Anomaly Detection in Television Digital Channel

Romulo Fabricio[1], Agemilson Pimentel[1], Ruan Belem[1], Anderson Sousa[2], Laura Martinho[2], Leo Araújo[3], Luan Silva[4] and Osmar Silva[2]
[1]TPV Technology Limited, Manaus-AM, Brazil
[2]Institute and Center for Development and Research in Software Technology - ICTS, AM-Brazil
[3]Federal University of Campina Grande (UFCG), PB-Brazil
[4]Federal University of Maranhão (UFMA), MA-Brazil
Emails: {romulo.fabricio, agemilson.pimentel, ruan.belem}@tpv-tech.com
{anderson.souza, laura.martinho, osmar.silva}@grupoicts.com.br
leo.araujo@ee.ufcg.edu.br
luan.souza@discente.ufma.br

***Abstract*—Detecting anomalies in industrial processes is a field in constant advancement. However, automating this task presents significant challenges due to the complexity of the problem. In this paper, Deep Learning techniques were employed to detect anomalies in video footage during digital channel testing of televisions on a production line. A 3D Convolutional Neural Network was trained on a dataset containing two classes of videos: those with simulated defects and those without defects. The resulting model achieved an accuracy of 98,45% with a processing speed of 648 FPS.**

***Index Terms*—Anomaly detection, Deep Learning, 3D Convolutional Neural Networks.**

I. Introdução

A indústria de manufatura tem passado por transformações significativas com a introdução da Indústria 4.0, marcada pela integração de sistemas ciber-físicos, automação e Inteligência Artificial (IA), que visam criar fábricas inteligentes e altamente eficientes [1]. Essa nova era industrial está redefinindo a maneira como os processos de manufatura são conduzidos, buscando não apenas aumentar a eficiência, mas também garantir altos padrões de qualidade.

Um dos maiores desafios que o setor enfrenta é realizar inspeções industriais rápidas e precisas para garantir os mais altos padrões de qualidade a preços competitivos [2]. Nesse contexto, a evolução dos sistemas tradicionais de manufatura para sistemas inteligentes e automatizados é fundamental, visto que desenvolvê-los com o uso de IA pode garantir a excelência dos processos industriais, sendo uma solução relevante [3]. Em problemas que envolvem a identificação e classificação de defeitos, a inspeção visual da qualidade é um importante tópico de pesquisa, e as imagens estão entre os tipos mais comuns de dados tratados.

Vários estudos propuseram soluções apoiadas pelo reconhecimento automatizado de imagens usando aprendizado de máquina para a detecção de defeitos, como a identificação de defeitos de materiais na fusão seletiva a laser de pós metálicos [4] e a classificação de defeitos na fabricação de semicondutores usando imagens de microscópio eletrônico [5]. Embora os diferentes aspectos abordados nos trabalhos que investigam a identificação de defeitos a partir de imagens possam ser extremamente úteis ao lidar com vídeo, a investigação desse tipo de problema usando dados de vídeo apresenta desafios únicos, especialmente ao considerar os padrões espaço-temporais das sequências de dados de entrada.

Anomalias identificadas durante o processo de manufatura são irregularidades na qualidade dos equipamentos produzidos [6]. Neste artigo, é descrito um detector de anomalias em vídeos durante testes de canal digital em televisores. Essa solução é baseada em Redes Neurais Convolucionais 3D para o processamento de vídeos, inspecionando as sequências de imagem reproduzidas pelas televisões e identificando fenômenos defeituosos.

A fim de melhorar o desempenho do sistema e aumentar a taxa de *frames* por segundo (FPS), propõe-se realizar testes nos parâmetros de uma CNN 3D customizada [3], que incluem ajustes finos na arquitetura da rede e na configuração de hiperparâmetros, visando otimizar o processamento dos vídeos sem comprometer a acurácia na detecção de anomalias. Testes com diferentes configurações permitem processar vídeos com uma taxa significativamente maior de FPS, mantendo altas taxas de acerto e proporcionando melhorias significativas na velocidade de processamento.

Entre as contribuições deste artigo destacam-se a aplicação prática de técnicas avançadas de aprendizado de máquina no contexto industrial e a otimização do desempenho do sistema de detecção de anomalias, que atua na melhoria da qualidade e confiabilidade dos produtos e automação dos processos de inspeção. Esse avanço alinha as práticas de manufatura com as expectativas de rapidez e precisão da Indústria 4.0, ao mesmo tempo que abre novos caminhos para pesquisas e desenvolvimentos futuros no campo da IA aplicada.

O artigo está estruturado da seguinte forma: na Seção II, são apresentados os trabalhos relacionados ao contexto da pesquisa. Na Seção III, apresenta-se o sistema proposto. Na seção IV estão os procedimentos experimentais, métricas utilizadas e os resultados obtidos. Por fim, na Seção V, as conclusões do trabalho.

II. Trabalhos Relacionados

Nos ambientes de produção modernos, são necessárias estratégias avançadas e inteligentes de monitoramento de pro-

cessos para permitir um diagnóstico da situação do processo e, portanto, da qualidade do componente final. A detecção de anomalias em vídeos é uma área de pesquisa necessária, marcada pela escassez de dados rotulados e pela necessidade de técnicas avançadas para uma análise eficaz.

Diversas abordagens significativas na literatura empregam técnicas do estado-da-arte para aprimorar a detecção de anomalias em diferentes contextos. O uso de Redes Neurais Convolucionais 3D foi explorado para enfrentar a complexidade dos dados de vídeo e a ambiguidade das anomalias. Nayak et al. [7] investiga arquiteturas de aprendizagem profunda de conjunto com base em redes neurais convolucionais (CNN) e unidades recorrentes controladas (GRU) combinadas com algoritmos de classificação de alto desempenho, como KNN e SVM. Além disso, a análise comparativa dos métodos mais avançados em termos de conjuntos de dados é discutida para descrever os desafios e as direções promissoras para pesquisas no campo de processamento de vídeo.

Por meio da estrutura de classificação profunda de várias instâncias, Sultani et al. [8] propõe um modelo que facilita a detecção sem a necessidade de rótulos detalhados. Em vez de rótulos de treinamento no nível de clipe, os rótulos (anômalos ou normais) são aplicados no nível do vídeo. Essa abordagem considera os vídeos normais e anômalos como pacotes e os segmentos de vídeo como instâncias no aprendizado de várias instâncias (MIL) e aprende automaticamente um modelo de classificação profunda de anomalias que prevê altas pontuações de problemas para segmentos de vídeo defeituosos.

O aprendizado profundo tem promovido soluções promissoras, explorando a capacidade de modelos complexos em identificar padrões irregulares em grandes conjuntos de dados. Ren et al. [9] oferece uma visão compreensiva dos desafios e oportunidades na detecção de anomalias em vídeo, apresentando várias possíveis direções de pesquisa futura do sistema inteligente de detecção de anomalias em vídeo em vários domínios de aplicação. A pesquisa de Zhao et al. [10] propôs um novo modelo chamado *Spatio-Temporal AutoEncoder* (ST AutoEncoder ou STAE), que utiliza redes neurais profundas para aprender a representação de vídeo automaticamente e extrai recursos de dimensões espaciais e temporais por meio de convoluções tridimensionais.

Adicionalmente, Yang et al. [11] introduz uma abordagem inovadora baseada em *keyframes* para restaurar eventos em vídeos de anomalias. Ao propor a restauração de múltiplos *frames* ausentes a partir de *keyframes* de vídeo, essa técnica incentiva redes profundas a explorar e aprender relações contextuais temporais abrangentes e características visuais de alto nível. A arquitetura proposta oferece uma nova forma de restaurar vídeos utilizando atenção cruzada e conexões residuais de *upsampling*. Este avanço destaca a eficácia de restaurações baseadas em eventos no contexto da detecção de anomalias.

No contexto de classificação ou detecção de irregularidades, Luo et al. [12] avança com uma rede de predição de quadros futuros que ajusta rapidamente seu modelo a novas cenas, e Chang et al. [13] explora uma arquitetura de autoencoder que separa as representações espaço-temporais para uma detecção mais eficaz de eventos anormais.

Essas contribuições destacam a importância e o impacto do aprendizado profundo na detecção de anomalias em vídeos, indicando avanços significativos na forma como os sistemas inteligentes podem automatizar e aprimorar as inspeções de qualidade em contextos industriais.

III. Metodologia

A abordagem proposta consiste em um sistema de detecção de anomalias em vídeos utilizado para avaliar canais digitais em televisores fabricados industrialmente. A metodologia emprega Redes Neurais Convolucionais 3D (CNN-3D) para analisar os vídeos exibidos nos aparelhos, visando detectar irregularidades visuais.

A. Base de dados

Para treinar e validar o sistema, foi compilado um conjunto de dados que inclui vídeos de dois tipos: um representando condições defeituosas simuladas e outro sem defeitos, ambos capturados das telas dos televisores durante os testes.

Na linha de produção de televisores, durante o teste do canal digital, quatro tipos principais de anomalias podem ocorrer: mosaico, congelamento, perda de *frames* e tela preta. O vídeo utilizado no teste é composto por 4 cenas de peixes em seu habitat natural, conforme o exemplo na Figura 1. A anomalia de mosaico causa distorções na imagem com formas geométricas, enquanto o congelamento resulta na repetição de *frames*, a perda de *frames* adianta alguns *frames* do vídeo, e a tela preta faz com que alguns *frames* fiquem completamente escuros, como pode ser visto na mesma sequência do vídeo, desta vez afetada pelas anomalias, nas Figuras 2 e 3.

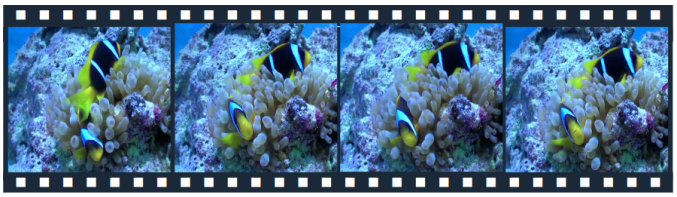

Figura 1. Sequência de *frames* sem defeito

Figura 2. Sequência de *frames* anômala com defeitos de congelamento, tela preta e mosaico simulados.

A base de dados utilizada para o treinamento da rede foi criada especificamente para o teste de detecção de anomalias em canais digitais de televisores. Para aumentar o conjunto de

Figura 3. Sequência de *frames* anômala com defeitos de mosaico e perda de *frames* simulados.

dados reais com defeitos, foram simulados os defeitos através de métodos de processamento de imagem, inseridos em vídeos que foram gravados diretamente de televisores. Como o dataset é sintético, foi fundamental garantir que cada classe tivesse quantidades iguais de amostras, assegurando um equilíbrio entre vídeos defeituosos e sem defeitos para evitar viés na classificação. A coleta foi realizada com gravações feitas a partir de 20 posições de câmera diferentes (10 para cada dispositivo). Para uma representação variada e abrangente dos possíveis defeitos, foram capturados 78 segmentos de vídeo para cada posição de câmera, resultando em um total de 3120 amostras. Essas amostras foram divididas equitativamente entre vídeos com defeitos simulados e vídeos sem defeitos.

B. Sistema Proposto

O sistema de detecção de anomalias foi implementado na linha de produção de uma fábrica de televisores para automatizar a identificação de defeitos no teste do canal digital. O processo começa com a captura dos vídeos da tela da televisão durante o teste de canais digitais, cujas sequências de imagens serão analisadas pelo sistema proposto, que identifica automaticamente qualquer ocorrência de anomalias. Para isso, uma câmera de alta resolução grava a tela da TV enquanto esta reproduz um vídeo padrão de teste transmitido via antena. Tem-se, na Figura 4, uma visão geral do sistema proposto.

C. Arquitetura da Rede

Para a detecção de anomalias, utilizamos uma rede neural convolucional 3D (CNN-3D) [3] como base. A partir dessa arquitetura inicial, realizamos ajustes para adaptá-la ao nosso contexto específico. As entradas foram configuradas com tamanho $128\times128\times3$, o que representa uma redução em relação às dimensões originais dos vídeos ($224 \times 224 \times 3$). A escolha da configuração de 55 *frames* de vídeo com 128×128 pixels em cada *frame* e 3 canais de cor foi motivada pelo objetivo de aumentar o FPS sem comprometer a acurácia do modelo. A arquitetura resultante é representada na Figura 5.

O processo começa com a aplicação de convoluções 3D em várias camadas sucessivas, onde cada camada utiliza filtros tridimensionais que capturam características espaciais e temporais dos vídeos. À medida que o vídeo passa pelas camadas, o número de filtros aumenta, permitindo que a rede extraia padrões cada vez mais complexos. Em cada etapa, os dados são normalizados para acelerar o treinamento e uma função de ativação ReLU [14] é aplicada, introduzindo não-linearidade e permitindo que a rede aprenda representações mais ricas, sendo a operação de convolução representada na Equação 1.

$$\mathbf{X}_{i+1} = f\left(\sum_{k=1}^{K} \mathbf{W}_k^{(i)} * \mathbf{X}_i + \mathbf{b}^{(i)}\right) \tag{1}$$

Sendo $\mathbf{X}_i$ a entrada para a i-ésima camada, $\mathbf{W}_k^{(i)}$ seus filtros tridimensionais, K o número de filtros na camada e $f(\cdot)$ a função de ativação ReLU, a operação de convolução extrai características espaciais e temporais. A saída $\mathbf{X}_{i+1}$ é então passada para a próxima camada da rede, onde o processo se repete com um número crescente de filtros, permitindo a extração de padrões cada vez mais complexos.

Após a extração de características pelas camadas convolucionais, uma camada de *Global Max Pooling* 3D [15] condensa todas as informações extraídas em um único valor por canal. Isso resulta em uma representação compacta e informativa dos dados, para reduzir a dimensionalidade e evitar que o modelo fique excessivamente especializado nos dados de treino (*overfitting*). Além disso, o *Dropout* espacial é utilizado para desligar aleatoriamente mapas de características inteiros durante o treinamento, o que ajuda a rede a generalizar melhor. Esta representação final é então processada por uma camada totalmente conectada (densa), que combina as características e prepara os dados para a classificação final.

A etapa final é realizada por uma camada de saída com 2 neurônios, que usa a função softmax para prever a classe do vídeo, indicando se ele é normal ou apresenta anomalias. Na Equação 2, $\mathbf{p}_c$ representa a probabilidade prevista para a classe c após a aplicação da função *softmax* na camada de saída da rede neural.

$$\mathbf{p}_c = \frac{\exp(\mathbf{z}_c)}{\sum_{j=1}^{C} \exp(\mathbf{z}_j)} \tag{2}$$

Este valor $\mathbf{z}_c$ corresponde à ativação do neurônio c na camada de saída, que reflete a força da evidência de que a entrada pertence à classe c.

D. Treinamento

Para treinar o modelo proposto, o dataset simulado foi particionado na proporção 4:3:3 para treino, validação e teste, respectivamente. O treinamento foi conduzido por 100 épocas, utilizando o otimizador Adam com uma taxa de aprendizagem inicial de 0,001 e aplicou *dropout* entre 30% e 50% para evitar *overfitting*. Além disso, foram realizadas mais de 30 execuções de treinamento, ajustando hiperparâmetros para otimizar a performance do modelo.

Três *callbacks* foram empregados para garantir a eficiência do treinamento. O *ModelCheckpoint* monitorava a acurácia de validação e salvava os pesos do modelo sempre que havia melhoria, assegurando que a melhor versão do modelo fosse preservada. O *ReduceLROnPlateau* utilizou um parâmetro de *patience* de 10 épocas, significando que a taxa de aprendizagem era reduzida se não houvesse melhoria na acurácia de

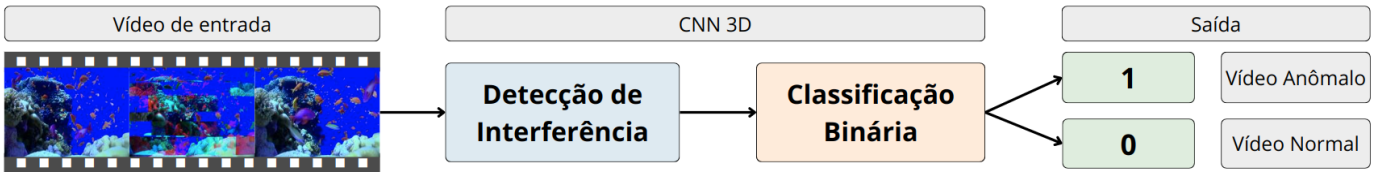

Figura 4. Fluxograma do sistema proposto.

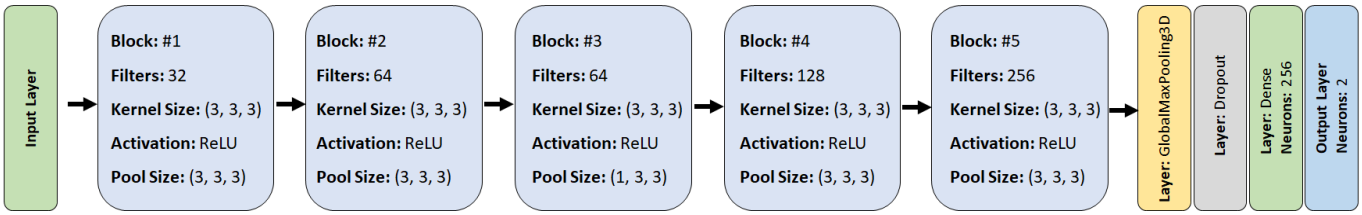

Figura 5. Arquitetura dos modelos treinados construída.

validação durante 10 épocas consecutivas, permitindo refinamentos progressivos. O *EarlyStopping* interrompia o treinamento ao detectar a estagnação da acurácia de validação, prevenindo *overfitting* e economizando recursos computacionais, após 15 épocas consecutivas (*patience* de 15 épocas).

Para aumentar a confiabilidade dos resultados, foi utilizada validação cruzada (k-fold), variando os dados de teste a cada iteração. As principais métricas de avaliação incluíram acurácia, erro (*loss*) e eficiência em termos de *frames* por segundo (FPS) durante o processamento.

IV. Experimentos

Para treinar e validar o sistema, foi utilizado o conjunto de dados capturados das telas dos televisores. A configuração de $128 \times 128 \times 3$ foi escolhida após diversos testes, nos quais constatou-se que, além de manter uma acurácia superior a 98%, essa configuração permite processar vídeos a uma taxa de 648 FPS, significativamente superior a outras arquiteturas testadas. Essa otimização foi possível através de ajustes nos parâmetros da rede, como o número de filtros, a taxa de *dropout* e o uso de *pooling* global, garantindo um modelo robusto e eficiente para a detecção de anomalias em tempo real. Seis modelos foram testados exaustivamente para determinar qual arquitetura oferecia o melhor equilíbrio entre precisão na detecção de anomalias e eficiência de processamento (FPS). Dentre eles estão a CNN-3D customizada, C3D [16], MoViNet [17], 3D ResNet [18], Auto Enconder 3D [10] e a 3D-GAN [19].

Após a fase de construção, treinamento, validação e teste das arquiteturas modeladas, uma variedade de métricas foi aplicada para avaliar e comparar o desempenho de cada uma. As métricas escolhidas foram: Acurácia (ACC), F1 *Score*, Área sob a curva ROC (AUC) e FPS. Os resultados dessa avaliação de ACC e F1 *Score* estão apresentados na Tabela I, e os resultados de AUC e FPS na Tabela II.

Dentre os modelos avaliados, a arquitetura que emprega CNN 3D Customizadas mostrou-se superior, evidenciando

Tabela I
Resultados de ACC e F1 Score

Arquitetura	ACC (%)	F1 Score (%)
C3D	95,79	95,72
MoViNet	85,83	83,33
3D ResNet	98,26	98,23
AutoEncoder3D	68,97	60,46
3D-GAN	94,50	94,37
Cust. 3D-CNN	**98,43**	**98,43**

Tabela II
Resultados de AUC e FPS

Arquitetura	AUC (%)	FPS
C3D	95,78	181
MoViNet	85,70	198
3D ResNet	98,25	217
AutoEncoder3D	68,92	122
3D-GAN	95,30	237
Cust. 3D-CNN	**98,45**	**648**

uma acurácia excepcional de 98,43%. Este resultado mostra a precisão do modelo e destaca sua capacidade de processamento rápido. Com *frames* de 128×128 pixels, essa arquitetura consegue processar 648 *frames* por segundo. Este desempenho é maior do que o alcançado pelas arquiteturas testadas, uma diferença significativa que demonstra a eficiência do modelo proposto em termos de velocidade de processamento.

A Tabela III apresenta ainda uma comparação de desempenho entre diferentes arquiteturas de redes neurais convolucionais 3D utilizando dois tamanhos de entrada distintos: 224×224 e 128×128 pixels. O objetivo desta comparação foi mostrar que os testes na configuração oferecem o melhor equilíbrio entre acurácia e eficiência computacional, mesmo

que testados na mesma rede. Os resultados indicam que a 3D-CNN com entrada de 128×128 apresentou um desempenho superior, alcançando uma acurácia de 98,30% e um FPS de 692, enquanto a configuração de 224×224 atingiu 95,79% de acurácia com um FPS de 181.

Tabela III
COMPARAÇÃO DE DESEMPENHO ENTRE DIFERENTES TAMANHOS DE ENTRADA.

Arquitetura	Tamanho do Input	Acurácia (%)	FPS
3D-CNN	224×224	95,79	181
3D-CNN	128×128	98,30	692

As curvas, observados nas Figuras 6 e 7, também mostram uma diminuição consistente da perda e um aumento na acurácia ao longo das épocas para o modelo com entrada de 224×224 pixels. No entanto, é perceptível que, embora a acurácia final do modelo com 224×224 seja alta, o processo de treinamento é menos eficiente, com maior oscilação na perda e na acurácia durante as primeiras épocas.

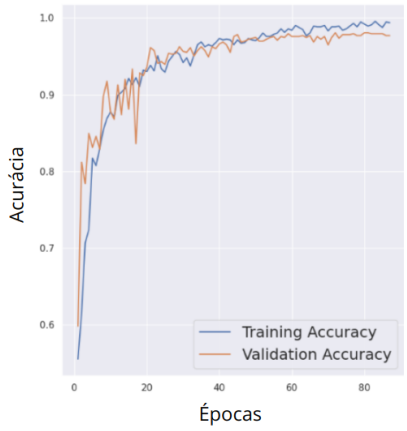

Figura 6. Curva de acurácia do modelo de CNN 3D 224×224

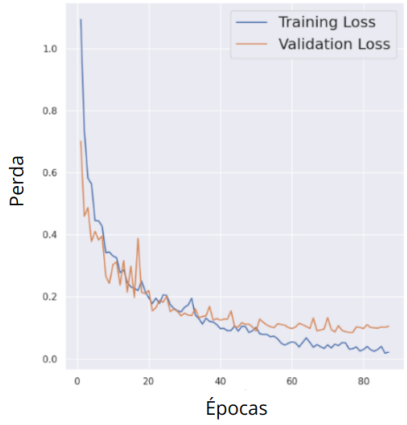

Figura 7. Curva de perda do modelo de CNN 3D 224×224

A Figura 8 apresenta as curvas de perda e a Figura 9 apresenta as curvas de acurácia ao longo das épocas durante o treinamento do modelo de CNN 3D com entrada de 128×128 pixels. A Figura 8 ilustra a evolução da perda tanto no conjunto de treino quanto no de validação. Observa-se uma diminuição significativa da perda nas primeiras épocas, indicando que o modelo está aprendendo de forma eficaz. A perda se estabiliza em valores baixos, com as curvas de treino e validação mantendo proximidade, o que sugere que o modelo não está sofrendo de overfitting. Na Figura 9 são apresentadas as curvas de acurácia para os conjuntos de treino e validação.

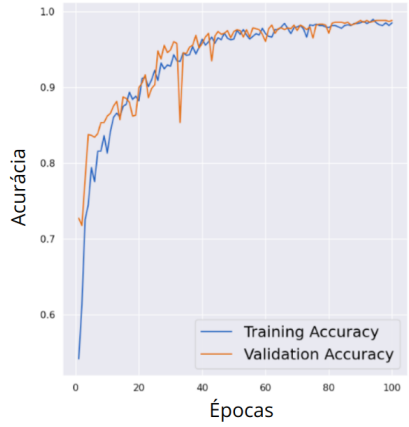

Figura 8. Curva de acurácia do modelo de CNN 3D 128×128

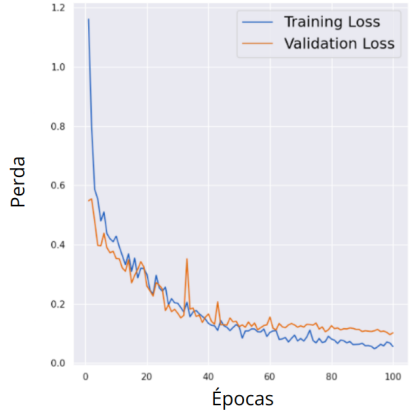

Figura 9. Curva de perda do modelo de CNN 3D 128×128

A acurácia aumenta rapidamente nas primeiras épocas e continua a melhorar em ambos os conjuntos. A semelhança entre as curvas de treino e validação reforça que o modelo está generalizando bem, aprende a partir dos dados de treino e performa adequadamente em dados novos que não foram vistos durante o treinamento.

A Tabela IV mostra os resultados do algoritmo de detecção de anomalias em vídeos, variando o número de *frames* perdidos para simular defeitos. O objetivo foi avaliar a capacidade do modelo em lidar com diferentes quantidades de *frames* ausentes. Os resultados indicam que, com poucos *frames* perdidos (1 ou 3), o modelo tem alto índice de falsos negativos (FN), No entanto, a partir de 5 *frames* perdidos, há uma

melhora significativa, com mais de 90% das ocorrências de defeito sendo identificadas quando 7 ou mais *frames* são perdidos. Para 11 *frames* ou mais, a acurácia atinge valores altos em alguns casos, embora a taxa de falsos positivos (FP) ainda seja elevada em certas situações, podendo acurácia diminuir, devido à maior variabilidade nas anomalias e à confusão do modelo com padrões menos regulares.

Tabela IV
RESULTADOS OBTIDOS EM FUNÇÃO DO NÚMERO DE *frames* PERDIDOS POR VÍDEO.

Nº de *Frames*	VP	FP	FN
1	8,33%	4,17%	87,50%
3	57,69%	7,69%	34,62%
5	73,91%	0,00%	26,09%
7	91,30%	0,00%	8,70%
9	88,46%	11,54%	0,00%
11	100,00%	0,00%	0,00%
13	95,65%	4,35%	0,00%
15	91,30%	4,35%	4,35%
17	86,96%	4,35%	8,70%
19	95,65%	4,35%	0,00%
21	87,50%	8,33%	4,17%
23	86,96%	4,35%	8,70%
25	90,91%	0,00%	9,09%

V. CONCLUSÃO

A exploração de novas técnicas de treinamento e a expansão dos datasets utilizados para treinamento e validação são aspectos que impulsionam o desempenho de sistemas de detecção de anomalias. A arquitetura CNN-3D Customizada proposta demonstrou uma notável eficácia na detecção de anomalias em vídeos, alcançando uma acurácia de 98,43%. Este resultado destaca a precisão do modelo, sua eficiência operacional e a superioridade das métricas em comparação com outras arquiteturas examinadas.

Além disso, o estudo atual reforça a aplicabilidade das CNNs 3D Customizadas no contexto de grandes bases de dados de vídeos, onde a capacidade de processar e analisar eficientemente grandes volumes de informação é essencial. O desempenho superior do modelo desenvolvido valida a abordagem escolhida e abre caminho para futuras investigações e desenvolvimentos na área. Projetos futuros poderiam focar em ajustes e refinamentos adicionais do modelo, visando a redução de falsos negativos e o aumento ainda maior da taxa de acerto.

Assim, este trabalho sugere que tais tecnologias são particularmente adequadas para ambientes industriais onde a precisão e a confiabilidade são essenciais. A solução proposta para a detecção de anomalias em vídeos avança de maneira significativa na abordagem dos desafios industriais, especialmente ao automatizar a inspeção de qualidade em tempo real. A experiência adquirida no desenvolvimento deste sistema mostra que ajustes na arquitetura e a customização do modelo aumentam a eficiência de processamento sem perda de precisão, abrindo espaço para futuras pesquisas que expandam as técnicas para outras indústrias de inspeção automatizada. Dessa forma, a solução contribui para o avanço tecnológico na detecção e estabelece um ponto de partida para explorar o potencial das redes neurais convolucionais em aplicações práticas.

REFERÊNCIAS

[1] L. H. S. Passos, "A indústria 4.0: fundamentos e principais impactos na economia brasileira," *Revista de Administração e Negócios da Amazônia*, vol. 12, no. 2, pp. 53–63, 2020.

[2] J. Villalba-Diez, D. Schmidt, R. Gevers, J. Ordieres-Meré, M. Buchwitz, and W. Wellbrock, "Deep learning for industrial computer vision quality control in the printing industry 4.0," *Sensors*, vol. 19, no. 18, p. 3987, 2019.

[3] L. A. da Silva, E. M. dos Santos, L. Araújo, N. S. Freire, M. Vasconcelos, R. Giusti, D. Ferreira, A. S. Jesus, A. Pimentel, C. F. Cruz *et al.*, "Spatio-temporal deep learning-based methods for defect detection: An industrial application study case," *Applied Sciences*, vol. 11, no. 22, p. 10861, 2021.

[4] A. Caggiano, J. Zhang, V. Alfieri, F. Caiazzo, R. X. Gao, and R. Teti, "Machine learning-based image processing for on-line defect recognition in additive manufacturing," *CIRP Annals*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:149774461

[5] K. Imoto, T. Nakai, T. Ike, K. Haruki, and Y. Sato, "A cnn-based transfer learning method for defect classification in semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. PP, pp. 1–1, 09 2019.

[6] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2112–2119.

[7] R. Nayak, U. C. Pati, and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," *Image and Vision Computing*, p. 104078, 2020.

[8] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] J. Ren, F. Xia, Y. Liu, and I. Lee, "Deep video anomaly detection: Opportunities and challenges," in *2021 International Conference on Data Mining Workshops (ICDMW)*, 2021, pp. 959–966.

[10] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal autoencoder for video anomaly detection," 2017.

[11] Z. Yang, J. Liu, Z. Wu, P. Wu, and X. Liu, "Video event restoration based on keyframes for video anomaly detection," in *IEEE/CVFCVPR*, 2023, pp. 14 592–14 601.

[12] W. Luo, W. Liu, D. Lian, and S. Gao, "Future frame prediction network for video anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7505–7520, 2022.

[13] Y. Chang, Z. Tu, W. Xie, B. Luo, S. Zhang, H. Sui, and J. Yuan, "Video anomaly detection with spatio-temporal dissociation," *Pattern Recognition*, vol. 122, p. 108213, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320321003940

[14] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[15] "Keras layers globalaveragepooling2d: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: https://www.tensorflow.org/

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," 2015. [Online]. Available: https://arxiv.org/abs/1412.0767

[17] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "Movinets: Mobile video networks for efficient video recognition," 2021. [Online]. Available: https://arxiv.org/abs/2103.11511

[18] R. He, Y. Xiao, X. Lu, S. Zhang, and Y. Liu, "St-3dgmr: Spatio-temporal 3d grouped multiscale resnet network for region-based urban traffic flow prediction," *Information Sciences*, vol. 624, pp. 68–93, 2023.

[19] S. Aigner and M. Körner, "Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing gans," *arXiv preprint arXiv:1810.01325*, 2018.