda proposal. **International journal of production research**, v. 55, n. 12, p. 3609-3629, 2017.

LIN, Feng-Cheng et al. Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection. **Sensors**, v. 21, n. 16, p. 5314, 2021.

LIU, Hai Chuan et al. Campus abnormal behavior recognition with temporal segment transformers. **IEEE Access**, v. 11, p. 38471-38484, 2023.

LIU, Qingtang; JIANG, Xinyu; JIANG, Ruyi. Classroom Behavior Recognition Using Computer Vision: A Systematic Review. **Sensors**, v. 25, n. 2, p. 373, 2025.

MARCONI, Marina de Andrade et al. Técnicas de pesquisa. São Paulo: **Atlas**, 2002.

MARR, David. **Vision: A computational investigation into the human representation and processing of visual information**. MIT press, 2010.

MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, v. 5, p. 115–133, 1943. DOI: 10.1007/BF02478259.

MISSAGLIA, Nelson et al. AMBCON: virtual learning environment connected to smart technologies for student monitoring. **Concilium**, v. 23, n. 13, p. 19-38, 2023. DOI: 10.53660/CLM-1534-23H57.

MITCHELL, T. M. *Machine learning*. New York: McGraw-Hill, 1997.

MOHER, David et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. **Bmj**, v. 339, 2009.

OROSOO, Myagmarsuren et al. Enhancing English Learning Environments Through Real-Time Emotion Detection and Sentiment Analysis. **International Journal of Advanced Computer Science & Applications**, v. 15, n. 7, 2024.

PEDRINI, Hélio; SCHWARTZ, William Robson. **Análise de imagens digitais: princípios, algoritmos e aplicações**. São Paulo: Cengage Learning, 2007. E-book. Disponível em: <https://app.minhabiblioteca.com.br/books/9788522128365>. Acesso em: 17 jul. 2025.

REDMON, Joseph et al. You only look once: Unified, real-time object detection. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2016. p. 779-788.

RIAD, Mustapha; QBADOU, Mohammed; AOULA, Es-Saâdia. Learner's attention detection in connected smart classroom using internet of things and convolutional neural networks. **International Journal of Electrical & Computer Engineering (2088-8708)**, v. 14, n. 3, 2024.

RUSSELL, S. J.; NORVIG, P. *Inteligência artificial*. 4. ed. Rio de Janeiro: Campus, 2022.

SCHWARTZ, Robert M.; SCHMITT, Maribeth C.; LOSE, Mary K. Effects of teacher-student ratio in response to intervention approaches. **The Elementary School Journal**, v. 112, n. 4, p. 547-567, 2012.

SHEN, Yi; SUN, Sixian. Design of International Chinese Education Promotion Platform Based on Artificial Intelligence and Facial Recognition Technology. **Computational Intelligence and Neuroscience**, v. 2022, n. 1, p. 6424984, 2022.

SÜMER, Ömer et al. Multimodal engagement analysis from facial videos in the classroom. **IEEE Transactions on Affective Computing**, v. 14, n. 2, p. 1012-1027, 2021.

SZELISKI, Richard. **Computer vision: algorithms and applications**. Springer Nature, 2022.

UÇAR, Mustafa Uğur; ÖZDEMİR, Ersin. Recognizing students and detecting student engagement with real-time image processing. **Electronics**, v. 11, n. 9, p. 1500, 2022.

ULTRALYTICS. **YOLO Architecture Explained**. [S. l.], 2024. Disponível em: <https://docs.ultralytics.com/guides/yolo-architecture/>. Acesso em: 2 ago. 2025.

VANIN, Anderson Silva. **Otimização do sistema de iluminação pública por meio de visão computacional**. 2021. 88 f. Dissertação (Mestrado em Informática e Gestão do Conhecimento) – Universidade Nove de Julho, São Paulo, 2021.

WANG, Zhifeng et al. SBD-Net: Incorporating Multi-Level Features for an Efficient Detection Network of Student Behavior in Smart Classrooms. **Applied Sciences**, v. 14, n. 18, p. 8357, 2024.

WANG, Zhifeng et al. Students' classroom behavior detection system incorporating deformable detr with swin transformer and light-weight feature pyramid network. **Systems**, v. 11, n. 7, p. 372, 2023.

WEI, Z., & MA, X. (2024). Research on the Informatization Strategy and Practical Effect of Chinese Traditional Culture Education. **Applied Mathematics and Nonlinear Sciences**, 9(1). <https://doi.org/10.2478/amns-2024-3170>

YIN ALBERT, Chuck Chung et al. Identifying and monitoring students' classroom learning behavior based on multisource information. **Mobile Information Systems**, v. 2022, n. 1, p. 9903342, 2022.

YIN, Robert K. **Estudo de Caso-: Planejamento e métodos**. Bookman editora, 2015.

ZHANG, Jian et al. Application of Video Abnormal Behavior Detection Algorithm in Evaluation of Track and Field Teaching and Training Effect. **International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)**, v. 19, n. 1, p. 1-18, 2024.

ZHANG, Qian et al. An Empirical Study on Teaching Management and Quality Feedback Based on Human Behavior Detection. **Journal of Organizational and End User Computing (JOEUC)**, v. 36, n. 1, p. 1-16, 2024.

ZHOU, Qi; SURAWORACHET, Wannapon; CUKUROVA, Mutlu. Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions. **Education and Information Technologies**, v. 29, n. 1, p. 1071-1098, 2024.

APÊNDICE A – CÓDIGO USADO NO PROCESSAMENTO DAS IMAGENS

```
# ===== SETUP =====
from google.colab import drive
drive.mount('/content/drive')

!uv pip install ultralytics

import os, re, cv2
import numpy as np
import pandas as pd
from datetime import datetime
from ultralytics import YOLO
from google.colab.patches import cv2_imshow

# ===== MODELOS =====
model_people = YOLO('yolo11m.pt')

# Baixa o modelo de face apenas se não existir
if not os.path.exists('yolov11m-face.pt'):
    !wget -O yolov11m-face.pt "https://github.com/YapaLab/yolo-
face/releases/download/1.0.0/yolov11m-face.pt"

model_face = YOLO('yolov11m-face.pt')

# ===== UTILIDADES =====

def extrair_data_hora(nome_imagem):
    base = nome_imagem.split('.')[0]
    m = re.search(r'(\d{8})_(\d{6})', base)
    if not m:
        return None, None
    data_str, hora_str = m.groups()
    data = datetime.strptime(data_str, "%Y%m%d").date()
    hora = datetime.strptime(hora_str, "%H%M%S").time()
    return data, hora

def contar_pessoas(img_bgr):
    r = model_people(img_bgr, classes=[0])[0] # só "person"
    return 0 if (r.bboxes is None) else len(r.bboxes)

def contar_faces_notas(img_bgr, limiar_frontal=0.60):
    """
    Retorna:
        n_faces_total, n_notas1 (<= limiar), n_notas2 (> limiar)
    """
    r = model_face(img_bgr)[0]
    if r.bboxes is None or len(r.bboxes) == 0:
        return 0, 0, 0
```

```

n1 = 0
n2 = 0
for box in r.bboxes:
    conf = float(box.conf[0])
    if conf > limiar_frontal:
        n2 += 1
    else:
        n1 += 1
return len(r.bboxes), n1, n2

# ===== PIPELINE (PASTA -> CSV) =====
pasta_imagens = '/content/drive/MyDrive/12 - Pesquisa
Dissertação/Imagens'
ext_validas = ('.jpg', '.jpeg', '.png')

lista_imagens = sorted([
    os.path.join(pasta_imagens, f)
    for f in os.listdir(pasta_imagens)
    if f.lower().endswith(ext_validas)
])

print("Total de imagens encontradas:", len(lista_imagens))

registros = []

for caminho in lista_imagens:
    nome = os.path.basename(caminho)
    data_foto, hora_foto = extrair_data_hora(nome)

    img = cv2.imread(caminho)
    if img is None:
        continue

    # 1) pessoas
    n_pessoas = contar_pessoas(img)

    # 2) faces -> notas 1 e 2
    n_faces_total, n_notas1, n_notas2 = contar_faces_notas(img,
limiar_frontal=0.60)

    # 3) nota 0 (sem face)
    n_notas0 = max(n_pessoas - n_faces_total, 0)

    registros.append({
        'imagem': nome,
        'data': data_foto,
        'hora': hora_foto,

```

```

        'nota_0': n_notas0,
        'nota_1': n_notas1,
        'nota_2': n_notas2,
        'pessoas_total': n_pessoas,
        'faces_total': n_faces_total,
        'h': img.shape[0],
        'w': img.shape[1]
    })

# Se você parou no debug, não vai ter registros; senão, salva CSV
if len(registros) > 0:
    df = pd.DataFrame(registros)
    print(df.head())

    caminho_csv = os.path.join(pasta_imagens,
    'indices_atencao_por_imagem(rev2).csv')
    df.to_csv(caminho_csv, index=False)
    print("CSV salvo em:", caminho_csv)

```

APÊNDICE B – TERMO DE CONFIDENCIALIDADE ASSINADO

TERMO DE CONFIDENCIALIDADE

Pelo presente instrumento particular e na melhor forma de direito,

I. COLÉGIO SAPUCAIA LTDA, sociedade limitada sediada na Rua Sapucaia, nº 1123, Mooca, São Paulo/SP, devidamente inscrita no CNPJ sob o nº 04.697.479/0001-18, neste ato representada na forma de seus atos constitutivos, doravante denominada simplesmente “COLÉGIO”; e,

II. MATHEUS MARTINS DORETTO, pessoa física, brasileiro, pesquisador e mestrando do Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho – UNINOVE, inscrito no CPF sob o nº 425.196.718-66, doravante denominado simplesmente “Pesquisador”.

RESOLVEM, as partes (“Partes” ou individual e indistintamente, “Parte”), celebrar o presente Termo de Confidencialidade (“Termo de Confidencialidade”) como condição para a transmissão de informações pelas envolvidas, que se obrigam a manter o mais absoluto sigilo com relação às informações por elas trocadas.

1. Objeto. O presente Termo de Confidencialidade objetiva a proteção às Informações Confidenciais (conforme definido abaixo) relacionadas à pesquisa acadêmica intitulada “Recursos Educacionais Baseados em Visão Computacional: O Impacto da Atenção no Rendimento Escolar”, a serem formalmente disponibilizadas pelo COLÉGIO ao “Pesquisador” e vice-versa, em razão da realização do projeto acadêmico (“Projeto”).
2. Informações Confidenciais. Todas as informações relacionadas a este Termo de Confidencialidade ou adquiridas em seu curso, reveladas pelo COLÉGIO ao “Pesquisador” ou vice-versa, serão consideradas informações confidenciais (“Informações Confidenciais”), conforme definidas abaixo, devendo ser protegidas por ambas as Partes, abrangendo, inclusive, as respectivas empresas controladoras, controladas, coligadas e/ou sob controle comum (“Afiliadas”) que eventualmente tenham acesso às informações confidenciais.

2.1. As Partes se comprometem a manter as Informações Confidenciais em sigilo, utilizando os mesmos mecanismos de cuidado e discrição para evitar a divulgação, publicação ou disseminação dessas informações a qualquer terceiro estranho à relação mantida entre as Partes.

2.2. Informação Confidencial no âmbito deste Termo de Confidencialidade significa, sem se limitar, toda e qualquer informação, patenteada ou não, de natureza técnica, operacional, financeira, comercial, jurídica, acadêmica, estatística ou metodológica, incluindo imagens, fotografias, registros escolares, indicadores de desempenho, documentos, contratos, relatórios, estudos, pareceres e pesquisas transmitidas entre as Partes:

- (i) por qualquer meio físico (e.g., documentos impressos, manuscritos, fac-símile, mensagens eletrônicas via e-mail, fotografias, etc.);
- (ii) por qualquer forma registrada em mídia eletrônica, tal como CD's, DVD's, pen drive ou qualquer outro meio magnético;
- (iii) oralmente; ou
- (iv) aquelas cujo conteúdo da informação torna óbvia a natureza confidencial.

2.3. As Partes reconhecem que são confidenciais, e de propriedade da Parte que as disponibilizou, todas as informações adquiridas sobre o Projeto.

2.4. As obrigações contidas no presente Termo de Confidencialidade não se aplicarão às Informações Confidenciais divulgadas pelas Partes nas seguintes hipóteses:

- (i) informações que encontram-se disponíveis ao público em geral ou tornaram-se, após a sua divulgação, parte do domínio público por meio de publicação ou por outro meio qualquer, sem ter havido culpa por parte de alguma das Partes;
- (ii) informações que já eram do conhecimento da Parte, antes de sua divulgação;
- (iii) informações que foram, após sua divulgação, adquiridas de boa-fé, sem infringir o presente Termo de Confidencialidade ou qualquer outro instrumento de confidencialidade com terceiros;

e/ou

- (iv) informações que já não são mais tratadas como confidenciais pelas Partes.

3. Utilização Informações Confidenciais. As Partes obrigam-se, direta ou indiretamente, a manter sigilo, bem como a limitar a utilização das informações disponibilizadas exclusivamente para fins acadêmicos e científicos relacionados ao Projeto, não utilizando tais Informações Confidenciais em proveito próprio ou alheio.

3.1. As Informações Confidenciais não poderão ser utilizadas para práticas anticoncorrenciais, desleais ou de competição desleal, permanecendo a obrigação válida mesmo após o término do prazo deste Termo de Confidencialidade.

3.2. O COLÉGIO tem ciência de que o Pesquisador poderá realizar um trabalho de pesquisa acadêmica, podendo ter acesso e utilizar as Informações Confidenciais para a sua realização. As Informações Confidenciais poderão ser utilizadas para a realização dessa pesquisa e divulgadas, desde que preservado o anonimato das pessoas naturais e jurídicas envolvidas na coleta das informações.

4. Vigência. O presente termo torna-se válido a partir da data de sua efetiva assinatura por ambas as Partes e perdurará por 5 (cinco) anos após o encerramento da pesquisa acadêmica, independentemente da forma na qual a relação se encerrou.
5. Infração. A inobservância de quaisquer das disposições de confidencialidade estabelecidas no presente Termo de Confidencialidade sujeitará a Parte infratora, bem como o agente causador ou facilitador, por ação ou omissão de quaisquer daqueles relacionados no referido termo, à reparação de todas as perdas e danos materiais sofridos pela outra Parte, inclusive as de ordem moral, bem como as responsabilidades civis e criminais respectivas.
6. Uso da Marca. Nenhuma das Partes divulgará, utilizará ou permitirá o uso de nomes, logotipos, marcas comerciais ou outros dados de identificação da outra Parte sem o consentimento prévio e por escrito da outra Parte.
7. Não Vínculo Trabalhista. O presente termo não estabelece entre as Partes nenhuma forma de sociedade, vínculo trabalhista ou previdenciário, relação de emprego ou responsabilidade solidária ou conjunta.
8. LEI GERAL DE PROTEÇÃO DE DADOS PESSOAIS. As Partes, suas Afiliadas, seus conselheiros, sócios, diretores, prepostos, funcionários, representantes ou terceiros contratados, em comunhão de esforços, comprometem-se a observar a Lei 13.709 de 2018 ("Lei Geral de Proteção de Dados Pessoais - LGPD").

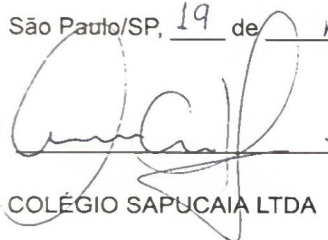
8.1. As Partes, na qualidade de agentes de tratamento, adotarão todas as medidas necessárias para que as operações realizadas durante a vigência deste Termo de Confidencialidade respeitem as diretrizes estipuladas pela LGPD, bem como os seus seguintes princípios: finalidade; adequação;

necessidade; livre acesso; qualidade dos dados; transparência; segurança; prevenção; responsabilização; e prestação de contas.

9. Independência. Se alguma disposição deste Termo de Confidencialidade for considerada inválida em virtude de qualquer lei aplicável ou decisão judicial, tal invalidade não afetará as demais disposições deste instrumento.
10. Irrevogabilidade e Alterações. Este instrumento é celebrado entre as Partes em caráter irrevogável e irretratável, obrigando-se as Partes e seus sucessores a qualquer título, sendo que qualquer alteração do presente instrumento somente produzirá efeitos jurídicos se efetuada por escrito e assinada por ambas as Partes.
11. Assinatura Eletrônica. As Partes desde já acordam que o presente instrumento poderá ser assinado eletronicamente, produzindo todos os efeitos legais aplicáveis.
12. Interpretação e Foro. O presente Termo de Confidencialidade será regido e interpretado de acordo com as leis do Brasil, elegendo as Partes o foro da Comarca de São Paulo/SP para dirimir quaisquer controvérsias oriundas deste instrumento.

E, por estarem assim justas e acordadas, as Partes assinam o presente Termo de Confidencialidade na presença de 02 (duas) testemunhas abaixo assinadas.

São Paulo/SP, 19 de maio de 2026.



COLÉGIO SAPUCAIA LTDA

Matheus M. Doretto

MATHEUS MARTINS DORETTO

Testemunhas:

Nome: Enika Roda Araujo
CPF: 406.982.048-59

Enika Roda

Nome: Rafaello Pereira Roman
CPF: 538.198.198-86

Rafaello P. Roman