UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE TECNOLOGIA E GEOCIÊNCIA

DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

CURSO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

RAFAEL GOMES PAES DE LIRA

CONSTRUÇÃO DE UM MODELO DE APRENDIZADO POR REFORÇO PROFUNDO PARA UM PROBLEMA DE CONTROLE DE LINHA

Recife

2024

RAFAEL GOMES PAES DE LIRA

CONSTRUÇÃO DE UM MODELO DE APRENDIZADO POR REFORÇO PROFUNDO PARA UM PROBLEMA DE CONTROLE DE LINHA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Produção da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Graduado em Engenharia de Produção.

Orientador (a): Márcio José das Chagas Moura, DSc

Recife

2024

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Lira, Rafael Gomes Paes de.
Construção de um modelo de aprendizado por reforço profundo para um problema de controle de linha / Rafael Gomes Paes de Lira. - Recife, 2024.
51 p. : il., tab.

Orientador(a): Márcio José das Chagas Moura
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, Engenharia de Produção - Bacharelado, 2024.

1. Deep Reinforcement Learning. 2. simulação de linha de produção. 3. Deep Q-learning. 4. inovação tecnológica. I. Moura, Márcio José das Chagas. (Orientação). II. Título.

620 CDD (22.ed.)

RAFAEL GOMES PAES DE LIRA

CONSTRUÇÃO DE UM MODELO DE APRENDIZADO POR REFORÇO PROFUNDO PARA UM PROBLEMA DE CONTROLE DE LINHA

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Produção da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Graduado em Engenharia de Produção.

Aprovado em: 18/03/2024

BANCA EXAMINADORA

_____

Prof. DSc. Márcio José das Chagas Moura (Orientador)

Universidade Federal de Pernambuco

_____

Profa. Dra. Eduarda Asfora Frej (Examinador Interno)

Universidade Federal de Pernambuco

_____

Prof. Dr. Anderson Lucas Carneiro de Lima da Silva (Examinador Interno)

Universidade Federal de Pernambuco

AGRADECIMENTOS

Em primeiro lugar, expresso minha gratidão a Deus, pois acredito firmemente que sem Ele nada disto seria possível. À minha família, que sempre me ofereceu apoio incondicional, e à minha namorada e companheira, Joanne, por ter compartilhado comigo os desafios destes últimos meses tão intensos. Agradeço também aos meus amigos e colegas de trabalho Leandro, Lucas e William, pela constante troca de conhecimentos e por tornarem os dias de trabalho mais leves e alegres. Não posso deixar de reconhecer o Instituto de Inovação Tecnológica por proporcionar um ambiente colaborativo e inspirador.

Sou imensamente grato ao meu orientador, Márcio, pelos ensinamentos valiosos que me proporcionou, também como professor, durante a graduação. Aproveito para estender meus agradecimentos aos demais professores da graduação, cujo conhecimento compartilhado foi fundamental para o meu crescimento acadêmico.

Quero reconhecer também o apoio imprescindível que recebi ao longo da graduação, seja de Yan, Giovana, Paulo Victor ou Thassyo. Aos amigos que sempre estiveram ao meu lado, desde os primeiros anos de colégio, Bianca e Victor, assim como Brian e Beatriz, cuja amizade e apoio foram incomparáveis durante o ensino médio. Meus agradecimentos se estendem aos amigos de infância e irmãos de coração, Hugo e Carlos.

Agradeço profundamente a Danillo, Tamires e Paulinha pelo apoio inestimável nestes últimos meses. E a todos aqueles que, de alguma forma, cruzaram o meu caminho e contribuíram para o meu crescimento, seja através de amor ou aprendizado, expresso meus melhores votos para o seu futuro.

RESUMO

Este estudo aborda a aplicação inovadora do Aprendizado por Reforço Profundo (Deep Reinforcement Learning - DRL) na otimização do controle de produção de embalagens de alumínio em uma empresa multinacional líder. Explorando a interseção entre inovação, tecnologia e produção, o trabalho concentra-se na integração do DRL com a simulação da linha de produção, visando aumentar a eficiência, reduzir paradas não planejadas e promover inovações na indústria. Com base em testes e experimentações, o modelo de DRL revelou desafios significativos em relação à convergência e desempenho em comparação com o baseline estabelecido pelo ambiente de simulação, destacando a complexidade dos problemas enfrentados na modulação da linha de produção e a necessidade de considerar uma abordagem mais abstrata na definição de parâmetros. Além disso, surgiram questões sobre a validação do ambiente de simulação e a lógica de auto-restart, sugerindo a importância de estudos mais aprofundados para garantir a eficácia das soluções propostas. Apesar dos desafios encontrados, o estudo contribui para o avanço do conhecimento em aprendizado de máquina aplicado à manufatura, destacando a necessidade de considerar cuidadosamente os elementos do ambiente de produção ao desenvolver e implementar soluções inovadoras.

Palavras-chave: Deep Reinforcement Learning, simulação de linha de produção, Deep Q-learning, inovação tecnológica.

ABSTRACT

This study explores the innovative application of Deep Reinforcement Learning (DRL) in optimizing the production control of aluminum packaging in a leading multinational company. Focusing on the intersection of innovation, technology, and manufacturing, the research integrates DRL with production line simulation to enhance efficiency, reduce unplanned downtime, and foster industry innovations. Through experimentation, the DRL model revealed significant challenges regarding convergence and performance compared to the established baseline simulation environment, highlighting the complexity of production line modulation and the need for a more abstract approach in parameter definition. Issues concerning simulation environment validation and auto-restart logic also emerged, suggesting the importance of further investigation to ensure the effectiveness of proposed solutions. Despite encountered challenges, the study contributes to advancing knowledge in machine learning applied to manufacturing, underscoring the need to carefully consider production environment elements when developing and implementing innovative solutions.

Keywords: Deep Reinforcement Learning, production line simulation, Deep Q-Learning, technological innovation.

SUMÁRIO

1. INTRODUÇÃO

Este estudo se concentra na interseção vital entre inovação, tecnologia e otimização da produção em uma empresa multinacional de grande porte, líder na fabricação global de embalagens de alumínio para bebidas. A empresa enfrenta desafios em sua busca contínua por aprimorar a eficiência da produção, apesar de já operar em um patamar de excelência em suas fábricas. A pesquisa se concentra especificamente em uma das unidades localizadas na América do Sul, explorando as oportunidades e desafios ligados à implementação de tecnologias da Indústria 4.0, visando aprimorar processos, aumentar eficiência, reduzir custos e fomentar a inovação.

No contexto atual, a modulação da linha de produção é baseada em regras empíricas e limitadas, o que resulta em múltiplas paradas das máquinas em cenários atípicos, prejudicando a produtividade e aumentando os custos de manutenção. Com a introdução das tecnologias da Indústria 4.0, como o aprendizado por reforço profundo, abre-se um novo horizonte para aprimorar a modulação da linha, permitindo uma resposta mais adaptável e eficiente às demandas variáveis da produção.

No foco desta investigação, encontra-se o desafio subjacente à transição na produção de uma linha contínua, especialmente quando se trata da fabricação de diferentes tamanhos de latas de alumínio. A empresa produz latas em diversos tamanhos, e o trabalho terá um foco nas latas de 12oz apenas para ilustração, contudo, as descobertas serão aplicáveis a qualquer tamanho de lata. No entanto, a adaptação da linha de produção para diferentes formatos de latas nem sempre é uma tarefa simples. Essas mudanças podem resultar em interrupções não programadas, perdas de produção e paradas na operação.

A abordagem do estudo reside na aplicação do Aprendizado por Reforço Profundo (DRL - Deep Reinforcement Learning) para aprimorar o controle da linha de produção, especificamente pela máquina Body Maker (BM), onde se encontra o gargalo da linha atualmente. O DRL é integrado ao modelo de simulação já utilizado pela empresa, transformando o ambiente de aprendizado em uma representação virtual da própria planta. Isso permite que um agente, no caso o sistema DRL, tome decisões melhores sobre as velocidades das máquinas na linha de produção, com o objetivo de minimizar paradas não planejadas e aumentar a produção global de latas.

Com base nos estudos revisados, destacam-se os seguintes pontos cruciais na área de controle da produção e otimização de processos com o uso do Aprendizado por Reforço Profundo (DRL). Primeiramente, o DRL está sendo empregado com sucesso para lidar com desafios complexos em ambientes de manufatura, como job-shops e linhas de montagem. Pesquisas

demonstram o potencial do DRL na otimização do despacho de pedidos, gerenciamento de restrições de tempo, tomada de decisões para veículos guiados automaticamente (AGVs), rearranjo de lotes de cores, agendamento de produção, e políticas de controle em tempo real. Essas abordagens inovadoras contribuem significativamente para a automação inteligente, eficiência operacional e flexibilidade das linhas de produção, alinhadas com os princípios da Indústria 4.0.

Este estudo, portanto, emerge como uma solução pioneira para um desafio operacional complexo. Ele abraça o potencial da Indústria 4.0, combinando inovação tecnológica com processos industriais tradicionais, a fim de alavancar a capacidade de melhoria da produção. Ao integrar o aprendizado por reforço profundo a um modelo de simulação, a pesquisa explora a convergência entre inteligência artificial e fabricação, mostrando um caminho promissor para aprimorar a eficiência e a competitividade em um ambiente de produção desafiador.

1.1. JUSTIFICATIVA E RELEVÂNCIA

A pesquisa aborda um cenário complexo e desafiador no contexto da indústria de produção, onde uma empresa líder na fabricação de embalagens de alumínio busca otimizar sua linha de produção. Essa pesquisa reside na integração inovadora de conceitos de inteligência artificial avançada, como o DRL, com simulação de processos produtivos. A aplicação do DRL para melhoria da produção, considerando a interconexão complexa de diferentes máquinas e processos, representa um avanço significativo na aplicação prática da aprendizagem de máquina em ambientes industriais. Além disso, ao modelar a linha de produção como um ambiente de aprendizado, a pesquisa contribui para o desenvolvimento de métodos mais robustos e eficazes de melhoria de processos industriais, que podem ser estendidos para outros setores.

Os procedimentos propostos abordam questões práticas cruciais para a empresa em estudo, como a possibilidade de ganho da produção, a redução de paradas não planejadas e o aumento da eficiência e introduzem um paradigma inovador que pode influenciar outras áreas de conhecimento e indústrias. A abordagem de combinar simulação de linha de produção com DRL [SHIUE; LEE; SU, 2018; WASCHNECK *et al*. (2018)] não apenas oferece soluções para desafios específicos enfrentados pela empresa, mas também contribui para o avanço da pesquisa em aprendizado de máquina aplicado à manufatura. Além disso, os resultados obtidos podem ter implicações para a indústria 4.0, onde a integração de tecnologias digitais e inteligentes é fundamental. Em última análise, a pesquisa pode ser uma ponte entre o mundo acadêmico e as necessidades práticas das indústrias, gerando conhecimento que beneficia tanto a ciência quanto a aplicação industrial.

A realização deste estudo traz consigo uma série de benefícios tanto teóricos quanto

práticos. Em termos teóricos, o trabalho contribui para a expansão do conhecimento em aprendizado por reforço profundo, ao aplicar essa abordagem avançada em um contexto específico de manufatura real. Além disso, a integração do DRL com simulação industrial pode servir como base para futuras pesquisas que visem melhorar processos complexos e interconectados em outras indústrias. Do ponto de vista prático, os procedimentos propostos têm o potencial de melhorar significativamente a eficiência e a produtividade da empresa estudada. Além disso, o sucesso dessa abordagem pode estabelecer um precedente para a adoção de soluções inovadoras baseadas em DRL em outras áreas da empresa e em setores industriais mais amplos, acelerando a transformação digital e a busca por eficiência em processos de produção.

1.2. OBJETIVO

Este trabalho tem como finalidade a aplicação inovadora de Aprendizado por Reforço Profundo (DRL) na otimização do controle de produção de embalagens de alumínio em uma empresa multinacional de grande porte. Através da integração do modelo de DRL com a simulação da linha de produção, o objetivo é alcançar uma produção mais eficiente, reduzindo paradas não planejadas e aumentando a capacidade de resposta da linha às mudanças de produção. Espera-se um aumento da eficiência e qualidade da produção, bem como a promoção de inovações na área industrial.

Objetivo Geral:

Desenvolver e testar um modelo de Aprendizado por Reforço Profundo (DRL) integrado ao ambiente de simulação da linha de produção de embalagens de alumínio, com foco na máquina Body Maker, visando melhorar o controle de velocidades das máquinas e minimizar paradas não planejadas, aprimorando assim a eficiência e a produtividade da produção.

Objetivos Específicos:

Nesse contexto, pretende-se, primeiramente, realizar um levantamento dos principais desafios operacionais enfrentados na máquina Body Maker, identificando oportunidades específicas para melhorar o controle das velocidades das máquinas. Em seguida, será desenvolvido um modelo de Aprendizado por Reforço Profundo (DRL) customizado para atender às demandas da produção de embalagens de alumínio, considerando os aspectos únicos da máquina Body Maker e da linha de produção em geral.

Posteriormente, o objetivo é avaliar empiricamente a eficácia desse modelo de DRL na redução das paradas não planejadas e no aumento da produção, utilizando métricas de desempenho pertinentes ao contexto da indústria de embalagens.

Por fim, com base nos resultados obtidos e nas conclusões da análise, serão elaboradas

recomendações práticas e diretrizes para a aplicação efetiva do modelo de DRL na indústria de embalagens de alumínio, com o objetivo de promover uma gestão mais eficiente e inovadora da produção, visando aprimorar continuamente os processos e alcançar níveis superiores de excelência operacional.

1.3. CONTEXTO E DESCRIÇÃO DO PROBLEMA

1.3.1. Contexto do problema

Neste estudo, é abordada uma empresa de grande porte, atuante em escala global como uma multinacional líder na fabricação de embalagens de alumínio para bebidas. O enfoque da pesquisa recai sobre uma das unidades localizadas na América do Sul. Além de sua principal atividade de produção de embalagens para bebidas, a empresa também se dedica à fabricação de outros tipos de produtos.

A empresa possui uma cultura orientada para a inovação, incentivando seus colaboradores a identificarem desafios e oportunidades e buscarem soluções criativas. Ela também está fortemente comprometida com a adoção de iniciativas de transformação digital, incorporando aplicativos como Qlik Sense, Tecnomatix Plant Simulation e outros, para aprimorar sua eficiência e otimizar seus processos. Além disso, a empresa mantém um sólido compromisso com a sustentabilidade, sendo reconhecida por suas práticas e conquistas nessa área. Através do uso estratégico de tecnologia, a organização impulsiona a inovação em seus produtos e procedimentos, mantendo um objetivo contínuo de melhorar a eficiência energética e produtiva, reduzir as emissões de carbono e promover a economia circular.

O foco do estudo está relacionado às áreas de inovação e tecnologia, que desempenham um papel fundamental na busca por soluções de aprimoramento. Contudo, a empresa atualmente enfrenta desafios em propor e implementar novas soluções para otimizar a produção no chão de fábrica. Devido ao seu atual nível operacional elevado, estudos de melhoria não são triviais, mesmo que a liderança de mercado e os processos avançados estejam presentes, tornando difícil alcançar essa melhoria incremental.

Diante dessa situação, a empresa busca explorar projetos de curto e longo prazos relacionados à indústria 4.0, visando manter seu ritmo constante e ascendente de evolução, aplicando tecnologias digitais e inteligentes na cadeia de produção, com o objetivo de aumentar a eficiência, melhorar a qualidade, reduzir custos e promover a inovação. Portanto, a empresa está apostando em projetos que envolvem a implementação dessas tecnologias para impulsionar sua capacidade de otimização da produção.

O objetivo deste trabalho, portanto, é analisar o uso de novas tecnologias para expandir o

conjunto de ferramentas disponíveis na busca pela otimização da produção e aproveitar os benefícios oferecidos por essas abordagens inovadoras.

1.3.2. Descrição do Problema

A empresa em questão trabalha com um sistema de produção contínua e com baixa variedade de produtos, com isso surgem alguns desafios. Quando há necessidade de mudar o tipo de produto fabricado na linha, podem surgir problemas de controle e gestão de processos, pois a linha de produção, normalmente, está otimizada para um determinado tipo de produto, que precisa passar por modificações para atender a um novo tipo, o que resulta em desequilíbrios e dificuldades de controle.

A indústria produz latas de alumínio em seis tamanhos: 7.5oz, 9.1oz, 12oz, 13.9oz, 16oz e 24oz; porém, o foco deste trabalho é a produção de latas de 12oz. A alteração da linha de produção para fabricar latas de um mesmo tamanho, no geral, é relativamente simples, necessitando apenas de pequenos ajustes internos em algumas máquinas para a troca de rótulos, minimizando o impacto nas operações do *line control*.

No entanto, essa simplicidade nem sempre é uma realidade. Em muitas situações, implementar mudanças nas lógicas de operação e controle da linha pode ser um desafio complexo. Essas mudanças podem levar a paradas não planejadas, perdas de produção e até mesmo quebras de máquinas. Isso ocorre porque os processos na linha de produção são interligados por esteiras e isso exige uma série de pequenos ajustes na programação e no controle das máquinas, a fim de equilibrar o fluxo de produtos e evitar interrupções.

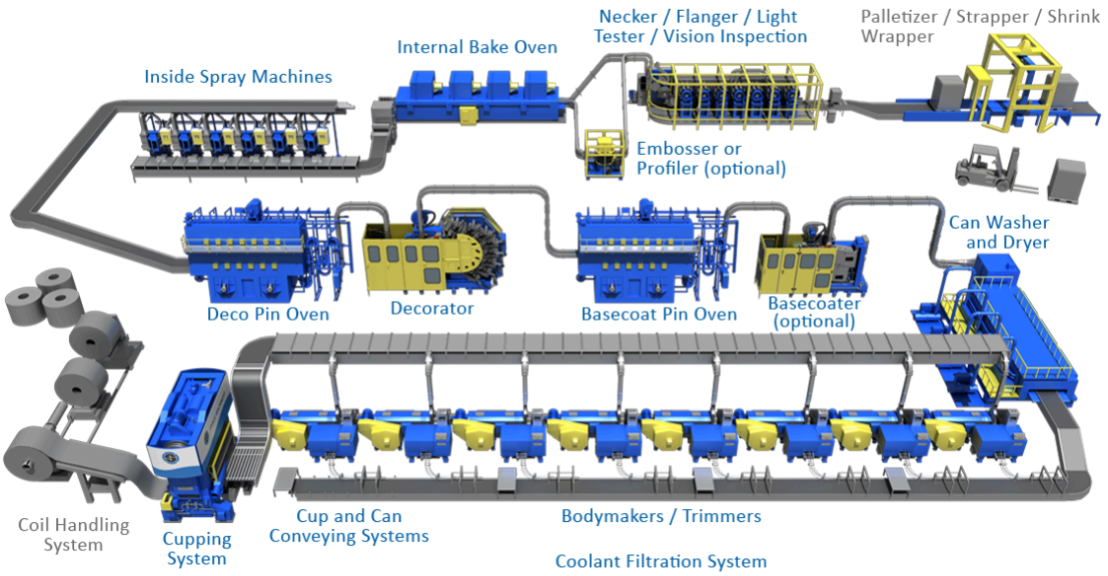
Considerando esse contexto, as decisões relacionadas ao controle da linha de produção e ao balanceamento das operações geralmente são tomadas com base em experiências empíricas dos profissionais responsáveis pelo line control. Contudo, especialmente em momentos de mudanças significativas na produção, essas decisões podem não ser tão simples de serem definidas, devido à complexidade das lógicas de controle da produção e às interligações entre os processos.

Essa variabilidade na definição das regras de balanceamento de linha pode resultar em limitações para abranger todas as restrições específicas da linha de produção. Como consequência, frequentemente são necessários o conhecimento empírico e a intervenção dos operadores para corrigir e criar regras de controle adaptadas a cada situação específica, a fim de evitar prejuízos na eficiência e produtividade da planta.

A Figura 1 abaixo demonstra o processo que segue o seguinte fluxo: começa na Cupper, que é a máquina responsável por cortar a bobina de alumínio em pequenas seções circulares. Em seguida, as seções circulares passam pelas Body Makers, que são 8 máquinas trabalhando simultaneamente para expandir o formato das latas. Após essa etapa, o corpo das latas é

encaminhado para a Washer, onde recebe um tratamento químico de lavagem antes de seguir para a Printer, onde é impresso o rótulo. O passo seguinte consiste na aplicação de uma camada interna de verniz na máquina Inside Spray, para proteger o líquido. Então, a parte superior da lata é colocada na máquina chamada Necker, e, por fim, as latas são embaladas e dispostas em pallets pela Paletizadora.

Figura 1 - Linha de Produção de Latas de Alumínio



Fonte: https://www.stollemachinery.com/pt-br/linhas-para-latas-de-bebidas#stolle-product-group___88, acessado em 10/02/2024

Esses processos são conectados através de esteiras, destacando a importância do controle de linha (line control). Caso uma máquina opere em uma velocidade maior do que a máquina anterior, ocorrerá uma demanda maior de latas do que a máquina anterior é capaz de fornecer, levando a possíveis paradas por falta de abastecimento. Por outro lado, se a máquina estiver fornecendo um grande número de latas em relação à velocidade de operação da máquina seguinte, haverá acúmulo de latas na esteira de saída, o que também pode resultar em paradas na produção.

Ambas as situações são prejudiciais para a operação, podendo gerar ineficiências, queda de latas das esteiras, perda de qualidade das latas, entre outros problemas. Em um cenário ideal, todas as máquinas operariam em velocidades iguais, porém, é na ocorrência de falhas mecânicas que a operação enfrenta desafios significativos. Diante dessa complexidade, este trabalho propõe a adoção de um modelo de aprendizado por reforço profundo (Deep Reinforcement Learning - DRL), visando determinar as velocidades ideais inicialmente das BMs na linha de produção. O objetivo é minimizar a quantidade de paradas não planejadas e aumentar a produção total de latas, enfrentando os desafios decorrentes da variabilidade operacional.

A abordagem do DRL será integrada ao modelo de simulação da linha de produção, que já

é utilizado pela empresa, permitindo assim que o ambiente de aprendizado do DRL seja a própria planta já modelada. Essa integração será realizada através de uma modelagem em Python, permitindo a comunicação em tempo real entre a simulação e o DRL.

No aprendizado por reforço profundo, os componentes essenciais incluem o agente (o sistema de DRL), o ambiente (a planta simulada), os estados (situações em que o agente se encontra), as ações (decisões tomadas pelo agente) e as recompensas (feedback dado ao agente com base nas ações realizadas). Ao longo do texto, será explicado em detalhes como esse modelo de DRL será implementado para lidar com o problema específico de controle da etapa inicial da produção, que será aplicado às BMs, que é o gargalo do line control.

1.4. METODOLOGIA

De acordo com Gil (2022), pesquisa aplicada é voltada à obtenção de conhecimentos visando à sua aplicação em uma situação específica. Diante disso, podemos dizer que esta pesquisa é classificada quanto à finalidade como uma pesquisa aplicada, pois o objetivo é solucionar um problema real e aplicá-lo.

Segundo Gil (2022), pesquisas explicativas são aquelas que se concentram principalmente em descobrir os elementos que influenciam ou participam na ocorrência dos fenômenos. Esse estilo de pesquisa aprofunda-se na compreensão da realidade, pois explora as razões e os motivos por trás dos eventos. Consequentemente, trata-se de uma abordagem mais complexa e delicada, uma vez que a possibilidade de incorrer em equívocos aumenta consideravelmente. Com base nisso, podemos classificar esta pesquisa como explicativa, uma vez que se concentra em descobrir os elementos que podem influenciar na produção e na disponibilidade das máquinas.

A abordagem quantitativa na pesquisa científica concentra-se em medir variáveis de maneira objetiva, usando linguagem matemática e métodos rigorosos. Ela enfatiza a mensurabilidade, causalidade, generalização e replicação. Hipóteses são formuladas, variáveis são mensuradas, dados são analisados estatisticamente e os resultados interpretados. A mensuração é importante, mas não exclusiva dessa abordagem (MORABITO, R. *et al*., 2018). A aplicação de um modelo de RL consiste em elaborar métodos para medir variáveis, analisar dados, simular situações e estabelecer relações causais. Consiste também em avaliar e generalizar os resultados, permitindo tomar decisões mais informadas e otimizar a produção de forma eficiente.

Berto e Nakano (1999) resumem a modelagem como a utilização de métodos matemáticos para descrever o desempenho de um sistema ou de uma parcela de um sistema de produção. Com base nisso, é possível categorizar essa pesquisa como uma modelagem, cujo propósito é desenvolver um modelo de aprendizado por reforço (RL) que represente o desempenho do sistema de produção, permitindo a realização de testes para melhorias subsequentes no desempenho.

Para este estudo, serão utilizadas como técnicas de pesquisa a pesquisa documental, que segundo Lakatos & Andrade Marconi (2010), pode abranger documentos escritos ou não, de fonte primária ou secundária, usados para obter informações históricas e compreender o contexto em estudo. No caso em questão, serão utilizados documentos não escritos de fonte secundária, ou seja, dados coletados pelos sensores da planta e disponibilizados em um banco de dados. Também será utilizada a observação, que é uma técnica de coleta de dados para obter informações por meio dos sentidos (LAKATOS; ANDRADE MARCONI, 2010). A observação será empregada para acompanhar o treinamento do modelo de RL; durante o processo, será possível analisar sua função de recompensa, que determina o progresso do aprendizado do modelo, possibilitando avaliar se o modelo está aprendendo eficazmente durante o treinamento.

Segundo Lakatos & Andrade Marconi (2010), o método dedutivo é um procedimento que, ao partir das teorias e leis, geralmente prevê a ocorrência de fenômenos específicos. Portanto, esta pesquisa utiliza o método dedutivo, considerando que o objetivo é partir de modelos de aprendizado por reforço já estabelecidos e desenvolver um modelo exclusivo para esta pesquisa.

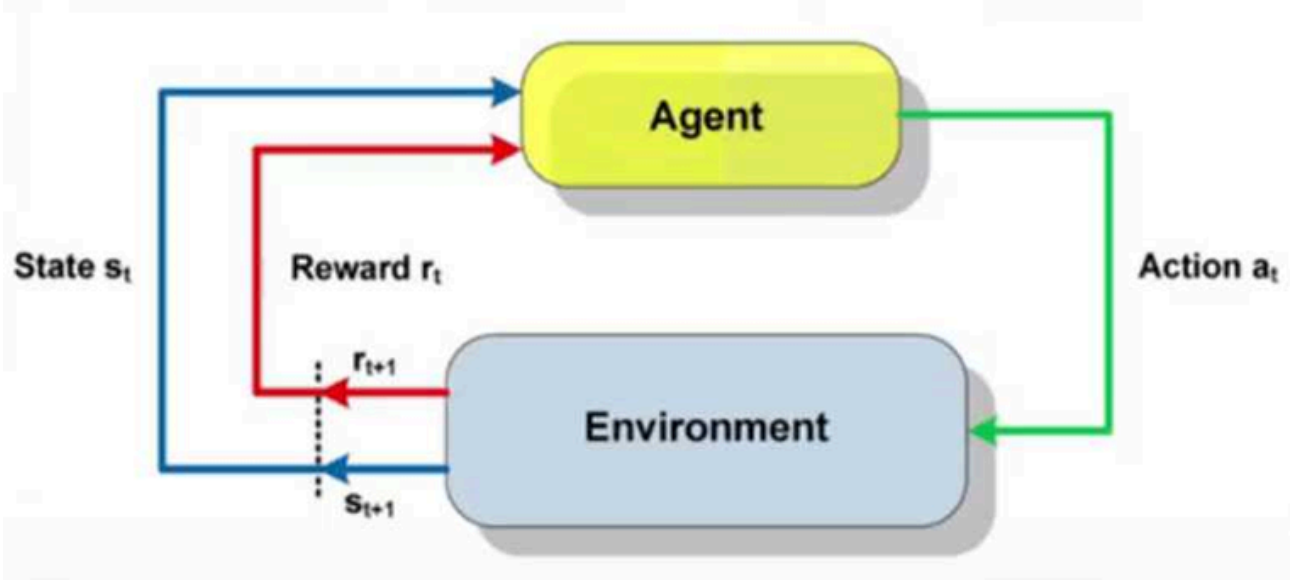
2. FUNDAMENTAÇÃO TEÓRICA

No contexto da otimização do controle de produção de embalagens de alumínio, o uso do Aprendizado por Reforço Profundo (DRL) é fundamental para desenvolver estratégias eficientes de tomada de decisão. O aprendizado por reforço visa encontrar uma política ótima que maximize as recompensas a longo prazo por meio da interação do agente com o ambiente. Este capítulo aborda conceitos-chave do aprendizado por reforço, como o processo de aprendizado em ambientes estocásticos, o equilíbrio entre exploração e exploração, e métodos como Q-Learning e Deep Q-Learning, que desempenham um papel crucial na resolução de problemas complexos e escaláveis como o controle de produção industrial. Ao compreender esses conceitos, torna-se possível explorar abordagens avançadas de aprendizado por reforço para aprimorar o controle e a eficiência das operações de produção.

2.1. REINFORCEMENT LEARNING

Existem diversos trabalhos na literatura que falam sobre o RL. Segundo Agostinelli *et al*. (2018), o aprendizado por reforço tem como objetivo calcular uma estratégia de comportamento, ou política, que busca maximizar um critério de satisfação, representado por uma soma de recompensas a longo prazo, através de interações por tentativa e erro com um ambiente específico. O problema de aprendizado por reforço envolve um agente, chamado tomador de decisões, que opera em um ambiente composto por estados. O agente é capaz de executar ações específicas, dependendo do estado atual. Após selecionar uma ação, o agente recebe uma recompensa escalar e, em seguida, se encontra em um novo estado, que é determinado pelo estado atual e a ação escolhida como mostrado na Figura 2 abaixo:

Figura 2 - Aprendizado por Reforço



Fonte: https://theaisummer.com/Reinforcement_learning/, acesso em 27/08/2023

A cada etapa, o agente segue uma estratégia chamada política, que é uma função que mapeia estados para a probabilidade de escolher cada ação possível. O objetivo do aprendizado por reforço é utilizar as interações do agente com o ambiente para derivar (ou aproximar) uma política ótima, a fim de maximizar a quantidade total de recompensas recebidas pelo agente a longo prazo.

Um dos principais desafios enfrentados no aprendizado por reforço, que não está presente em outros tipos de aprendizado, é o delicado equilíbrio entre *exploration* e *exploitation*. Para maximizar suas recompensas, um agente de aprendizado por reforço precisa preferir ações que já foram testadas no passado e que provaram ser eficazes em obter recompensas. No entanto, para descobrir novas ações potencialmente mais vantajosas, ele também deve se arriscar e experimentar ações que ainda não foram exploradas. O dilema é que se o agente se concentrar apenas no *exploration* ou apenas na *exploitation*, ele pode falhar na tarefa de maximizar suas recompensas. Para lidar com esse dilema, o agente deve realizar uma variedade de ações e, ao longo do tempo, favorecer aquelas que parecem ser mais promissoras em termos de recompensas (SUTTON; BARTO, 2018).

2.2. Q-LEARNING E DEEP Q-LEARNING

2.2.1. **Q-learning**

O método do Q-learning tem se tornado cada vez mais popular para o aprendizado de reforço *off-policy*, ou, livre de modelo (WATKINS; DAYAN, 1992). Trata-se de um algoritmo genuíno de aprendizado, pois gradualmente aprende a função de política ótima interagindo com o ambiente, sem pressupor conhecimento prévio das recompensas ou probabilidades de transição. Em suma, o Q-learning visa aprender uma função Q, que associa o estado atual s com uma ação a e o valor de utilidade Q(s,a). Esse valor prediz a recompensa total futura descontada que se espera receber ao tomar a ação atual a e seguir a política ótima π^* em todas as ações subsequentes. A política ótima π^*, segundo Watkins e Dayan (1992), é determinada pela função Q, sendo

$$\pi^*(s) = argmax_a Q(s,a) \qquad (1)$$

O objetivo do Q-learning é aprender diretamente a função Q, sem a necessidade de explícito conhecimento das probabilidades de transição ou valores esperados das recompensas. Para contextualizar, imagine que conhecemos a função Q e que o ambiente está atualmente no estado s. Nesse cenário, o agente escolhe uma ação a para maximizar Q(s,a). Essa ação leva a uma transição do estado s para o estado s’ no ambiente. Em seguida, o agente escolhe outra ação b para maximizar Q(s’,b).

A recompensa imediata de tomar a ação “a” a partir do estado s, seguindo a política ótima, é representada por Q(s,a), acrescida da utilidade máxima possível do próximo estado s’, descontada pelo fator de desconto γ. Portanto, temos a seguinte equação:

$$Q(s,a) = R(s,a) + \gamma * \sum_{s'} P_{s,s'}(a) * max_b Q(s',b) \quad (2)$$

Nesse contexto, os métodos de diferença temporal (TD), Sutton e Barto (2018), são usados para construir incrementalmente a função Q por meio de estimativas. Durante o processo de aprendizado, a equação (2) pode não ser exatamente verdadeira, e a atualização TD é projetada para mudar a estimativa (s,a) de Q(s,a) em uma direção que reduza a desigualdade. A Equação (3) é um exemplo específico de uma atualização de diferença temporal:

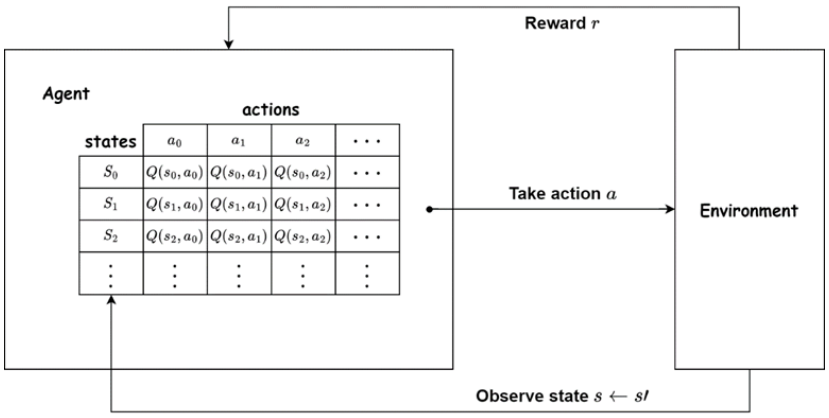
$$Q_{n+1}(s,a) = Q_n(s,a) + \eta * (R(s,a) + \gamma * max_B Q_n(s',b) - Q_n(s,a)) \quad (3)$$

Observe que a Equação (3) também pode ser expressa como:

$$Q_{n+1}(s,a) = (1-\eta) * Q_n(1,a) + \eta * (R(s,a) + \gamma * max_B Q_n(s',b)) \quad (4)$$

Aqui, η é uma taxa de aprendizado que pode ser uma constante pequena ou um valor que tende a zero à medida que o passo de tempo η aumenta. Ao examinar a Equação (3), é possível perceber que o Q-learning realiza um backup de um passo, usando a função Q estimada no estado do próximo passo de tempo. Estudos de Watkins e Dayan (1992) demonstraram que esse procedimento converge sob condições muito frágeis, principalmente quando o estado s’ depende apenas do estado s e da ação tomada enquanto no estado s, e quando Q(s,a) pode ser expresso em forma de tabela (ou seja, s e a pertencem a conjuntos finitos), como demonstrado na Figura 3:

Figura 3 - Q-Learning



Fonte: https://wikidocs.net/174536, acessado em 27/08/2023

Uma característica interessante do Q-learning é que a função Q pode ser atualizada de forma completamente assíncrona, permitindo um alto grau de paralelismo em sua construção. Em outras palavras, é possível dedicar um processador diferente a cada estimativa (s,a) e cada processador responsável pela atualização não precisa coordenar suas ações com os demais processadores.

A maioria dos algoritmos de RL disponíveis podem ser agrupados em duas categorias distintas: as abordagens baseadas em modelos e as abordagens livres de modelos. As abordagens baseadas em modelos têm como objetivo aprender um conjunto de regras subjacentes e dinâmicas do ambiente em questão, buscando assim encontrar uma política ótima para a tomada de decisões. Em contraste, as abordagens livres de modelos buscam determinar a melhor ação para cada estado, seja para construir uma política global que cubra todos os estados ou para atingir rapidamente um estado desejado (ACHIAM; MORALES, 2018; KAELBLING et al., 1996).

O Q-Learning é um exemplo notável de abordagem livre de modelo. Neste método, o agente aprende um valor (chamado de Q-Value) para cada par de estado e ação, permitindo assim a aproximação de uma função de valor $Q(s,a)$ a partir da qual é possível derivar uma política ótima (WATKINS; DAYAN, 1992).

Para sistemas complexos e dinâmicos, especialmente aqueles que enfrentam incerteza estocástica, como os sistemas modulares de produção, que é o problema do trabalho, as abordagens baseadas em modelos frequentemente mostram-se inadequadas. Isso se deve ao fato de que o espaço de estados é muito extenso, a construção do modelo interno do algoritmo pode ser imprecisa e o esforço computacional necessário pode ser inviável (GOSAVI, 2015).

2.2.1.1. Processos de Decisão de Markov

Tan, Yan & Guan (2017) ressaltam que a essência do aprendizado por reforço é um processo de decisão de Markov. O aprendizado por reforço fornece apenas uma função de recompensa, que é a decisão em risco de uma condição específica para executar uma ação (que pode resultar em algo melhor ou possivelmente pior). No entanto, o desempenho do aprendizado por reforço depende da tecnologia de extração de características artificiais, e o aprendizado profundo apenas compensa essa limitação.

Um Processo de Decisão de Markov (MDP) é um modelo de controle estocástico em tempo discreto que fornece uma estrutura matemática para a tomada de decisões em situações em que os resultados não estão totalmente sob o controle do agente (BELLMAN, 1957). Portanto, os MDPs são úteis para resolver problemas de otimização, como o controle de condução, por meio da programação dinâmica (ou seja, DRL) (TALPAERT *et al*., 2019).

O MDP pode ser definido por um conjunto de elementos: o espaço de estados S, o espaço de ações A, o modelo de probabilidade de transição P, a função de recompensa R e o fator de desconto γ. No MDP, o agente executa uma ação e recebe uma recompensa numérica do ambiente com base na função de recompensa R.

2.2.1.2. A Equação de Bellman

É importante lembrar que o agente recebe uma recompensa de feedback r_{t+1} a cada etapa de tempo t até alcançar o estado terminal T. No entanto, essa recompensa imediata r_{t+1} não representa o ganho de longo prazo. Em vez disso, utilizamos um valor de retorno generalizado R_t no momento t, em que γ é um fator de desconto, variando de 0 a 1. Quando γ se aproxima de 1, o agente se torna mais focado no longo prazo, enquanto que quando γ se aproxima de 0, o agente se concentra mais no curto prazo.

$$R_t = r_{t+1} + \gamma * r_{t+2} + \gamma^2 * r_{t+3} + \ldots + \gamma^{T-t-1} * r_T = \sum_{i=0}^{T-t-1} \gamma^i * r_{t+i+1} \quad (5)$$

O próximo passo é definir uma função de valor que é utilizada para avaliar o quão "bom" é um determinado estado s ou um determinado par estado-ação (s, a). Especificamente, a função de valor do estado s sob a política π é calculada obtendo-se o valor de retorno esperado a partir de s: $V_\pi(s) = E[R_t | s_t = s, \pi]$. Da mesma forma, a função de valor do par estado-ação (s, a) é $Q_\pi(s, a) = E[R_t | s_t = s, a_t = a, \pi]$. Utilizamos as funções de valor para comparar o quão "bom" são duas políticas π e π', seguindo a seguinte regra (SUTTON; BARTO, 2018):

$$\pi \leq \pi' \Longleftrightarrow [V_\pi(s) \leq V_{\pi'}(s)\ \forall\ s] \vee [Q_\pi(s, a) \leq Q_{\pi'}(s, a)\ \forall\ (s, a)] \quad (6)$$

Com base na Eq. (6), podemos expandir Vπ (s) e Qπ (s, a) para representar a relação entre dois estados consecutivos s = st e s' = st+1 como (SUTTON; BARTO, 2018):

$$V_\pi(s) = \sum_a \pi(s, a) * \sum_{s'} p(s' | s, a) * (W_{s \to s' | a} + \gamma * V_\pi(s')) \quad (7)$$

E

$$Q_\pi(s, a) = \sum_{s'} p(s' | s, a) * (W_{s \to s | a} + \gamma * \sum_{a'} \pi(s', a') * Q_\pi(s', a') \quad (8)$$

onde $W_{s \to s' | a} = E[r_{t+1} | s_t = s, a_t = a, s_{t+1} = s']$.

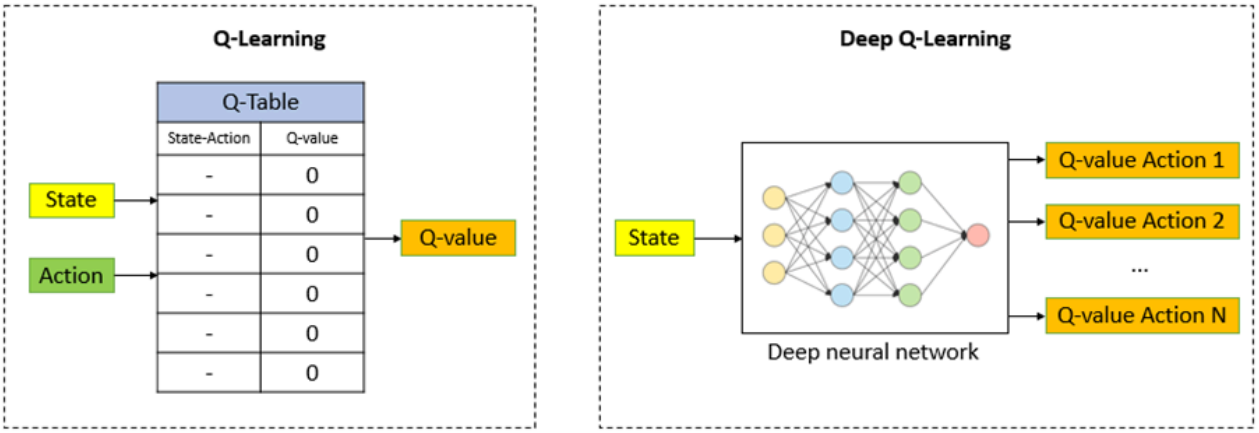
Resolvendo (7) ou (8), podemos encontrar a função de valor V(s) ou Q(s,a),

respectivamente. As Equações (7) e (8) são conhecidas como equações de Bellman.

2.2.2. Deep Q-Learning

Uma extensão apropriada do Q-Learning tradicional é o Deep Q-Learning (DQL), que utiliza uma rede neural convolucional profunda para mapear a relação entre estados e ações, como exemplificado na Figura 4. Essa abordagem possibilita a aplicação do princípio do Q-Learning mesmo em espaços de estado e ação consideravelmente grandes (MNIH et al., 2015; SHRESTHA; MAHMOOD, 2019).

Figura 4 - Q-Learning vs. Deep Q-Learning



Fonte:https://www.baeldung.com/cs/q-learning-vs-deep-q-learning-vs-deep-q-network, acessado em 27/08/2023

De acordo com Alabadi e Albayrak (2020), o controle da escalabilidade em termos de número de estados e ações torna-se desafiador ao utilizar o Q-learning tradicional. Em vez disso, os pesquisadores têm recorrido ao DQL, empregando redes neurais para aproximar a função de valor Q. Em virtude da complexidade do problema em questão, a utilização do Q-learning convencional seria praticamente inviável. Conforme a complexidade do problema aumenta, a Q-table cresce de forma exponencial, resultando em tempos de execução impraticáveis.

Por essa razão, este trabalho adotará o Deep Q-Learning, que se utiliza de uma rede neural para calcular o valor Q para a próxima ação possível. Enquanto em problemas menores a Q-table pode ser vantajosa, uma vez que a construção de uma tabela pequena é mais rápida do que o treinamento de uma rede neural, à medida que o problema se torna mais complexo, essa vantagem se inverte. O Deep Q-Learning se mostra mais adequado para lidar com problemas de maior complexidade, oferecendo uma abordagem mais eficaz e escalável.

3. REVISÃO DE LITERATURA

Estudos recentes vêm sendo desenvolvidos na área de controle da produção, otimização de processos juntamente com o DRL. As oportunidades e desafios da manufatura em tempos de digitalização generalizada levam a uma nova era de gerenciamento de operações. Abordagens de controle inteligente e autônomo de job-shops complexos são aprimorados pelo aprendizado por reforço (ALTENMÜLLER et al., 2020). Este trabalho contribui para pesquisa por meio de uma investigação do despacho de pedidos sob restrições de tempo em job-shops complexos. Os resultados computacionais são baseados em uma simulação de eventos discretos e revelam o alto potencial de um algoritmo Deep Q-learning (DQN). Os resultados mostram que os agentes do RL podem ser aplicados com sucesso para controlar o envio de pedidos em um caso de uso de aplicativo de tamanho real. Além disso, as restrições de tempo são gerenciadas melhor do que os benchmarks estabelecidos em toda a indústria, ou seja, uma heurística que otimiza as restrições de tempo ou uma heurística orientada para first-in, first-out (FIFO).

Atualmente, com o notável crescimento da complexidade e incerteza nos sistemas de manufatura, tem havido uma necessidade urgente de métodos dinâmicos de reescalonamento multiobjetivo com a capacidade de lidar com distúrbios aleatórios, como demandas dinâmicas, quebras de máquinas e tempos de processamento incertos em tempo real e simultaneamente considerando diferentes objetivos como makespan e atraso total (LUO; ZHANG; FAN, 2021). O autor utiliza um modelo de DQN de duas hierarquias, sendo a primeira como controlador para localizar o objetivo de otimização mais alto e os recursos do estado atual como entrada para o segundo DQN, o que comprovou ser um modelo mais eficiente que os modelos clássicos. Fatores bem importantes não foram considerados como quebras de máquinas e tempos de processamento variáveis, o que pode mudar consideravelmente o desempenho do modelo.

Diversos estudos abordam a aplicação do RL em indústrias automotivas. Feldkamp, Bergmann e Strassburger (2020) utilizam o DQN para melhorar a tomada de decisão de veículos guiados automaticamente (AGVs) em uma indústria automotiva. O DQN é treinado para concluir tarefas de produção em sequência, superando regras heurísticas simples.

Leng et al. (2020) resolvem o problema de rearranjo de lotes de cores em oficinas automotivas de pintura, utilizando um modelo de Histograma de Cores (CH) integrado a um algoritmo DQN. O agente DQN é treinado para minimizar os custos de troca de cores entre pedidos de produção consecutivos, e os resultados mostram sua eficácia em comparação com algoritmos convencionais.

Em um esforço para melhorar a flexibilidade das linhas de montagem, especialmente em

cenários de produção de pequenos lotes e múltiplas espécies, Li et al. (2022) apresentam um trabalho que utiliza algoritmos de aprendizado por reforço e modelos de gêmeos digitais. O objetivo é automatizar o planejamento das ações de montagem e monitorar as linhas de produção em tempo real. Para atingir esse objetivo, os pesquisadores desenvolveram um modelo de gêmeo digital da linha de montagem e treinaram um agente de aprendizado por reforço para executar a montagem das peças com base nas informações fornecidas pelo modelo. Na etapa de produção, o gêmeo digital é empregado para monitorar a linha de montagem e prever falhas. O sistema proposto foi validado em um experimento prático de montagem de encaixe de pino em furo, alcançando uma taxa de sucesso de 90% em uma única tentativa de montagem, sem ocorrência de colisões. Essa abordagem inovadora visa alcançar uma montagem mais flexível e eficiente, reduzindo a necessidade de intervenção humana e melhorando a taxa de sucesso nas tarefas de montagem em linhas de produção.

O trabalho de Marchesano et al. (2022) apresenta uma abordagem baseada em Aprendizado por Reforço Profundo (DRL) para o agendamento de produção em linhas de fluxo com restrições de datas de vencimento das tarefas. Nessa pesquisa, os autores propõem o uso de um algoritmo DQN, empregado como um controlador para determinar a posição adequada na qual cada tarefa deve ser inserida na fila de produção. O objetivo principal é criar um algoritmo de agendamento que seja capaz de atender aos prazos das tarefas e, ao mesmo tempo, maximizar a taxa de produção. A abordagem proposta permite flexibilidade, pois não requer a repetição da fase de treinamento caso a configuração do sistema de produção mude. Com isso, busca-se fornecer uma ferramenta de agendamento geral que possa ser utilizada em diversas situações de fabricação, inclusive aquelas imprevistas que podem ocorrer no ambiente de produção.

A pesquisa de Huang, Chang e Chakraborty (2019) se concentra na resolução do problema de substituição preventiva de máquinas em linhas de produção seriadas. Com o objetivo de encontrar uma política ótima de substituição das máquinas que leve em consideração a perda de produção avaliada nas linhas de produção, os autores formulam o problema como um problema de aprendizado por reforço e utilizam o algoritmo Q-learning para resolvê-lo. A modelagem baseada em dados históricos das linhas de produção é utilizada durante o treinamento do agente. A pesquisa apresenta uma abordagem que considera o estado de envelhecimento das máquinas e o estado das linhas de produção para obter políticas de manutenção ótimas. O objetivo final é fornecer um método eficiente e efetivo de substituição preventiva de máquinas em linhas de produção, e o estudo é avaliado por meio de simulações que demonstram a viabilidade da abordagem proposta.

Zheng, Lei e Chang (2017) propõem um estudo abrangente sobre a política de controle em tempo real de sistemas de produção de duas máquinas e um buffer. Eles abordam o desafio de reduzir o custo total, que é composto principalmente pelo custo de produção, as penalidades por

perda permanente de produção e o custo do nível de estoque em processo (WIP). Para resolver esse problema, os autores propõem duas políticas de decisão de controle com base em aprendizado por reforço. As políticas, chamadas LSP I e TH, foram encontradas por meio de amostras coletadas de um modelo simulado. Embora a política TH seja uma forma simplificada da LSP I, ambos os métodos subótimos se mostraram eficientes na redução do custo total de produção. O estudo destaca que essas políticas são facilmente implementáveis em sistemas de manufatura, demandando poucos cálculos em tempo real, tornando-as viáveis e eficazes para a aplicação em ambientes de produção com múltiplas máquinas e buffers.

A literatura engloba uma ampla variedade de estudos direcionados à utilização da Aprendizagem por Reforço profundo (Deep Reinforcement Learning - DRL) em contextos de produção industrial. No entanto, até o momento, não foi encontrado nenhum estudo que aborde a aplicação de um algoritmo de Aprendizagem por Reforço, como o Deep Q-Learning (DQL), em um cenário onde uma máquina é capaz de analisar o ambiente e tomar decisões em tempo real, considerando todos os eventos que ocorrem na fábrica.

Este trabalho apresenta um notável potencial de inovação ao empregar o Aprendizado por Reforço Profundo na otimização da produção industrial. Ao integrar essa tecnologia avançada com simulação industrial, cria-se um sistema autônomo capaz de aprender e se adaptar às mudanças nas condições de produção, impulsionando a automação inteligente e a eficiência da indústria. Além disso, a pesquisa contribui para a expansão do conhecimento em DRL aplicado à manufatura, estabelece critérios de avaliação para soluções de otimização e promove abordagens interdisciplinares para resolver desafios complexos na indústria, em consonância com os princípios da Indústria 4.0.

4. AMBIENTE DE SIMULAÇÃO

Neste capítulo, será apresentada uma descrição detalhada do ambiente de produção de latas de alumínio, composto por diversas máquinas e esteiras que formam o sistema de produção. Compreender esse ambiente é crucial para analisar o fluxo de trabalho e as interações entre as diferentes etapas do processo de fabricação. O ambiente de produção, ilustrado na Figura 1, é composto por 25 máquinas que desempenham funções específicas, desde a estampagem inicial até a paletização das latas. Além das máquinas, as esteiras desempenham um papel fundamental no transporte das latas entre as etapas da produção, atuando também como buffer em caso de falhas das máquinas e modulando o fluxo de produção. Esse ambiente complexo e interativo é essencial para garantir eficiência, qualidade e integridade do produto final na linha de produção de latas de alumínio.

4.1. DESCRIÇÃO DO AMBIENTE DE SIMULAÇÃO

Nesta seção, será apresentada uma descrição detalhada do ambiente de produção, composto por diferentes máquinas e esteiras, que formam o sistema de produção de latas de alumínio. Este ambiente é fundamental para compreender o fluxo de trabalho e as interações entre as diversas etapas do processo de fabricação.

O ambiente de produção, como visto na figura 1, é composto por 25 máquinas que desempenham funções específicas no processo de fabricação de latas de alumínio. Dentre as 25 máquinas, temos:

- **1 Cupper (CUP):** Responsável por estampar ou conformar folhas de alumínio para criar o formato inicial das latas.
- **8 Body Makers (BM):** Projetadas para formar o corpo da lata a partir das folhas processadas pela máquina Cupper.
- **1 Washer (WSH):** Remove resíduos, sujeira e contaminantes das latas antes do enchimento com produtos.
- **2 Printers (PRT):** Imprime designs, logotipos e informações gráficas nas latas.
- **8 Inside Sprays (IS):** Aplica um revestimento interno de verniz nas latas para proteger o conteúdo embalado.
- **1 Internal Bake Oven (IBO):** Acelera o processo de secagem e cura do revestimento aplicado nas latas.
- **1 Necker (NCK):** Molda a parte superior da lata, criando uma borda reforçada conhecida como "pescoço".
- **2 Light Testers (LT):** Identifica vazamentos de luz na lata, indicando falhas na selagem.
- **1 Palletizer (PLTZ):** Organiza e empilha as latas em paletes para embalagem e

transporte.

Adicionalmente às máquinas, as esteiras transportadoras desempenham um papel crucial no transporte de latas entre as diferentes etapas do processo de produção. Essas esteiras também atuam como um buffer de latas em caso de falha das máquinas, garantindo a continuidade do fluxo de produção. Além da função de transporte e de buffer, as esteiras desempenham outro papel muito importante na modulação da linha de produção; no entanto, este aspecto será abordado com mais detalhes posteriormente.

O fluxo de produção segue uma sequência lógica, começando com a estampagem das folhas de alumínio pela máquina CUP e terminando com a paletização na máquina PLTZ. Cada máquina realiza operações específicas, conforme descrito a seguir:

Cupper (CUP)

A máquina Cupper, como ilustrado na Figura 5, é uma peça fundamental no processo de fabricação de latas de alumínio. Sua principal função reside em estampar ou conformar as folhas de alumínio para criar o formato inicial das latas. Esse processo ocorre em várias etapas sequenciais.

Figura 5 - Máquina Cupper

Fonte: https://www.stollemachinery.com/pt-br/linhas-para-latas-de-bebidas#stolle-product-group___88, acessado em 10/02/2024

Inicialmente, a máquina é carregada com folhas de alumínio, que podem se apresentar em forma de bobinas ou folhas cortadas, dependendo da configuração adotada pela fábrica. Em seguida, a CUP aplica pressão controlada sobre as folhas, moldando-as no formato desejado para

as latas. Esta etapa é crucial, pois molda o alumínio na forma inicial da lata, definindo seu tamanho e formato.

Além disso, em certos casos, a máquina Cupper pode realizar operações adicionais, como cortar as bordas ou executar outras tarefas de acabamento, de acordo com as especificações do processo de produção. Essas operações adicionais visam garantir que as latas tenham uma qualidade consistente e atendam aos padrões exigidos. Por fim, após o processo de estampagem, as folhas de alumínio assumem a forma da base das latas e são retiradas da máquina para serem encaminhadas às próximas etapas do processo de fabricação. Essa saída das folhas conformadas representa a conclusão da etapa na Cupper, preparando o caminho para os procedimentos subsequentes na linha de produção de latas de alumínio.

Body Makers (BM)

Essas máquinas são projetadas para moldar o corpo da lata de acordo com as especificações desejadas, realizando uma série de etapas sequenciais, um exemplo desta pode ser visualizado na Figura 6.

Figura 6 - Máquina Body Maker

Fonte: https://www.stollemachinery.com/pt-br/linhas-para-latas-de-bebidas#stolle-product-group___88, acessado em 10/02/2024

Inicialmente, as folhas de alumínio processadas pela CUP são alimentadas na BM para dar início ao processo de formação do corpo da lata. Em seguida, a máquina executa operações de estampagem e conformação, moldando o corpo da lata de acordo com as dimensões e formato desejados. Essa etapa é crucial para garantir que as latas tenham uma forma consistente e atendam aos padrões de qualidade estabelecidos.

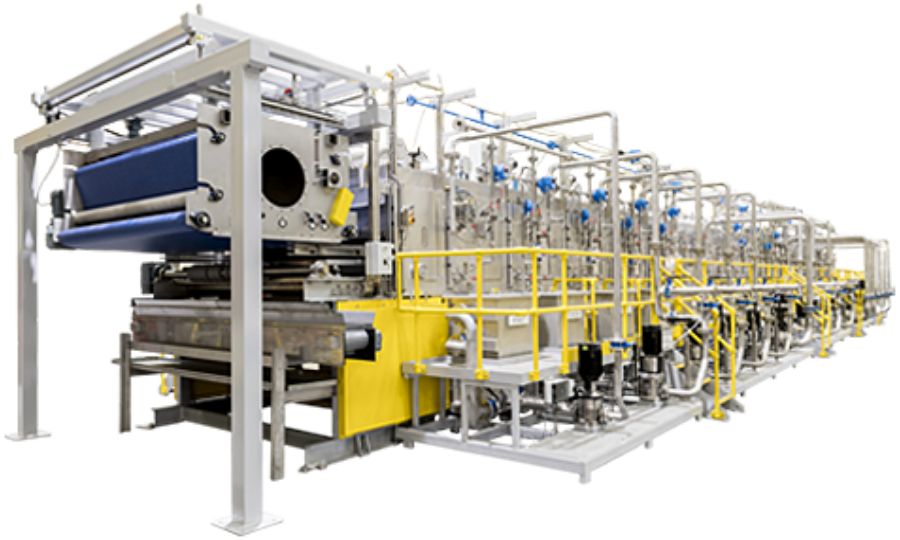
Após a formação do corpo da lata, as latas são retiradas da máquina e encaminhadas para as etapas subsequentes do processo de fabricação. Essa finalização marca o término do processo

realizado pelas Body Makers, preparando as latas para serem processadas nas próximas etapas da linha de produção de latas de alumínio.

Washer (WSH)

A máquina Washer, indicada na Figura 7, é responsável por garantir que as latas estejam completamente limpas e livres de quaisquer resíduos ou contaminantes antes do enchimento com produtos. As etapas de operação da Washer são cuidadosamente projetadas para assegurar uma limpeza eficaz das latas, seguindo os seguintes procedimentos:

Figura 7 - Máquina Washer



Fonte: https://www.stollemachinery.com/pt-br/linhas-para-latas-de-bebidas#stolle-product-group___88, acessado em 10/02/2024

Inicialmente, as latas vazias entram na máquina Washer através de uma esteira transportadora, dando início ao processo de limpeza. Antes da lavagem principal, algumas máquinas podem realizar uma etapa de pré-lavagem, onde jatos de água são aplicados para remover resíduos grosseiros e partículas soltas das superfícies das latas.

Em seguida, a Washer realiza a lavagem principal utilizando uma combinação de água e detergentes específicos, projetados para remover sujeira, poeira e outros contaminantes das latas de maneira eficiente. Durante esse processo, as latas podem ser submetidas a escovas ou outros métodos mecânicos para garantir uma limpeza mais profunda, especialmente em áreas de difícil acesso.

Após a lavagem, as latas passam por um processo de enxágue para remover qualquer resíduo de detergente ou partículas soltas que possam ter ficado nas superfícies. Algumas máquinas podem incluir também um processo de secagem, utilizando ar quente ou outros meios, para garantir que as latas estejam completamente secas antes de avançarem para as próximas

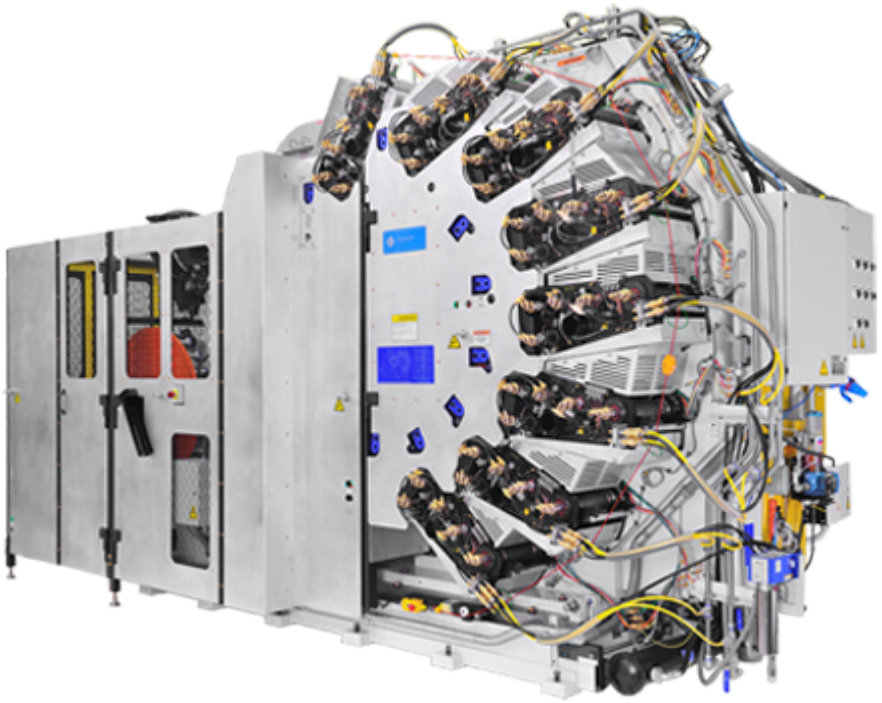
etapas do processo de produção.

Essas etapas cuidadosamente coordenadas da máquina Washer garantem que as latas estejam limpas e livres de contaminantes, cumprindo os mais altos padrões de qualidade e segurança para os produtos que serão posteriormente envasados nelas.

Printer (PRT)

A PRT é responsável por imprimir designs, logotipos, rótulos, códigos de barras e outras informações gráficas nas superfícies das latas, um exemplo pode ser observado na Figura 8. Suas etapas de operação são cuidadosamente executadas para garantir uma impressão precisa e de alta qualidade, seguindo os seguintes procedimentos:

Figura 8 - Máquina Printer



Fonte: https://www.stollemachinery.com/pt-br/linhas-para-latas-de-bebidas#stolle-product-group___88, acessado em 10/02/2024

Inicialmente, a Printer é configurada de acordo com o design específico que será aplicado nas latas. Isso envolve o carregamento de tintas coloridas ou outros materiais de impressão, bem como a programação da máquina para garantir a precisão na aplicação do design.

As latas entram na Printer por meio de uma esteira transportadora, posicionadas adequadamente para receber a impressão. Esse processo de alimentação das latas na máquina é fundamental para garantir uma impressão uniforme em todas as superfícies.

Em seguida, a máquina utiliza tecnologia de impressão avançada para aplicar a tinta diretamente nas latas. Isso pode envolver diferentes processos, como serigrafia, impressão por

transferência ou tecnologias de impressão digital, dependendo das especificações do processo e da máquina utilizada.

Após a aplicação da tinta, as latas passam por um processo de secagem para garantir que a tinta esteja completamente fixada nas superfícies das latas. Esse processo de secagem é essencial para garantir a durabilidade e a integridade do design impresso, preparando as latas para as etapas subsequentes do processo de produção, um exemplo de máquina responsável por este processo pode ser observado na Figura 10.

Figura 10 - Forno de Cura após a Printer

Fonte: https://www.stollemachinery.com/pt-br/linhas-para-latas-de-bebidas#stolle-product-group___88, acessado em 10/02/2024

Essas etapas cuidadosamente coordenadas da máquina Printer asseguram que as latas saiam da linha de produção com designs nítidos e de alta qualidade, prontas para serem preenchidas e distribuídas para os consumidores finais.

Inside Spray (IS)

A IS, exemplificada na Figura 11, aplica um revestimento interno de verniz que atua como uma barreira protetora entre o metal da lata e o conteúdo embalado, garantindo a qualidade e segurança dos alimentos ou bebidas armazenados. As atividades desta máquina são realizadas em várias etapas cuidadosamente coordenadas:

Figura 11 - Máquina Inside Spray

Fonte: https://www.stollemachinery.com/pt-br/linhas-para-latas-de-bebidas#stolle-product-group___88, acessado em 10/02/2024

Inicialmente, a máquina é preparada para o processo, garantindo que todos os sistemas estejam operacionais. Isso inclui a verificação e o abastecimento dos reservatórios de verniz, além da calibração dos bicos de spray e outros ajustes necessários para garantir a aplicação adequada do revestimento.

As latas vazias são alimentadas na máquina e posicionadas de maneira que o interior esteja acessível para receber o revestimento de verniz. Esse processo de alimentação das latas na máquina é fundamental para garantir uma aplicação uniforme do verniz em todas as superfícies internas das latas.

A máquina utiliza um sistema de pulverização para aplicar o verniz no interior das latas. O verniz é pulverizado de maneira uniforme, garantindo uma cobertura completa e consistente em todas as áreas internas das latas.

Em alguns casos, as latas podem ser giradas durante o processo para garantir uma distribuição uniforme do verniz em todas as superfícies internas. Esse giro das latas contribui para garantir a qualidade e eficácia do revestimento aplicado.

Após a aplicação do verniz, as latas saem da máquina prontas para as próximas etapas do processo de produção, como a aplicação de designs externos, o enchimento do produto e o fechamento da lata. O revestimento interno de verniz proporciona uma proteção essencial para os produtos armazenados nas latas, garantindo sua qualidade e segurança durante todo o processo de produção e distribuição.

Internal Bake Oven (IBO)

O IBO, representado na Figura 12, é responsável por garantir que o revestimento atinja as propriedades desejadas, como durabilidade, aderência e resistência química. As atividades desta máquina são realizadas em várias etapas cuidadosamente coordenadas:

Figura 12 - Máquina Internal Bake Oven

Fonte: https://www.stollemachinery.com/pt-br/linhas-para-latas-de-bebidas#stolle-product-group___88, acessado em 10/02/2024

Inicialmente, as latas que receberam algum tipo de revestimento, como tinta ou verniz, são alimentadas na entrada do forno de cura. Esse processo de alimentação das latas na máquina é fundamental para garantir que todas as latas sejam submetidas ao processo de cura de forma adequada.

No interior do forno, as latas são expostas a temperaturas elevadas controladas de acordo com as especificações do processo. O aquecimento é essencial para ativar o processo de cura do revestimento aplicado nas latas, garantindo sua aderência e durabilidade.

Durante a permanência no forno, o revestimento aplicado nas latas passa por um processo de cura que envolve a secagem e a reação química. Esse processo é crucial para garantir que o revestimento atinja as propriedades desejadas e proporcione a proteção necessária aos produtos armazenados nas latas.

Após o processo de cura, as latas passam por uma zona de resfriamento dentro do forno. Isso é feito para garantir que as latas não saiam do forno a uma temperatura que possa comprometer a qualidade do produto ou causar danos, proporcionando um resfriamento gradual e controlado.

Uma vez que as latas tenham completado o processo de cura e resfriamento, elas estão prontas para as próximas etapas do processo de produção, como a decoração, o enchimento do produto e o fechamento da lata, garantindo a qualidade e integridade do produto final.

Necker (NCK)

A NCK cria a borda característica na parte superior da lata, conhecida como "pescoço". Esta borda é fundamental para garantir uma selagem hermética durante o processo de vedação posterior. As atividades realizadas pela máquina NCK, representada na Figura 13, podem ser divididas em três etapas principais:

Figura 13 - Máquina Necker

Fonte: https://www.stollemachinery.com/pt-br/linhas-para-latas-de-bebidas#stolle-product-group___88, acessado em 10/02/2024

Inicialmente, as latas vazias são alimentadas na máquina por meio de um sistema de transporte, garantindo que estejam posicionadas adequadamente para o processo de formação do pescoço. Esta etapa de alimentação é fundamental para garantir uma produção contínua e eficiente.

Em seguida, a máquina realiza operações de estampagem e conformação para moldar a parte superior da lata, criando o pescoço reforçado necessário para a vedação hermética. Durante esse processo, podem ser utilizadas matrizes e ferramentas específicas para garantir que o formato do pescoço atenda às especificações desejadas.

Após a conclusão do processo de formação do pescoço, as latas estão prontas para seguir para as etapas subsequentes do processo de produção. Essas etapas podem incluir a decoração, o enchimento do produto e o fechamento da lata, completando assim o ciclo de produção.

É importante destacar que as máquinas Necker, juntamente com as demais mencionadas anteriormente, assim como as esteiras transportadoras, desempenham um papel crucial na composição de uma linha de produção completa para a fabricação de latas. Esses componentes trabalham em conjunto para garantir eficiência, qualidade e integridade do produto final,

atendendo às exigências do mercado e garantindo a satisfação dos clientes.

Light Testers (LT)

A máquina de inspeção de luz (LT), exemplificada na Figura 14, é crucial para assegurar a qualidade das latas produzidas, identificando potenciais vazamentos de luz que indicam falhas na selagem. Esse processo é essencial para preservar a qualidade e a integridade dos produtos armazenados nas latas, especialmente em bebidas carbonatadas ou outros itens suscetíveis à contaminação externa. O procedimento conduzido pela máquina LT pode ser dividido em cinco etapas distintas.

Figura 14 - Máquina Light Tester

Fonte: https://www.stollemachinery.com/pt-br/linhas-para-latas-de-bebidas#stolle-product-group___88, acessado em 10/02/2024

Inicialmente, as latas, após passarem por várias fases de produção, são encaminhadas para a máquina de inspeção de luz, onde serão submetidas ao teste de vedação. Em seguida, a máquina cria um ambiente controlado ao redor da lata, usualmente aplicando pressurização ou vácuo, condições essenciais para detectar vazamentos de gás ou líquido.

Durante a inspeção, a máquina utiliza um sistema sensível de detecção de luz para verificar vazamentos ao redor da lata. Isso é feito iluminando a lata e observando qualquer transmissão de luz externa que possa indicar uma falha na selagem.

Quaisquer falhas detectadas são registradas pela máquina para que as latas com problemas possam ser identificadas e retiradas da linha de produção, garantindo a rastreabilidade do processo e possibilitando correções adequadas.

Por fim, dependendo da configuração da máquina, as latas com falhas podem ser automaticamente removidas da linha de produção ou sinalizadas para que os operadores realizem inspeções mais detalhadas, assegurando que apenas as latas que atendam aos padrões de qualidade estabelecidos prossigam no processo de produção.

Palletizer (PLTZ)

A máquina paletizadora (PLTZ), como mostrado na Figura 15, é uma peça essencial no fechamento do ciclo de produção de latas, concentrando-se especialmente na etapa de empacotamento e preparação para o transporte. Seu propósito primordial é dispor e empilhar as latas de maneira otimizada sobre paletes, simplificando sua movimentação dentro da fábrica, a estocagem em depósito e o transporte para distribuição. O funcionamento desse equipamento pode ser delineado em uma sequência de seis passos distintos.

Figura 15 - Máquina Palletizer

Fonte: https://www.stollemachinery.com/pt-br/linhas-para-latas-de-bebidas#stolle-product-group___88, acessado em 10/02/2024

Inicialmente, as latas, após terem percorrido todas as fases de produção, são encaminhadas para a paletizadora através de uma esteira transportadora. Em seguida, a máquina é programada para organizar as latas em camadas sobre o palete, podendo ajustar a disposição para maximizar a capacidade de armazenamento ou atender a requisitos específicos de empilhamento.

Cada camada de latas é então elevada à altura apropriada e posicionada no palete em construção. Esse processo é realizado com o auxílio de sistemas de elevação e dispositivos de manipulação que garantem precisão e segurança durante a movimentação das latas.

Algumas paletizadoras podem intercalar camadas de latas com folhas de papelão ou outros materiais para fornecer estabilidade adicional e proteção contra danos durante o transporte. Após a formação de cada camada, a paletizadora pode aplicar pressão para compactar as latas e assegurar um empilhamento firme e estável.

Ao finalizar o arranjo das latas sobre o palete, o equipamento pode movê-lo para uma área

designada onde é envolto com filme plástico para estabilização e segurança. Assim, o palete fica pronto para ser transportado com eficiência e integridade até seu destino final, concluindo com sucesso o processo de paletização.

Esteiras

Como mencionado anteriormente, as esteiras desempenham múltiplos papéis na linha de produção de latas, indo além do simples transporte e buffer. Além disso, elas desempenham um papel essencial na modulação do fluxo de produção. Essa modulação é possível graças aos sensores estrategicamente posicionados ao longo das esteiras. Esses sensores têm a importante função de identificar acúmulos de latas em pontos específicos das esteiras, como ilustrado na Figura 16.

Figura 16 - Esteira

Fonte: https://www.stollemachinery.com/pt-br/linhas-para-latas-de-bebidas#stolle-product-group___88, acessado em 10/02/2024

Dependendo da detecção desses acúmulos, os sensores podem acionar medidas corretivas, como a diminuição da velocidade das máquinas ou até mesmo a interrupção temporária do fluxo de produção. No ambiente de simulação, foram implementadas 45 esteiras, algumas das quais equipadas com vários sensores, totalizando 69 sensores em toda a linha de produção. A seguir, está a Tabela 1 detalhando de maneira aleatória, por questões de confidencialidade, as esteiras e a quantidade de sensores presentes em cada uma:

Tabela 1 – Dados das Esteiras e Sensores

Esteira	Quantidade de Sensores	Esteira	Quantidade de Sensores
E1	0	E24	0
E2	5	E25	0
E3	0	E26	2
E4	0	E27	0
E5	0	E28	0
E6	1	E29	1
E7	4	E30	1
E8	0	E31	0
E9	1	E32	3
E10	0	E33	4
E11	4	E34	0
E12	0	E35	1
E13	0	E36	0
E14	0	E37	3
E15	4	E38	0
E16	1	E39	4
E17	6	E40	2
E18	0	E41	3
E19	0	E42	0
E20	1	E43	0
E21	0	E44	3
E22	4	E45	7
E23	4		

No trecho entre a máquina CUP e a Body Maker (BM), identificam-se quatro esteiras, das quais uma está munida de cinco sensores. Avançando para a etapa subsequente, entre a BM e a Washer (WSH), observam-se três esteiras e cinco sensores. Em seguida, na seção que vai da WSH até a Printer (PRT), encontram-se sete esteiras e cinco sensores. Ao longo do trajeto entre a Printer (PRT) e a Inside Spray (IS), são dispostas nove esteiras, equipadas com um total de onze sensores para monitoramento. Após a Inside Spray (IS), encontram-se apenas três esteiras, e um único sensor. Da Internal Bake Oven (IBO) até a Necker (NCK), conta-se com cinco esteiras e oito sensores. Finalmente, do percurso da Necker (NCK) até a paletizadora (PLTZ), passando pela

máquina de inspeção de luz (LT), são oito esteiras que hospedam um total de dezenove sensores. Além das esteiras mencionadas, a linha de produção integra duas esteiras de buffer adicionais, uma delas equipada com onze sensores e a outra com quatro sensores. Em um tópico subsequente, serão aprofundadas a função e a lógica de utilização desses sensores, assim como a relevância das esteiras adicionais na otimização do processo de produção.

4.2 MODULAÇÃO DA LINHA DE PRODUÇÃO

O presente tópico visa fornecer uma visão sucinta do funcionamento da modulação exclusiva da Body Maker (BM). Isso se dá devido à criação de um modelo de RL destinado ao controle dessas máquinas, com o intuito de avaliar o desempenho da linha utilizando uma inteligência artificial para controlar a velocidade de apenas um tipo de máquina. Por questões de segurança e confidencialidade, não é possível abordar detalhadamente a modulação. Entretanto, para manter o contexto, será delineada uma visão geral do processo, facilitando a compreensão dos tópicos subsequentes.

Passando agora para a importância das esteiras, estas desempenham um papel crucial além do simples transporte e buffer, integrando-se também à modulação das máquinas por meio de seus sensores. No caso da BM, a esteira E2 conta com um sensor que emite um sinal quando não há latas em determinada área da esteira. Consequentemente, diante dessa falta de latas, as BMs são interrompidas até que mais latas cheguem àquele trecho da esteira, caracterizando o que chama-se de auto-restart da máquina pela entrada. Já a esteira E6 possui um sensor que identifica um acúmulo de latas em seu trajeto, acarretando na paralisação das BMs e novamente ocorrendo o auto-restart, desta vez pela saída.

É importante ressaltar que esses auto-restarts representam um custo adicional para a empresa, pois se tratam de interrupções devido à falta de modulação entre as máquinas. Dado que a modulação atual opera com base em regras empíricas, fundamentadas em condições específicas, torna-se praticamente impossível contemplar todos os cenários possíveis, devido à complexidade da linha de produção.

Além dos auto-restarts, outro aspecto relevante é a configuração de velocidade das BMs, que atualmente operam apenas em velocidade mínima ou máxima. Simplificadamente, se uma determinada máquina à frente na linha estiver operando normalmente e não houver acúmulo de latas em um determinado trecho, as BMs operarão na velocidade máxima; caso contrário, operarão na velocidade mínima. Contudo, essa abordagem não contempla os diversos cenários do dia a dia de produção, resultando em subutilização dos recursos e eventual ocorrência de auto-restarts.

Em resumo, a modulação atual da linha baseia-se em um conjunto limitado de regras de

condições (*if-else*), o que acarreta em múltiplas paradas das máquinas em cenários atípicos. Isso prejudica a produtividade, o desempenho das máquinas, aumenta os custos de manutenção e resulta em desperdício de recursos, uma vez que, no caso das BMs, é necessário um tempo para que a máquina atinja a temperatura ideal após o religamento, antes de retomar a produção.

5. DESENVOLVIMENTO DO MODELO

Diante deste contexto, surge a provocação: considerando a quase impossibilidade de prever todos os cenários do cotidiano e modelar a correção destes, por que não desenvolver um modelo de aprendizado de máquina capaz de compreender esses cenários, enfrentar situações atípicas e, com isso, formular sua própria lógica de modulação? Essa reflexão nos conduz a uma segunda indagação: qual modelo seria mais adequado para enfrentar esse tipo de desafio? Apesar da ausência de uma base de dados estruturada e de rótulos para alimentar o modelo, dispõe-se de um ambiente de simulação já utilizado pela empresa ao longo dos anos. Nesse ambiente, cada alteração na velocidade das máquinas é baseada exclusivamente na situação presente, sem depender de dados históricos.

Dessa forma, podemos concluir que o Reinforcement Learning (RL), mais especificamente o Deep Reinforcement Learning (DRL), emerge como uma solução ideal para esse cenário. A complexidade do problema demanda uma abordagem que possa aprender com as interações com o ambiente e, a partir delas, desenvolver uma estratégia melhor de modulação. No contexto do DRL, o agente de aprendizado interage com o ambiente, recebendo recompensas ou penalidades com base em suas ações. Essa retroalimentação contínua permite que o sistema aprenda e ajuste seu comportamento de forma dinâmica, sem a necessidade de um conjunto prévio de dados rotulados. Portanto, o DRL apresenta-se como uma abordagem promissora para a criação de um sistema de modulação mais adaptável e eficiente, capaz de lidar com os desafios complexos e imprevisíveis enfrentados no ambiente de produção industrial.

5.1. DEFINIÇÃO DE ESTADOS, AÇÕES E RECOMPENSAS

Neste ponto, adentra-se à definição do modelo de Reinforcement Learning (RL), iniciando pela especificação dos estados. O estado representa a situação atual em que o ambiente e o agente se encontram, o que implica na necessidade de determinar quais parâmetros serão considerados no treinamento do modelo, visto que cada parâmetro reflete um aspecto do ambiente. Para este trabalho, optou-se por utilizar dados obtidos diretamente do Plant Simulation para o Python, por meio da biblioteca Plantsim. Essa biblioteca proporciona ferramentas que permitem a leitura e a modificação dos dados do Plant Simulation, além da execução de funções, como iniciar e pausar a simulação, por meio de comandos no Python.

Embora haja outras abordagens possíveis para este problema, como o uso de imagens do Plant Simulation durante a execução da simulação, entre outras, a decisão foi focar apenas nos dados coletados. Os parâmetros utilizados podem variar ao longo dos testes, mas serão detalhadamente especificados em cada experimento, incluindo a quantidade e quais parâmetros

foram empregados.

A seleção dos parâmetros é um aspecto crucial nos testes, podendo impactar significativamente o desempenho do modelo. É importante ressaltar que, para quem está desenvolvendo o código, pode parecer óbvio incluir todos os dados disponíveis do Plant Simulation, pois está familiarizado com o ambiente. No entanto, é essencial adotar uma perspectiva externa e imaginar-se em um ambiente completamente desconhecido, considerando o tipo de informação que cada parâmetro pode fornecer, como se estivéssemos em um ambiente completamente escuro.

Embora haja a crença de que mais informações resultam em um melhor aprendizado do modelo, na prática, isso nem sempre é verdadeiro. De fato, a inclusão de um grande número de parâmetros, como a velocidade das máquinas, se estas estão em falha ou não, e dados dos sensores, pode tornar o processo de associação entre o estado atual e as recompensas recebidas mais complexo para o modelo. Em algumas situações, a inclusão de muitos parâmetros pode prejudicar mais do que ajudar, tornando o sistema menos eficiente na aprendizagem e na tomada de decisões.

Quanto às ações disponíveis para o modelo, estas se limitam a três possibilidades, representadas por números inteiros: 0, 1 ou 2. No contexto do modelo proposto, o valor 0 indica que nenhuma alteração será feita na velocidade das Body Makers (BMs). Por sua vez, o valor 1 indica que o modelo aumentará a velocidade das BMs em uma unidade. Já o valor 2 implica na diminuição da velocidade das BMs em uma unidade. Contudo, é importante ressaltar que essas ações devem respeitar os limites estabelecidos para as velocidades das BMs. Assim, caso a velocidade atinja o limite máximo, tentativas de aumento não terão efeito, e o mesmo ocorrerá caso a velocidade atinja o limite mínimo, impedindo a redução adicional. Detalhes sobre como o modelo chega a esses valores inteiros de 0, 1 ou 2 serão abordados em outro tópico.

No que se refere à função de recompensa, esta poderá sofrer ajustes ao longo dos testes, mas inicialmente será projetada para punir o modelo de forma progressiva com um valor reduzido sempre que ocorrerem auto-restarts em todas as máquinas, não se restringindo apenas às BMs. Em contrapartida, o modelo receberá uma recompensa substancial sempre que uma hora de simulação transcorrer sem que ocorra nenhum auto-restart em qualquer máquina. A cada hora, os contadores de auto-restart de cada máquina serão zerados e reiniciados. Ademais, será atribuída uma recompensa significativa sempre que a produção em uma hora for superior à produção na hora anterior. A concepção inicial desta função de recompensa visa incentivar o modelo a evitar auto-restarts e a buscar constantemente o aumento da produção. Contudo, é importante ressaltar que nem sempre é tão simples assim, pois o modelo pode encontrar brechas não previstas, desafiando as expectativas e apresentando resultados surpreendentes.

5.2. ESTRUTURA DO MODELO DE DRL

O modelo está organizado da seguinte maneira: primeiro, são importadas as bibliotecas necessárias. Em seguida, algumas funções são criadas para iniciar a simulação, pausá-la e resetá-la, além de puxar dados do ambiente de simulação. Então, passa-se para a configuração do ambiente de simulação de acordo com cada experimento. Inicialmente, estabelecemos o tempo de simulação em que o Plant Simulation irá operar. Os primeiros testes serão realizados ao longo de 8 dias e 2 horas de simulação, com uma escala de 100 segundos de simulação para cada segundo real.

O treinamento do modelo ocorrerá ao longo de um período de 8 dias apenas. As 2 horas iniciais são designadas como um período de aquecimento (*warm up*) para que as esteiras sejam preenchidas por latas. Isso é essencial, pois, sem considerar esse tempo, os parâmetros das máquinas localizadas no final da linha de produção permaneceriam inalterados por muitas iterações, o que poderia prejudicar o aprendizado do modelo.

Ao iniciar a simulação, o python começa a coletar informações sobre o tempo de simulação e inicia o treinamento apenas quando identifica que se passaram 2 horas de simulação. A partir desse momento, os dados são coletados exclusivamente para o treinamento do modelo, desconsiderando as 2 horas iniciais de aquecimento para avaliação do desempenho do modelo. Prosseguindo, são criadas variáveis, incluindo listas e dicionários, que receberão os dados ao longo do treinamento para gerar uma tabela ao final. Além disso, é criada a rede neural responsável por receber os parâmetros e determinar a ação a ser tomada.

Feitas as configurações iniciais, são utilizados comandos para configurar o Plant Simulation. Após isso, a simulação é iniciada, e após 2 horas de simulação, o treinamento começa. Este é realizado em um loop *while*, no qual o modelo treinará até que o tempo de simulação alcance o tempo estabelecido inicialmente. Os dados de velocidade das máquinas, sensores e falhas são coletados e armazenados em variáveis.

Para cada máquina, avalia-se se há algum auto-restart pela entrada ou pela saída no momento atual. Caso afirmativo, é criada uma variável contador, que é incrementada a cada iteração em que a máquina estiver em auto-restart. Essa variável é utilizada na função de recompensa, onde o valor da recompensa recebe um valor negativo multiplicado pela quantidade de auto-restart de cada máquina. Esse valor negativo precisa ser pequeno, caso contrário, o modelo receberá penalidades significativas a cada iteração, dificultando a busca por alternativas para aumentar a recompensa.

A cada hora de simulação, os valores de auto-restart são zerados para evitar penalidades excessivas, e é avaliado se a produção durante esse período foi maior do que na hora anterior. Em

caso afirmativo, o modelo recebe uma recompensa positiva significativa, e se alguma máquina passar esse período de 1 hora sem auto-restart, o modelo também recebe uma recompensa alta.

Coletados os valores de velocidade das máquinas, dados dos sensores, informações de falhas e auto-restarts, uma lista contendo todas essas informações é criada e fornecida à rede neural, juntamente com a recompensa acumulada. Com base nessas informações, a rede neural retorna um vetor com três índices, dos quais apenas um recebe o valor 1, enquanto os demais são atribuídos como zero. O índice que recebe o valor 1 indica qual ação deve ser tomada. Por exemplo, se a rede neural retornar um vetor no qual o número 1 está na posição 2, então a ação escolhida será aquela associada ao valor 2, que corresponde à redução das velocidades das BMs em 1 unidade.

Após determinar a ação recomendada pela rede neural, são utilizados os comandos adequados para ajustar as velocidades no Plant Simulation. Em seguida, a recompensa acumulada é zerada e é verificado o tempo atual da simulação para determinar se as iterações devem continuar ou se devem ser encerradas. Este processo continua até que o tempo de simulação estabelecido inicialmente seja alcançado.

5.3. ALGORITMO DE TREINAMENTO

O algoritmo selecionado para este trabalho é o Deep Q-Learning (DQL), cuja rede neural é composta por 5 camadas. A primeira camada recebe a lista de parâmetros como entrada e transmite essas informações para uma camada intermediária com 256 neurônios. Em seguida, os dados são processados por uma camada com 128 neurônios, seguida por outra com 64 neurônios e mais uma com 32 neurônios. Por fim, temos a camada de saída, que retorna o vetor com 3 posições, conforme mencionado anteriormente.

O otimizador utilizado é o Adam, uma escolha comum para redes neurais devido à sua eficácia em problemas de otimização. Na camada de saída, é empregado o softmax, uma função de ativação frequentemente utilizada para problemas de classificação multi classe, que normaliza os valores de saída para formar uma distribuição de probabilidade sobre as possíveis ações.

O valor do hiperparâmetro epsilon, que controla a taxa de exploração do modelo, será ajustado ao longo dos testes. Inicialmente, será considerado como 0.85, mas pode ser modificado durante a fase de experimentação para encontrar a configuração mais adequada ao problema em questão.

5.4. AJUSTES NO MODELO E EXPERIMENTAÇÃO

Antes de tudo, é importante compreender como serão conduzidos os testes e quais ferramentas serão empregadas para avaliar o desempenho do modelo. Dentro do Plant Simulation,

cada máquina possui um método que é acionado a cada 5 segundos e é responsável pelo acionamento dos auto-restarts e pela configuração da velocidade.

O primeiro passo será estabelecer o baseline. Como estamos considerando inicialmente um período de simulação de 8 dias, será utilizada uma ferramenta nativa do Plant Simulation chamada Experiment Manager (EM). Essa ferramenta possibilita a realização de várias simulações, especificando os dados que se deseja observar. Ao final, o EM apresenta a média e o desvio padrão desses dados para o tempo especificado.

Para este trabalho, foram realizadas 1000 simulações no EM. Os dados que serão utilizados para comparação com o modelo de DRL serão a produção, a quantidade de auto-restart da CUP, a primeira BM, e as duas PRTs. Não será necessário coletar os dados de todas as BMs, pois quando uma entra em auto-restart, as demais também entram, resultando em valores idênticos ou muito próximos. Além disso, caso ocorra um auto-restart e a primeira BM apresente falha, apenas a falha dela será identificada, enquanto as demais serão contabilizadas como auto-restart, resultando em um valor muito próximo.

As demais máquinas possuem valores reduzidos de auto-restart ou não representam um problema significativo para a linha de produção, como é o caso da ISM. Uma vez que o EM tenha sido executado sem qualquer modificação, os resultados obtidos serão considerados como o valor de base para comparação com o modelo proposto.

Para iniciar os testes, é necessário desativar o controle de velocidade das BMs dentro do método, para que o modelo possa assumir essa responsabilidade. Além disso, o ambiente de simulação é baseado em "cockpits", que consiste em extrair dados da nuvem de uma determinada semana. Esses dados incluem as velocidades mínimas e máximas de cada máquina, bem como sua disponibilidade. O "cockpit" é então baixado como uma planilha e importado no Plant Simulation para configurar as máquinas de acordo com essa semana específica.

Para os testes realizados, foram utilizados 6 cockpits diferentes, representando situações desde as semanas mais produtivas até as semanas com muitas situações atípicas. No total, cerca de 40 testes foram realizados, e a cada nova iteração, ajustes foram feitos. Isso incluiu a alteração dos valores das recompensas, testes com diferentes valores de epsilon e a variação dos parâmetros utilizados no modelo. Inicialmente, os testes foram conduzidos com 160 parâmetros, os quais incluíam as velocidades de todas as máquinas, dados de falhas de todas as máquinas, informações de todos os sensores e a quantidade de auto-restart de quase todas as máquinas. No entanto, ao longo das iterações, descobriu-se que o melhor desempenho foi alcançado com apenas 46 parâmetros, restringindo as informações apenas para as máquinas mais próximas.

Essa descoberta contraria a intuição comum de que quanto mais parâmetros, melhor será o

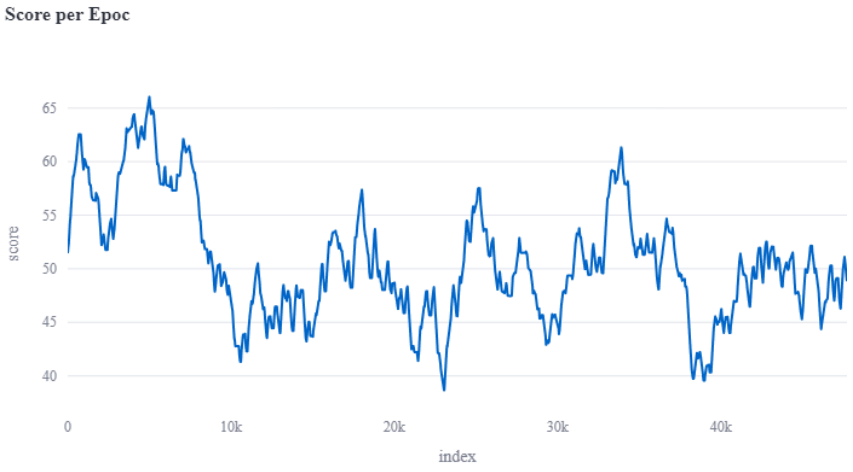
aprendizado do modelo. Este trabalho evidencia a necessidade de uma análise mais aprofundada da relação entre esses parâmetros e o que eles realmente representam para o modelo. Em alguns casos, a inclusão de mais parâmetros pode confundir mais do que ajudar, destacando a importância de compreender completamente o impacto de cada parâmetro no desempenho do modelo.

Durante os testes, observou-se que o modelo não conseguiu superar o baseline em termos de produção em nenhuma das simulações realizadas. As diferenças em relação ao baseline variaram entre -17,80% e -0,21%. No que diz respeito à quantidade de auto-restart, notou-se um aumento significativo nos modelos com produção bem abaixo do baseline, enquanto nos modelos com produção próxima ao baseline, não houve uma diferença considerável.

Os modelos que apresentaram os piores desempenhos foram aqueles configurados com 160, 131 e 129 parâmetros. Foi observado também que os testes variando apenas o valor de epsilon não resultaram em diferenças significativas nos resultados obtidos.

Um ponto de destaque durante os testes foi a falta de convergência do modelo, conforme evidenciado pelo gráfico do score, na Figura 17, que não apresentou um padrão de melhoria ao longo do tempo. Essa falta de convergência sugere que o modelo pode não estar aprendendo de forma eficaz ou que pode estar enfrentando dificuldades para generalizar os padrões encontrados durante o treinamento. A figura abaixo ilustra esse comportamento, em que o eixo y representa o *score* do modelo e no eixo x temos a quantidade de épocas:

Figura 17 - Gráfico de Score do DRL



Fonte: o autor (2024)

A cada ciclo de treinamento, os modelos eram salvos em arquivos utilizando a biblioteca Python chamada "Pickle". Devido à falta de convergência e à complexidade do modelo, além dos testes realizados criando novos modelos, foi adotada uma abordagem alternativa. Um modelo que

apresentou um desempenho próximo ao baseline foi selecionado e, em vez de criar um novo modelo, esse modelo salvo foi utilizado e submetido a treinamentos adicionais por mais épocas, na esperança de identificar alguma convergência ou melhoria no desempenho ao longo do tempo.

Durante os testes, também foi avaliada a mudança nas ações do modelo. Enquanto a ação 0 permaneceu como "não fazer nada", a ação 1 foi modificada para colocar as BMs para operar em velocidade máxima, no lugar de apenas aumentar 1 CPM (*Cans Per Minute*). Da mesma forma, a ação 2 foi ajustada para operar as BMs em velocidade mínima, ao invés de diminuir 1 CPM. No entanto, essa alteração não resultou em uma melhoria significativa no desempenho do modelo.

6. ANÁLISE DOS RESULTADOS E CONSIDERAÇÕES FINAIS

Ao longo deste estudo, explorou-se a aplicação inovadora do Aprendizado por Reforço Profundo (DRL) na otimização do controle de produção de embalagens de alumínio em uma empresa multinacional de grande porte. Este trabalho representou uma tentativa de integrar conceitos avançados de aprendizado de máquina com simulação de processos industriais, visando melhorar a eficiência e a competitividade em um ambiente de produção desafiador.

Uma análise inicial dos resultados nos leva a considerar a redução na quantidade de parâmetros como um aspecto significativo. Para alcançar essa redução, foi necessário adotar uma abordagem mais abstrata, considerando como qualquer ação tomada pelo modelo pode afetar diretamente cada parâmetro. Por exemplo, diminuir a velocidade pode ter um impacto rápido e significativo em máquinas como a NCK, onde certos parâmetros podem mudar a cada iteração, independentemente da velocidade das BMs. Concentrar os parâmetros apenas nas máquinas e sensores próximos às BMs ajudou a mitigar esse problema e a melhorar o treinamento do modelo.

Outro aspecto relevante é a lógica de auto-restart. A questão levantada é se simplesmente alterar a velocidade é suficiente para melhorar a modulação e a fluidez na linha de produção. Uma análise mais aprofundada do ambiente de simulação revelou que em situações onde as disponibilidades nas duas Printers estão baixas, ocorrerá auto-restart na BM pela saída devido ao acúmulo de latas. No entanto, se fosse possível parar apenas metade das BMs, ao invés de todas, qual seria o impacto disso? Essa é uma questão importante a ser considerada para otimizar ainda mais a modulação e a eficiência da linha de produção.

É importante notar que a lógica de modulação atualmente utilizada pela empresa funciona bem em condições normais, quando há pouca manutenção nas máquinas e tudo está funcionando conforme o esperado. No entanto, o problema surge em dias atípicos, quando a eficácia da lógica empírica é comprometida. Por esse motivo, esperava-se um desempenho melhor do modelo em semanas mais atípicas, onde a lógica empírica tende a falhar.

Diante dessas considerações, surge a questão sobre a validação do ambiente de simulação. Seria necessário realizar um estudo mais aprofundado para garantir que o ambiente de simulação reflita com precisão as condições reais da linha de produção e permita uma avaliação precisa do desempenho do modelo em diferentes cenários. Essa validação é crucial para garantir a eficácia e a aplicabilidade das soluções propostas.

No ambiente de simulação atual, observa-se uma discrepância em relação à mudança na velocidade das máquinas, que ocorre de forma instantânea, o que contrasta com a realidade operacional, especialmente no contexto da BM. Nesse cenário, é imprescindível considerar o

tempo necessário para que a máquina alcance a temperatura ideal após religada, o que demanda a implementação de uma lógica adequada para a rampa de velocidade. Além disso, a introdução da modelagem de falhas das máquinas representa uma oportunidade significativa para aprimorar a precisão do modelo de simulação. Ao analisar os dados de falhas e identificar a distribuição de probabilidade mais adequada para cada máquina, é possível incorporar essas informações ao modelo, aproximando-o ainda mais da realidade operacional. Essas melhorias têm o potencial não apenas de tornar o modelo mais robusto, mas também de otimizar seu desempenho, contribuindo assim para uma simulação mais fiel ao ambiente de produção real.

Apesar dos desafios encontrados e das limitações identificadas, este estudo representa um passo importante na jornada rumo à integração bem-sucedida de tecnologias avançadas de inteligência artificial na indústria de produção. Os insights obtidos ao longo deste trabalho fornecem uma base sólida para futuras pesquisas e desenvolvimentos nesta área, contribuindo para o avanço do conhecimento e para a busca contínua por eficiência e inovação na indústria.

Considerando os resultados e as conclusões deste estudo, é possível destacar diversos benefícios práticos que a empresa pode obter com a aplicação do Aprendizado por Reforço Profundo (DRL) na melhoria do controle de produção de embalagens de alumínio. Primeiramente, a implementação bem-sucedida do DRL pode levar a uma melhoria significativa na eficiência operacional da linha de produção, reduzindo custos e aumentando a produção. Além disso, a capacidade do modelo de aprender com as interações com o ambiente pode resultar em estratégias de modulação mais adaptáveis e eficientes, permitindo à empresa lidar de forma mais eficaz com desafios complexos e imprevisíveis no ambiente industrial. A utilização do DRL também pode contribuir para aprimorar a tomada de decisões, fornecendo insights valiosos e orientando ações para maximizar a utilização de recursos e melhorar a qualidade dos produtos. Dessa forma, o trabalho realizado neste estudo não apenas abre portas para avanços tecnológicos, mas também oferece oportunidades tangíveis para a empresa melhorar sua competitividade e alcançar maior excelência operacional em sua produção.

Para pesquisas futuras, é recomendável explorar outros algoritmos de aprendizado por reforço profundo, como o Double Deep Q-Learning, PPO (Proximal Policy Optimization), entre outros. Além disso, testar diferentes configurações de redes neurais, incluindo arquiteturas alternativas, pode proporcionar insights valiosos. A investigação de novas estratégias para a função de recompensa também é fundamental, uma vez que o objetivo deste trabalho não é necessariamente encontrar o melhor modelo, mas sim compreender melhor o comportamento do sistema e suas possíveis melhorias. Essas abordagens podem enriquecer ainda mais o campo de estudo e contribuir para avanços significativos na aplicação de inteligência artificial em ambientes industriais.

REFERÊNCIAS

ACHIAM, J.; MORALES, M. OpenAI Spinning Up – Part 2: Kinds of RL Algorithms. 2018. Disponível em: https://spinningup.openai.com/en/latest/etc/author.html. Acessado em 29 de julho de 2023.

AGOSTINELLI, F. *et al*. From Reinforcement Learning to Deep Reinforcement Learning: An Overview. In: COMPUTER SCIENCE, Irvine, 2018. **Lecure Notes...**, Irvine, 298-328, 2018.

ALTENMÜLLER, T. *et al*. Reinforcement learning for an intelligent and autonomous production control of complex job-shops under time constraints. **Production Management**, Production Engineering, 14: 319-328, 2020.

ANTONIO CARLOS GIL. **Como elaborar projetos de pesquisa**. [s.l.] Atlas, 2022.

BELLMAN, R. A Markovian Decision Process. **Journal of Mathematics and Mechanics**. JSTOR, 6(4): 679–684, 1957.

BERTO, R. M. VILLARES. S.; NAKANO, D. N. A produção científica nos anais do encontro nacional de engenharia de produção: um levantamento de métodos e tipos de pesquisa. **Production**, v. 9, n. 2, p. 65–75, dez. 1999.

FELDKAMP, N.; BERGMANN, S.; Strassburger, S. In: Simulation-Based Deep Reinforcement Learning For Modular Production Systems. In: WINTER SIMULATION CONFERENCE (WSC), Ilmenau, 2020. **Proceedings...** Ilmenau: Technische Universität Ilmenau, 2020.

GOSAVI, Abhijit. Simulation-Based Optimization. Parametric Optimization - Techniques and Reinforcement Learning. New York: Springer, 2015.

HUANG, J.; CHANG, Q.; CHAKRABORTY, N. Machine Preventive Replacement Policy for Serial Production Lines Based on Reinforcement Learning. In: 15th INTERNATIONAL CONFERENCE ON AUTOMATION SCIENCE AND ENGINEERING (IEEE 2019).

KAELBLING, L., P.; LITTMAN, M., L.; MOORE, A., W. Reinforcement Learning: A Survey. **Journal of Artificial Intelligence Research,** 4(1): 237-285, 1996.

LAKATOS, E. M.; ANDRADE MARCONI, M. DE. **Fundamentos de metodologia científica (7a. ed.)**. São Paulo: Editora Atlas S.A., 2010.

LENG, J. *et al*. Deep reinforcement learning for a color-batching resequencing problem. **Journal of Manufacturing Systems**, ScienceDirect, 56: 175-187, 2020.

LI, J. *et al*. A flexible manufacturing assembly system with deep reinforcement learning. **Control Engineering Practice**, ScienceDirect, 118: 104-957, 2022.

LUO, S.; ZHANG, L.; FAN, Y. Dynamic multi-objective scheduling for flexible job shop by deep reinforcement learning. **Computers & Industrial Engineering**, Science Direct, 159: 107-489, 2021.

MARCHESANO, M. G. *et al*. Dynamic scheduling of a due date constrained flow shop with Deep Reinforcement Learning. **IFAC PapersOnLine**, ScienceDirect, 55(10): 2932-2937, 2022.

MNIH, V., K. *et al*. Human-Level Control Through Deep Reinforcement Learning. **Nature**, 518(7540): 529–533, 2015.

MONTDHER ALABADI; ZAFER ALBAYRAK. Q-Learning for Securing Cyber-Physical Systems : A survey. 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications, 1 jun. 2020.

MORABITO, R. et al. **Metodologia de pesquisa em engenharia de produção e gestão de operações**. [s.l.] Elsevier Brasil, 2018.

SHIUE, Y., R.; LEE, K., C.; SU, C., T. Real-time scheduling for a smart factory using a reinforcement learning approach. **Computers & Industrial Engineering**, ScienceDirect, 125: 604-614, 2018.

SHRESTHA, A.; MAHMOOD, A. Review of Deep Learning Algorithms and Architectures. **IEEE Access**, 7: 53040–53065, 2019.

SUTTON, Richard; BARTO, Andrew. **Reinforcement Learning – An Introdution**. 2ª Edição. Cambridge, Massachusets: The MIT Press, 2018.

TALPAERT, V. *et al*. Exploring Applications of Deep Reinforcement Learning for Real-world Autonomous Driving Systems. In: 14th INTERNATIONAL JOINT CONFERENCE ON COMPUTER VISION, IMAGING AND COMPUTER GRAPHICS THEORY AND APPLICATIONS (VISIGRAPP 2019), 5: 564–572, 2019.

TAN, F.; YAN, P.; GUAN, X. Deep Reinforcement Learning: From Q-Learning to Deep Q-Learning. **Neural Information Processing**, p. 475–483, 2017.

WASCHNECK, B. *et al*. Optimization of global production scheduling with deep reinforcement learning. In: 51st CONFERENCE ON MANUFACTURING SYSTEMS (CIRP 2018), . **Procedia...**, 72, 1264-1269, 2018.

WATKINS, C., J., C., H.; DAYAN, P. Q-learning. In: MACHINE LEARNING, Boston, 1992. **Technical Note...**, Boston, Kluwer Academic Publisher, 1992, 8(3-4), 280-292, 1992.

ZHENG, W.; LEI, T.; CHANG, Q. Comparison Study of Two Reinforcement Learning Based Real-Time Control Policies For Two-Machine-One-Buffer Production System. In: 13th CONFERENCE ON AUTOMATION SCIENCE AND ENGINEERING (IEEE 2017).