UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS FLORIANÓPOLIS
CURSO DE GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

Matheus Henrique Schaly

Aplicação de Aprendizado de Máquina na Classificação de Litofácies

Florianópolis
2021

Matheus Henrique Schaly

Aplicação de Aprendizado de Máquina na Classificação de Litofácies

Trabalho de Conclusão de Curso do Curso de Graduação em Ciências da Computação do Campus Florianópolis da Universidade Federal de Santa Catarina para a obtenção do título de bacharel em Ciências da Computação.
Orientador: Prof. Dr. Mauro Roisenberg

Florianópolis
2021

Matheus Henrique Schaly

Aplicação de Aprendizado de Máquina na Classificação de Litofácies

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de "bacharel em Ciências da Computação" e aprovado em sua forma final pelo Curso de Graduação em Ciências da Computação.

Florianópolis, [dia] de [mês] de [ano].

Prof. Dr. Alexandre Gonçalves Silva
Coordenador do Curso

Banca Examinadora:

Prof. Dr. Mauro Roisenberg
Orientador

Profa. Dra. Jerusa Marchi
Avaliadora
Universidade Federal de Santa Catarina

Prof. Dr. Elder Rizzon Santos
Avaliador
Universidade Federal de Santa Catarina

RESUMO

Aprendizado de máquina (ML do inglês *Machine Learning*) vem se tornando uma ferramenta cada vez mais importante em vários campos da ciência, neste trabalho aplicaremos AM no ramo das geociências. O objetivo do trabalho consiste em propor um modelo eficaz de AM, incluindo a parte de manipulação dos dados, para a classificação de litofácies em poços geológicos. Acreditamos que a sequência de padrões sedimentares possa ajudar no processo de classificação, e para isso poderia ser utilizado uma versão modificada de uma redes neural recorrente (RNN do inglês *Recurrent Neural Network*). A classificação acurada de litofácies é de grande importância para obter informações geológicas úteis para a exploração e produção de hidrocarbonetos. A classificação automática de litofácies torna o processo de estudo da litologia dos poços mais rápido e menos oneroso. A classificação de litofácies é realizada estudando as propriedades litológicas das rochas encontradas em poços, que são características dos sedimentos atuais acumulados em determinadas condições físicas e geográficas. As propriedades litológicas podem incluir raio gama, resistividade, efeito fotoelétrico, densidade de porosidade de nêutrons, porosidade de densidade de nêutrons média, entre outras. Dado um banco de dados contendo as características e a classificação das litofácies, é esperado que o modelo proposto consiga, de maneira eficaz, realizar automaticamente a classificação de tais litofácies. A eficácia do método será medida através das métricas de classificação, como acurácia, precisão, *recall* e *F1-score*.

Palavras-chave: Aprendizado de máquina, classificação automática de litofácies.

ABSTRACT

Machine Learning (ML) has become an increasingly important tool in various fields of science, in this work we will apply ML in the field of geosciences. The objective of the work is to propose an effective model of ML, including data manipulation, for the classification of lithofacies in geological wells. We believe that the sequence of sedimentary patterns can help in the classification process, and for that a modified version of a recurrent neural network (RNN) could be used. Accurate classification of lithofacies is of great importance to obtain useful geological information for the exploration and production of hydrocarbons. The automatic classification of lithofacies makes the process of studying the lithology of the wells faster and less costly. The classification of lithofacies is performed by studying the lithological properties of the rocks found without wells, which are characteristic of the current sediments accumulated under certain physical and geographical conditions. Lithological properties can include gamma ray, resistivity, photoelectric effect, neutron porosity density, average neutron density porosity, among others. Given a database containing the characteristics and the classification of lithofacies, it is expected that the proposed model will be able to effectively carry out the classification of such lithofacies automatically. The effectiveness of the method will be measured through the classification metrics, such as accuracy, precision, recall and F1-score.

Keywords: Machine learning, automatic classification of lithofacies.

LISTA DE FIGURAS

Figura 1 – Ambientes Deposicionais 14
Figura 2 – Fácies Sedimentares 14
Figura 3 – Hierarquia das subáreas da inteligência artificial. 19
Figura 4 – Feedforward Neural Network. 20
Figura 5 – Função de ativação ReLU. 20
Figura 6 – Deep Feedforward Neural Network. 21
Figura 7 – Esquema da célula sigma recorrente padrão. 22
Figura 8 – Arquitetura da LSTM com um portão de esquecimento.. 24
Figura 9 – A arquitetura geral da 1D-CNN. 30
Figura 10 – Comparação dos resultados da classificação de fácies usando a abordagem proposta com RNN, LSTM, SVM e KNN. 31
Figura 11 – Arquitetura *FaciesNet*. 31
Figura 12 – Acurácia e acurácia balanceada da rede. 31
Figura 13 – Comparação de precisão, *recall*, *F1-score*, da *FaciesNet* com *Naive Bayes*. 32
Figura 14 – Arquitetura da CNN. 32
Figura 15 – Resultados das três diferentes estratégias de preenchimento, onde a acurácia é dada pelo *F1-score*. 32
Figura 16 – Resultados da competição 2016 SEG ML (https://github.com/seg/2016-ml-contest). 33
Figura 17 – Resultados da avaliação do desempenho dos modelos para o poço A. . . 33
Figura 18 – Resultados da avaliação do desempenho dos modelos para o poço B. . . 33
Figura 19 – Matriz de penalidade. 35
Figura 20 – Métrica de avaliação. 35
Figura 21 – Etapas pré-processamento dos dados. 36

LISTA DE QUADROS

LISTA DE TABELAS

Tabela 1 – Planejamento das etapas do trabalho de conclusão de curso 38

SUMÁRIO

1 INTRODUÇÃO 9
1.1 OBJETIVOS 10
1.1.1 Objetivo Geral 10
1.1.2 Objetivos Específicos 11
1.2 MÉTODO DE PESQUISA 11
1.3 ESTRUTURA DO TRABALHO 11
2 FUNDAMENTAÇÃO TEÓRICA 13
2.1 FÁCIES SEDIMENTARES 13
2.1.1 Litofácies 15
2.2 ATRIBUTOS DE LITOFÁCIES 15
2.3 APLICAÇÃO DE ML NA CLASSIFICAÇÃO DE LITOFÁCIES 17
2.4 INTELIGÊNCIA ARTIFICIAL 18
2.4.1 Aprendizado de Máquina 18
2.4.1.1 Aprendizagem Profunda 19
2.4.1.1.1 Rede Neural Recorrente 22
3 REVISÃO SISTEMÁTICA DA LITERATURA 25
3.1 LITHOLOGICAL FACIES CLASSIFICATION USING DEEP CONVOLUTIONAL NEURAL NETWORK 25
3.2 FACIESNET: MACHINE LEARNING APPLICATIONS FOR FACIES CLASSIFICATION IN WELL LOGS 26
3.3 CHARACTERIZING ROCK FACIES USING MACHINE LEARNING ALGORITHM BASED ON A CONVOLUTIONAL NEURAL NETWORK AND DATA PADDING STRATEGY 26
3.4 COMPARISON OF DIFFERENT MACHINE LEARNING ALGORITHMS FOR LITHOFACIES CLASSIFICATION FROM WELL LOGS 27
3.5 COMPARAÇÃO ENTRE OS TRABALHOS 27
4 PROPOSTA 34
4.1 MÉTODO DE REDE NEURAL RECORRENTE 34
4.2 CONJUNTO DE DADOS 34
4.3 MÉTRICA DE AVALIAÇÃO 34
5 EXPERIMENTOS E RESULTADOS 36
5.1 PRÉ-PROCESSAMENTO 36
6 CONCLUSÕES 38
6.1 TRABALHOS FUTUROS 38
REFERÊNCIAS 39

1 INTRODUÇÃO

Há diversas definições existentes para o termo fácies. Definimos fácies como qualquer parte restrita não comparável de uma unidade estratigráfica projetada que exibe caráter significativamente diferentes daqueles de outras partes da unidade (MOORE, 1949). Biofácies são fácies identificadas por características paleontológicas (conteúdo fóssil) sem levar em conta o caráter litológico. Litofácies são fácies identificadas com base em características litológicas (BOGGS, 2001). Usaremos litofácies como base de dados no presente trabalho.

A classificação de litofácies consiste em atribuir uma classe de rocha a uma amostra específica com base nas características medidas. A fonte ideal para classificação de litofácies são amostras de núcleo de rochas extraídas de poços. No entanto, devido aos custos associados, nem sempre as amostras de núcleo podem ser obtidas. Além disso, o método convencional é um processo tedioso e demorado, pois consiste em classificar litofácies manualmente por intérpretes humanos. Portanto, um método para classificar fácies a partir de medidas indiretas (por exemplo, gerar perfis utilizando cabo de aço) é necessário. Várias abordagens distintas para a questão da classificação de fácies utilizando dados de poços já foram propostas (MANDAL; REZAEE, 2019). Neste trabalho será investigado um conjunto de 118 perfis de poços que possui 7 tipos de metadados, 13 atributos e 12 classes.

Nos últimos anos, ML se tornou uma ferramenta interdisciplinar cada vez mais importante, que avançou vários campos da ciência, como biologia, química, medicina e farmacologia. Especificamente, o método de rede neural profunda (DNN do *Deep Neural Network*) encontrou ampla aplicação. Enquanto a geociência foi mais lenta na adoção, a bibliometria mostra adoção do aprendizado profundo (DL do inglês *Deep Learning*) em todos os aspectos da geociência (DRAMSCH, 2020).

Aprendizado de máquina é profundamente enraizada em estatísticas aplicadas, criando modelos computacionais que utilizam ML de inferência e reconhecimento de padrões em vez de conjuntos explícitos de regras (DRAMSCH, 2020). Aprendizado de máquina é o campo de estudo que fornece aos computadores a capacidade de aprender sem serem explicitamente programados (SAMUEL, 1959). Aprendizagem supervisionada consiste na tarefa de um algoritmo de ML em aprender uma função que mapeia uma entrada para uma saída com base em exemplos de pares de entrada e saída (RUSSELL; NORVIG, 2010). Uma função é inferida a partir de dados de treinamento rotulados que consistem em um conjunto de exemplos de treinamento (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012).

O DL é uma forma de ML que permite que os computadores aprendam com a experiência e entendam o mundo em termos de uma hierarquia de conceitos. A hierarquia de conceitos permite que o computador aprenda conceitos complicados construindo-os a

partir de outros mais simples (GOODFELLOW; BENGIO; COURVILLE, 2016).

Recentemente, as técnicas de DL foram desenvolvidas e amplamente adotadas para extrair informações de vários tipos de dados. Considerando as diferentes características dos dados de entrada, existem vários tipos de arquiteturas para DL, como a RNN, rede neural convolucional (CNN do inglês Convolucional Neural Network), e DNN. Geralmente a RNN e a DNN não podem lidar com as informações temporais de entrada de dados. Portanto, em áreas de pesquisa que contêm dados sequencias, como texto, áudio e vídeo, RNNs são dominantes. Contudo, RNNs são incapazes de aprender as informações relevantes dos dados de entrada quando a lacuna de entrada é grande. Ao introduzir funções de portão na estrutura da célula de uma RNN a long short-term memory (LSTM) poderia lidar bem com o problema das dependências de longo prazo. Desda a introdução da RNN quase todos os resultados interessantes baseados em RNNs foram alcançados pela LSTM. A LSTM se tornou o foco do DL (YU, Y. *et al.*, 2019).

O problema de classificação automática de litofácies deve ser explorado a fim de diminuir os custos envolvidos na classificação manual de litofácies. Existem competições envolvendo a classificação automática acurada de perfis de poços por meio de algoritmo de AM. No presente trabalho avaliaremos a competição, já encerrada, chamada *Force 2020 Machine Learning Competition* (2020, s.d.), na qual modelo de ML vencedor da competição utilizou o algoritmo XGBoost. Nossa solução ao problema de classificação de litofácies utilizará um algoritmo de ML supervisionado. Acreditamos que a sequência de padrões sedimentares possa ajudar no processo de classificação. Portanto, podemos, mais especificamente, criar uma nova topologia de RNN ou LSTM que venha a considerar este aspecto sequencial do nosso banco de dados. Além disso, criaremos um pipeline para a manipulação dos dados para organizar e melhorar os dados de entrada ao modelo.

1.1 OBJETIVOS

A proposta deste trabalho é estudar análises prévias de modelos de ML aplicadas para a classificação de litofácies. Com uma introdução dos fundamentos de litofácies e IA, assim como o estudo de trabalhos correlatos.

Após a introdução, realizamos o treinamento do modelo de AM, apresentamos o banco de dados utilizado e concluímos o trabalho com os resultados obtidos pelo modelo criado em comparação com o resultado de outros modelos da literatura.

Os objetivos são divididos em:

1.1.1 Objetivo Geral

Desenvolver uma solução para a classificação automática de litofácies, partindo da análise, limpeza e organização dos dados e chegando ao desenvolvimento de um modelo de ML supervisionado.

1.1.2 Objetivos Específicos

a) Estudo sobre litofácies e classificação automática de litofácies utilizando AM. Assim como o estudo de algoritmos de ML que levam em consideração a sequencia dos dados.

b) Levantamento da literatura buscando técnicas que já foram utilizadas para esta tarefa e tarefas similares.

c) Construção de um *pipeline* para a manipulação dos dados que serão utilizados como entrada para o algoritmo de ML proposto.

d) Desenvolvimento de um algoritmo de classificação de ML que leve em consideração a sequencia dos dados.

e) Análise dos resultados obtidos com o modelo criado e comparação do modelo criado com outros modelos já existentes.

1.2 MÉTODO DE PESQUISA

Iniciamos o trabalho com o estudo teórico de litofácies e a importância da utilização de modelo de ML para a classificação automática de litofácies. Além disso, também realizamos um estudo sobre IA e seus subcampos, partindo de AM, passando pela RNN e chegando a LSTM.

Em seguida realizamos o levantamento da literatura, na área de classificação de litofácies, onde fizemos uma análise crítica dos trabalhos que tentam solucionar o problema de classificação de litofácies utilizando modelos de AM.

Finalmente implementamos o algoritmo de ML proposto, apresentamos o banco de dados utilizado e seguimos com a conclusão e comparação dos resultados obtidos pelo algoritmo proposto e outros algoritmos da literatura.

1.3 ESTRUTURA DO TRABALHO

O capítulo 2 aborda alguns conhecimentos necessários para o entendimento do trabalho, relacionado-os com o problema em questão.

O capítulo 3 apresenta trabalhos já existentes na área de classificação de litofácies por modelos de AM. Neste capítulo também é realizado a análise crítica entre os trabalhos apresentados.

O capítulo 4 detalha o modelo escolhido para o cumprimento dos objetivos de pesquisa. Também são descritos o ambiente e o linguagem de programação utilizada e como a rede neural foi escolhida e configurada.

O capítulo 5 apresenta os dados utilizados e como o tratamento dos dados foi realizado.

Por fim, o capítulo 6 analisa os resultados obtidos realizando a comparação entre o modelo proposto e outros modelos da literatura. Além disso, é sugerido possíveis trabalhos futuros e feito as considerações finais.

2 FUNDAMENTAÇÃO TEÓRICA

Aqui serão apresentados os conceitos principais para a realização deste trabalho. Começaremos apresentando a definição de fácies sedimentares, litofácies, e uma breve descrição dos atributos utilizados para a classificação de litofácies. Em seguida, será feito uma breve introdução aos conceitos de IA, ML, DL, RNN e LSTM.

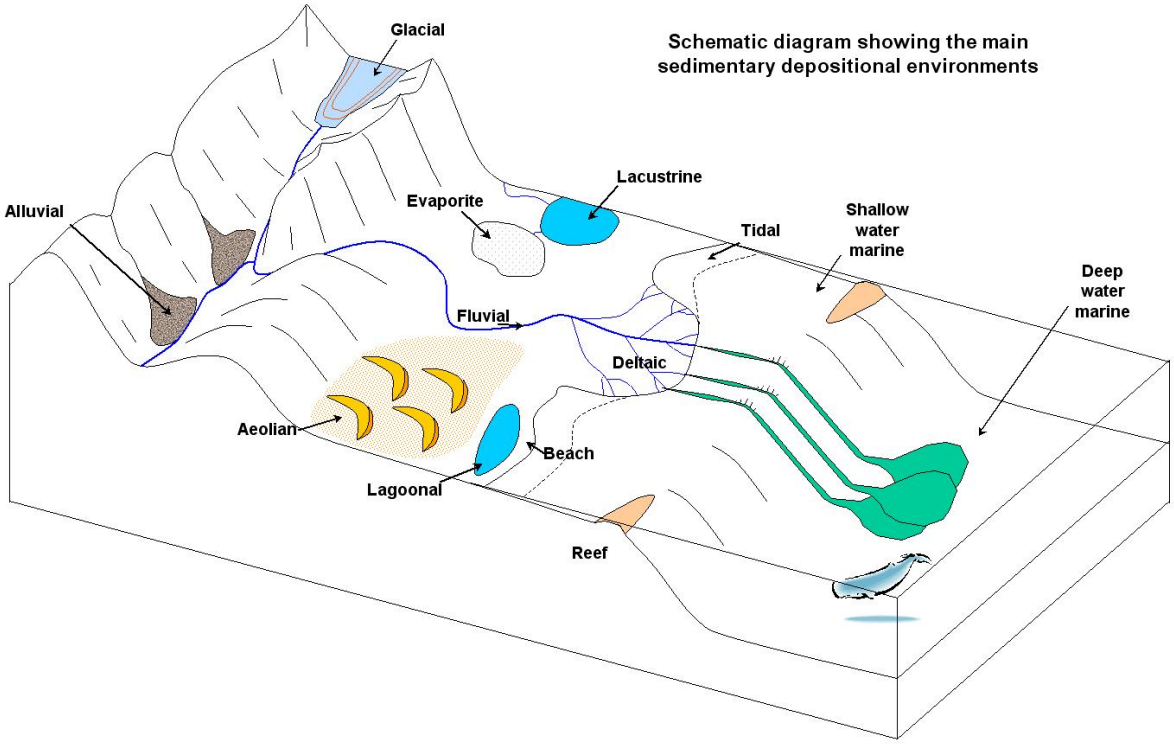
2.1 FÁCIES SEDIMENTARES

Um ambiente de deposição (ou ambiente sedimentar) (Figura 1) é um tipo específico de local no qual os sedimentos são depositados, como um canal de riacho, um lago ou o fundo do oceano profundo (COLLEGE, s.d.). Rochas sedimentares podem ser formadas apenas onde os sedimentos são depositados por tempo suficiente para se compactar e cimentar em camadas ou estratos duros. A sedimentação normalmente ocorre em áreas onde o sedimento permanece intacto por muitos anos em bacias sedimentares. Enquanto algumas dessas bacias são pequenas, outras ocupam milhares de quilômetros quadrados e geralmente possuem vários ambientes locais deposicionais diferentes. Fatores físicos, químicos e biológicos influenciam esses ambientes e as condições que eles produzem determinam em grande parte a natureza dos sedimentos que se acumulam. Vários ambientes locais diferentes (sedimentares) podem, portanto, existir lado a lado dentro de uma bacia, à medida que as condições mudam lateralmente; as rochas sedimentares que, em última instância são ali produzidas, podem estar relacionadas a esses ambientes deposicionais. Essas rochas sedimentares diferentes, mas contemporâneas e justapostas, são conhecidas como fácies sedimentares (BRITANNICA, s.d.).

Por exemplo, uma fácies de praia geralmente pode ser distinguida de uma fácies plana de maré, ambas as quais foram depositadas ao mesmo tempo adjacentes uma à outra. Em comparação com a fácies da praia, a fácies plana da maré terá um tamanho médio de grão de sedimento menor, mais fósseis de bioturbação, conterá camadas cruzadas e ondulações criadas por correntes de maré e terá mais moluscos ou outros fósseis de águas rasas preservados em seu lugar original, em forma ininterrupta. Não haverá uma fronteira nítida entre as duas fácies preservadas no registro sedimentar. Em vez disso, a fronteira entre eles será uma zona com camadas de sedimentos que se interpenetram e se misturam lateralmente de uma fácies para outra (COLLEGE, s.d.).

Abaixo (Figura 2) está um diagrama simplificado de três fácies sedimentares adjacentes entre si: uma fácies plana de praia e maré (combinadas), uma porção marinha ou perto da costa de uma plataforma continental e uma plataforma carbonática ou recife de alto mar. Os sedimentos da fácies plana da praia e da maré são principalmente areia, a fácies da baía é principalmente lama, e a fácies do recife é composta principalmente por conchas e corais que são feitos de minerais carbonáticos. Se esses sedimentos forem enterrados e litificados em rochas sedimentares, as areias da praia se transformam em

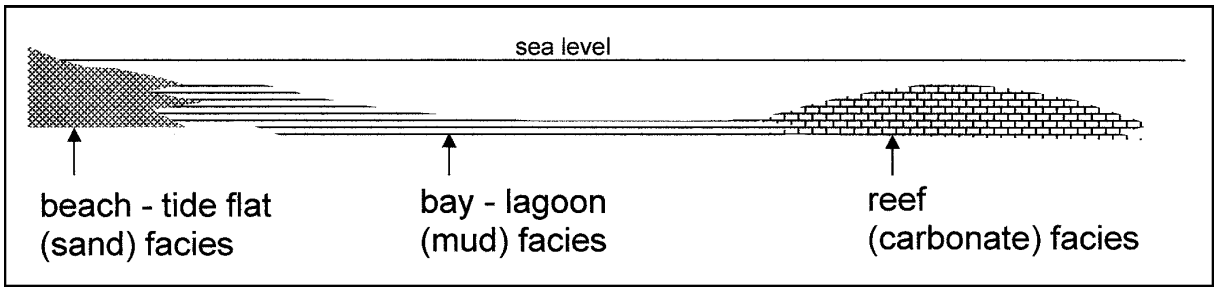
Figura 1 – Ambientes Deposicionais



Fonte: (COLLEGE, s.d.)

arenito, a lama da baía se transforma em xisto e os sedimentos do recife se transformam em calcário (COLLEGE, s.d.).

Figura 2 – Fácies Sedimentares



Fonte: (BRITANNICA, s.d.)

2.1.1 Litofácies

Existem várias maneiras de descrever ou designar fácies sedimentares. Ao observar as características físicas (ou litológicas) primárias, é possível reconhecer litofácies. Os atributos biológicos (ou mais corretamente, paleontológicos) - os fósseis - definem biofácies. Ambos são o resultado direto da história deposicional da bacia. Ao atribuir modos de origem a diferentes fácies (ou seja, interpretando as litofácies ou biofácies), pode-se visualizar um sistema genético de fácies (BRITANNICA, s.d.).

2.2 ATRIBUTOS DE LITOFÁCIES

Levantamentos de cabo de aço determinam propriedades físicas dentro e além da parede de um poço por dispositivos conectados a um cabo ou cabo de aço. As condições geológicas de subsuperfície e as características de engenharia podem ser derivadas direta ou indiretamente de uma ampla variedade de propriedades mensuráveis disponíveis por levantamento de cabos de aço. Os dados de vários métodos são frequentemente combinados para avaliar uma única característica geológica ou de engenharia (INTERIOR BUREAU OF RECLAMATION, 1998). A *Force 2020 Machine Learning Competition* utiliza 24 atributos de registro de poços, além de uma coluna de confiança de interpretação, e outra coluna contendo a classe das litofácies. A seguir apresentaremos uma breve descrição sobre os cada um deles (SCHLUMBERGER, s.d.).

- Medida qualitativa de confiança de interpretação: 1 para alta, 2 para média, 3 para baixa;
- Litofácies interpretadas: classe da litologia;
- Medição de resistividade de leitura profunda/média/rasa/micro: uma medição da resistividade da formação feita no tubo de perfuração a uma frequência na faixa de 100 kHz a 10 GHz, mais comumente 2 MHz. Na prática, vários transmissores podem ser usados para obter diferentes profundidades de investigação e obter compensação de poço;
- Medição de resistividade de zona lavada: o volume próximo à parede do poço no qual todos os fluidos móveis foram deslocados pelo filtrado de lama;
- Potencial espontâneo: potencial elétrico de ocorrência natural (estático) na Terra. Os potenciais espontâneos são geralmente causados pela separação de carga na argila ou outros minerais, pela presença de uma interface semipermeável que impede a difusão de íons através do espaço dos poros das rochas, ou pelo fluxo natural de um fluido condutor (água salgada) através das rochas;

- Tempo de trânsito de onda compressiva e de onda cisalhante (seg/ft): um tipo de registro acústico que exibe o tempo de viagem das ondas em relação à profundidade. Os perfis sônicos são normalmente registrados puxando uma ferramenta em um cabo de aço até o furo de poço. A ferramenta emite uma onda sonora que viaja da fonte para a formação e de volta para um receptor;

- Registro de porosidade de nêutrons: referindo-se a um registro de porosidade com base no efeito da formação em nêutrons rápidos emitidos por uma fonte. Uma vez que o hidrogênio é encontrado principalmente nos fluidos dos poros, o registro da porosidade do nêutron responde principalmente à porosidade. O registro é calibrado para ler a porosidade correta assumindo que os poros são preenchidos com água doce e para uma dada matriz (calcário, arenito ou dolomita). O registro da porosidade do nêutron é fortemente afetado por argila e gás;

- Registro do fator fotoelétrico: um registro das propriedades de absorção fotoelétrica. Como os fluidos têm números atômicos muito baixos, eles têm muito pouca influência, de modo que o fator fotoelétrico é uma medida das propriedades da matriz da rocha. Arenitos têm baixo fator fotoelétrico, enquanto dolomitos e calcários têm alto fator fotoelétrico. Argilas, minerais pesados e minerais contendo ferro têm alto fator fotoelétrico. Assim, o registro é muito útil para determinar a mineralogia;

- Registro de raios gama: um registro da radioatividade natural total. Os xistos e as argilas são responsáveis pela maior parte da radioatividade natural, de modo que o registro de raios gama costuma ser um bom indicador dessas rochas. No entanto, outras rochas também são radioativas, notadamente alguns carbonatos e rochas ricas em feldspato. O perfil também é usado para correlação entre poços, para correlação de profundidade entre orifícios abertos e revestidos e para correlação de profundidade entre execuções de perfilagem;

- Registro de densidade aparente: uma medição da densidade aparente da formação, com base na redução no fluxo de raios gama entre uma fonte e um detector devido ao espalhamento Compton. A medição responde à densidade média do material entre a fonte e o detector;

- Registro de correção de densidade: uma correção para variações na densidade ou espessura da crosta terrestre. As correções isostáticas são comumente aplicadas aos dados de gravidade e são feitas de acordo com um modelo específico para isostasia;

- Registro do calibrador/registro de calibrador diferencial: uma representação do diâmetro medido de um poço ao longo de sua profundidade;

- Tamanho do poço: O próprio tamanho do poço, incluindo o poço aberto ou parte não revestida do poço. O furo de poço pode se referir ao diâmetro interno da parede do furo de poço, a face da rocha que limita o furo perfurado;
- Taxa média de penetração: a velocidade média na qual a broca pode quebrar a rocha sob ela e, assim, aprofundar o furo de poço.
- Registro de raio gama espectral: é a última variante do registro de raios gama. A energia do raio gama captado pelo detector é proporcional ao brilho do pulso de luz que ele produz, e esse brilho, por sua vez, determina o tamanho do pulso elétrico produzido pelo fotomultiplicador. A energia dos raios gama é determinada por qual elemento os emitiu. As medições de raios gama espectrais oferecem várias vantagens. Eles podem ajudar na digitação com argila (PETROWIKI, s.d.);
- Peso da lama de perfuração: a massa por unidade de volume de um fluido de perfuração;
- Taxa de penetração: a velocidade na qual a broca pode quebrar a rocha sob ela e, assim, aprofundar o furo de poço;
- Profundidade medida: O comprimento do furo de poço, como se determinado por uma régua de medição;
- Localização X da amostra: localização da amostra na coordenada X;
- Localização Y da amostra: localização da amostra na coordenada Y;
- Profundidade Z (TVDSS) da amostra: localização da amostra na profunzidade Z. Profundidade vertical verdadeira SS (TVDSS do inglês True Vertical Depth SS).

2.3 APLICAÇÃO DE ML NA CLASSIFICAÇÃO DE LITOFÁCIES

Para resolver o problema de alto custo e tempo despendido na classificação de litofácies, vários estudos incorporaram algoritmos de ML, alimentado por dados e de baixo custo, usando apenas registros de poços para classificar fácies. Propriedades físicas de perfis de poço são usadas como atributos, enquanto fácies interpretadas de testemunhos são usadas como categoria verdadeira. Classificar fácies com base exclusivamente em características de perfis de poço é um desafio devido às suas diferenças nas resoluções, bem como valores de características sobrepostos para diferentes fácies. Embora as abordagens de estudos anteriores sejam robustas e capazes de prever as fácies com certo grau de precisão, as informações geológicas e as sequências de fácies estão ausentes, o que faz com que os modelos prevejam sequências irrealísticas de fácies. É reconhecido que as fácies em camadas vizinhas são correlacionadas e os padrões de empilhamento de fácies são

significativos para a interpretação geológica. Um modelo de ML baseado em sequência é, portanto, mais apropriado do que a abordagem tradicional de classificação multiclasse usada em estudos anteriores. Ele detecta naturalmente a sequência aprendendo com as fácies anteriores antes de fazer uma previsão (JAIKLA *et al.*, 2019).

Sendo assim, este trabalho pretende resolver o problema de custo associado a classificação manual de litofácies usando um algoritmo de ML capaz de levar em consideração a sequencia das camadas de rocha. Em seguida será apresentado em maiores detalhes os conceitos básicos de algoritmos de ML que levam em consideração as sequencias dos dados.

2.4 INTELIGÊNCIA ARTIFICIAL

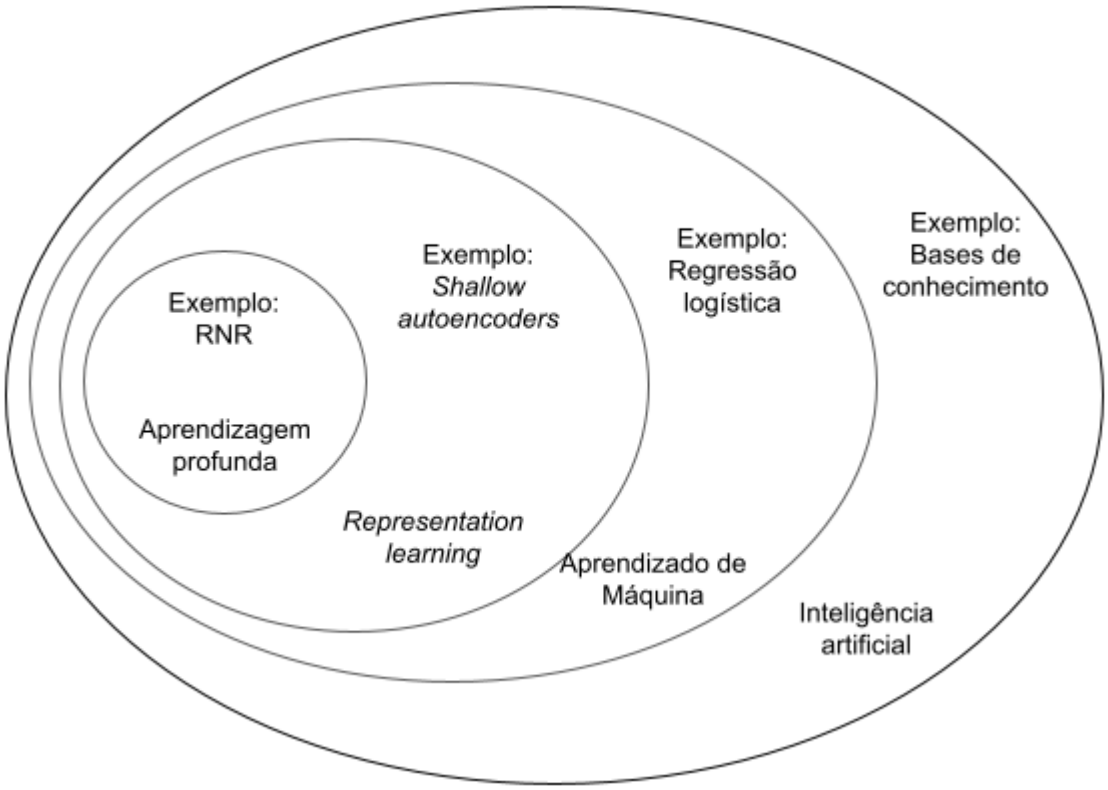
Antes de entrar em detalhes sobre RNN, vamos primeiramente definir inteligência artificial (IA) e algumas de suas subáreas.

Inteligência artificial é a automação de atividades que nós associamos com o pensamento humano, atividades como tomada de decisão, resolução de problemas e aprendizado. A IA é um dos campos mais recentes na ciência e engenharia. O termo originou-se em 1956 e seus estudos começaram logo após a Segunda Guerra Mundial. Atualmente, a IA engloba uma grande variedade de atividades, que vão do geral (aprendizagem e percepção) ao específico, como jogar xadrez, provar teoremas matemáticos, escrever poesia, dirigir um carro em uma rua movimentada e diagnosticar doenças. Essa variedade de atividades levou ao desenvolvimento da hierarquia de subáreas de IA, ilustradas na Figura 3 junto com exemplos representativos de cada subárea (RUSSELL; NORVIG, 2010). A seguir, a hierarquia será apresentada, partindo do termo mais abrangente ML, seguido por DL, RNN e chegando ao termo mais específico LSTM.

2.4.1 Aprendizado de Máquina

As dificuldades enfrentadas por sistemas que dependem do conhecimento manualmente codificados sugerem que os sistemas de IA precisam ser capazes de adquirir seu próprio conhecimento, extraindo padrões de dados brutos. Esse recurso é conhecido como ML. A introdução de ML permitiu aos computadores lidar com problemas que envolvem o conhecimento do mundo real, assim como tomar decisões que parecem subjetivas. Um algoritmo de ML é um algoritmo capaz de aprender com dados. De acordo com (MITCHELL, 1997) dizemos que um programa de computador aprende com experiência E com relação a alguma classe de tarefas T e medida de desempenho P, se seu desempenho nas tarefas em T, conforme medido por P, melhora com a experiência E (GOODFELLOW; BENGIO; COURVILLE, 2016). Por exemplo, um modelo que aprende a tarefa de classificar rochas pode melhorar seu desempenho, medido pela acurácia da classificação, obtendo experiência a partir de dados de rochas já previamente categorizadas.

Figura 3 – Hierarquia das subáreas da inteligência artificial.



Fonte: adaptado de (GOODFELLOW; BENGIO; COURVILLE, 2016)

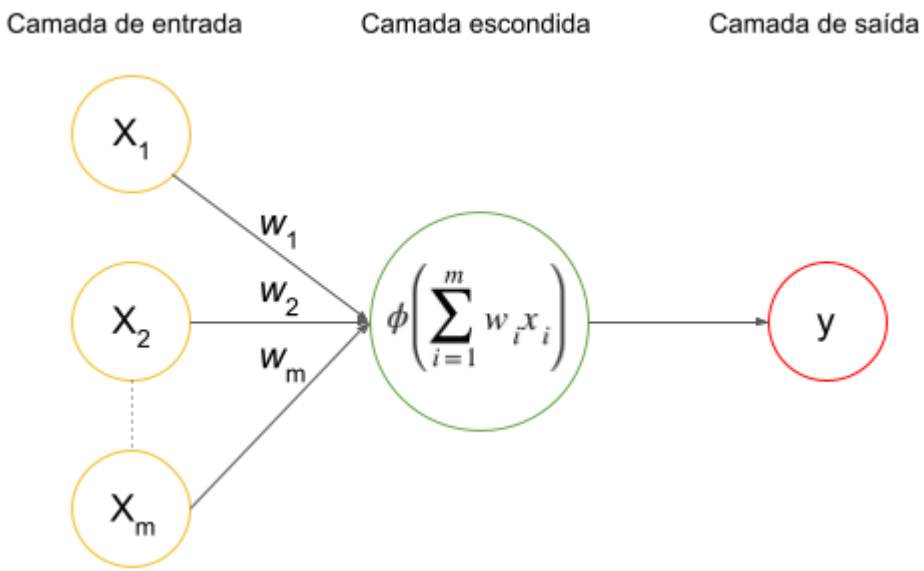
2.4.1.1 Aprendizagem Profunda

Os algoritmos tradicionais de ML funcionam bem em uma ampla variedade de problemas importantes. Eles não conseguiram, no entanto, resolver os problemas centrais de IA, como reconhecer a fala ou reconhecer objetos. O desenvolvimento de DL foi motivado em parte pela falha dos algoritmos tradicionais em generalizar bem essas tarefas de IA. O DL é um tipo particular de ML que atinge grande poder e flexibilidade ao representar o mundo como uma hierarquia aninhada de conceitos, com cada conceito definido em relação a conceitos mais simples e representações mais abstratas computadas em termos de conceitos menos abstratos (GOODFELLOW; BENGIO; COURVILLE, 2016).

Uma *deep feedforward network* sem camadas ocultas é chamada de *feedfoward network*. O objetivo de uma *feedfoward network* (Figura 4) é de aproximar uma função f^*. Por exemplo, para um classificador $y = f^*(x)$ que mapeia uma entrada x para uma categoria y. Uma *feedfoward network* define um mapeamento $y = f(x; w)$ e aprende o valor dos parâmetros w que resultam na melhor aproximação da função (GOODFELLOW; BENGIO; COURVILLE, 2016).

O nodo da *neural network* realiza a soma dos pesos w_i e dos atributos de entrada x_i e passa o resultado da soma como parâmetro para a função de ativação ϕ. A *rectified*

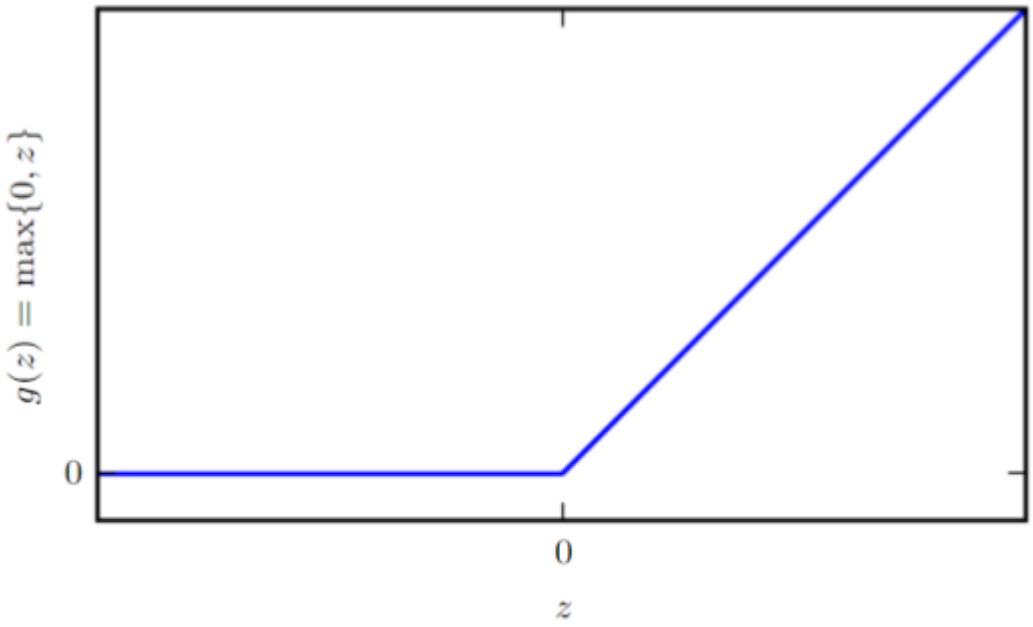
Figura 4 – Feedforward Neural Network.



Fonte: autor

linear unit (ReLU) (Figura 5) é a função de ativação padrão recomendada para uso com a maioria das *feedforward neural networks*. Aplicar essa função à saída de uma transformação linear produz uma transformação não linear (GOODFELLOW; BENGIO; COURVILLE, 2016).

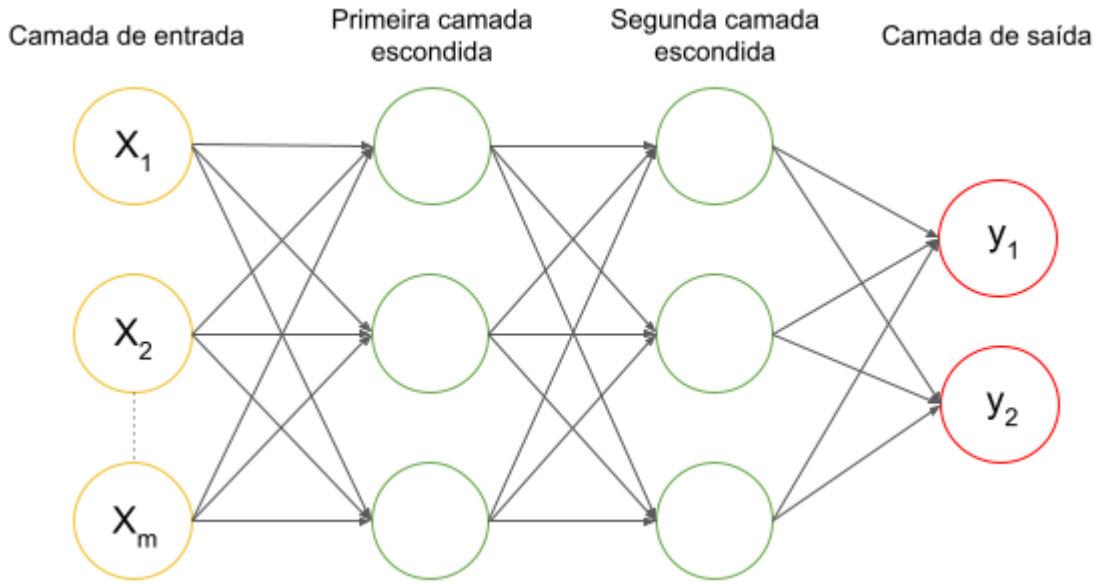
Figura 5 – Função de ativação ReLU.



Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016)

Uma *deep feedforward network* é o exemplo mais típico de um modelo de DL. As *deep feedforward neural network* (Figura 6) são chamadas de redes pois são normalmente representadas pela composição de muitas funções diferentes. O modelo está associado a um grafo acíclico direcionado que descreve como as funções são compostas juntas. Por exemplo, podemos ter três funções $f^{(1)}$, $f^{(2)}$ e $f^{(3)}$ conectadas em uma cadeia, para formar $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$. Essas estruturas em cadeia são as estruturas mais comumente usadas de redes neurais. Nesse caso, $f^{(1)}$ é chamada de primeira camada da rede, $f^{(2)}$ é chamada de segunda camada e assim por diante (GOODFELLOW; BENGIO; COURVILLE, 2016).

Figura 6 – Deep Feedforward Neural Network.



Fonte: autor

O comprimento total da rede fornece a profundidade do modelo. O nome "aprendizagem profunda" surgiu dessa terminologia. A camada final de uma rede feedforward é chamada de camada de saída. Durante o treinamento da rede neural, dirigimos $f(x)$ para corresponder a $f^*(x)$. Os dados de treinamento nos fornecem exemplos aproximados e ruidosos de $f^*(x)$ avaliados em diferentes pontos de treinamento. Cada exemplo x é acompanhado por um rótulo $y \approx f^*(x)$. Os exemplos de treinamento especificam diretamente o que a camada de saída deve fazer em cada ponto x; ela deve produzir um valor que seja próximo de y. O comportamento das outras camadas não é especificado diretamente pelos dados de treinamento. O algoritmo de aprendizado deve decidir como usar essas camadas para produzir a saída desejada, mas os dados de treinamento não dizem o que cada camada individual deve fazer. Em vez disso, o algoritmo de aprendizagem deve decidir como usar essas camadas para melhor implementar uma aproximação de f^*. Como os dados de treinamento não mostram a saída desejada para cada uma dessas camadas, elas são

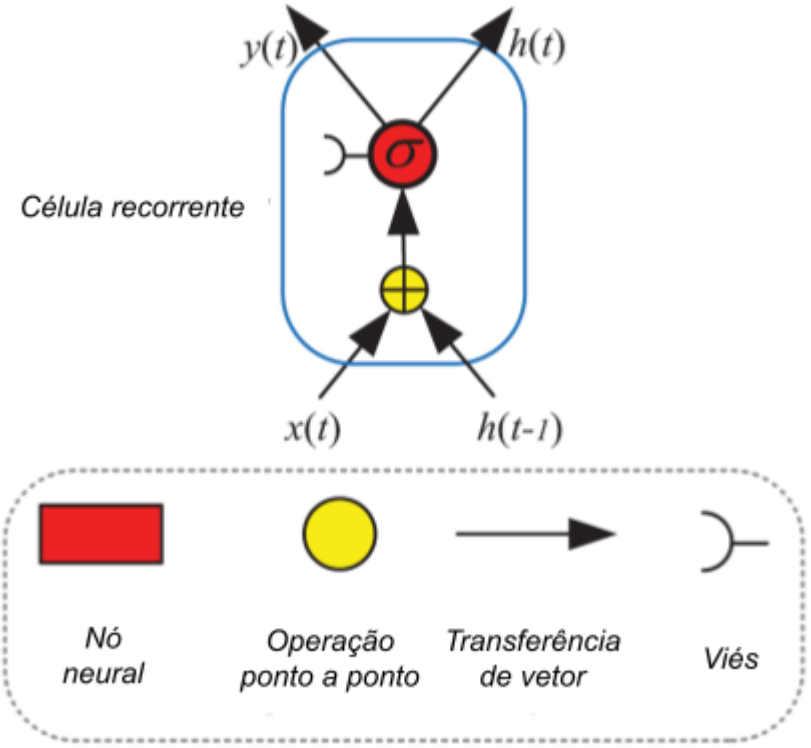
chamadas de camadas ocultas (GOODFELLOW; BENGIO; COURVILLE, 2016).

2.4.1.1.1 Rede Neural Recorrente

Os modelos descritos acima são chamados de *feedforward* pois a informação flui através da função que está sendo avaliada de x através dos cálculos intermediários usados para definir f e, finalmente, para a saída y. Ou seja, não há conexões de *feedback* nas quais as saídas do modelo são realimentadas. Quando as redes neurais *feedforward* são estendidas para incluir conexões de *feedback*, elas são chamadas de RNN (GOODFELLOW; BENGIO; COURVILLE, 2016).

As RNNs têm sido amplamente adotadas em áreas de pesquisa relacionadas com dados sequenciais, como texto, áudio e vídeo. Nas RNNs, as camadas recorrentes ou camadas ocultas consistem em células recorrentes cujos estados são afetados tanto pelos estados passados quanto pela entrada atual a partir de conexões de *feedback*. Normalmente RNNs são redes que consistem em células recorrentes padrão, como células sigma e células tanh. A Figura 7 mostra um esquema da célula sigma recorrente padrão. A expressão matemática da célula sigma recorrente padrão é definida pela equação 1, onde x_t, h_t, and y_t denotam a entrada, a informação recorrente e a saída da célula no tempo t, respectivamente; W_h e W_x são os pesos; e b é o viés (YU, Yong *et al.*, 2019).

Figura 7 – Esquema da célula sigma recorrente padrão.



Fonte: adaptado de (YU, Yong *et al.*, 2019)

$$\begin{aligned} h_t &= \sigma(W_h h_{t-1} + W_x x_t + b) \\ y_t &= h_t \end{aligned} \tag{1}$$

As células recorrentes padrão obtiveram algum sucesso em alguns problemas. No entanto, RNNs que consistem em células sigma ou células tanh são incapazes de aprender as informações relevantes dos dados de entrada quando o intervalo de entrada é grande: à medida que a lacuna entre as entradas relacionadas aumenta, é difícil aprender as informações de conexão (YU, Yong *et al.*, 2019).

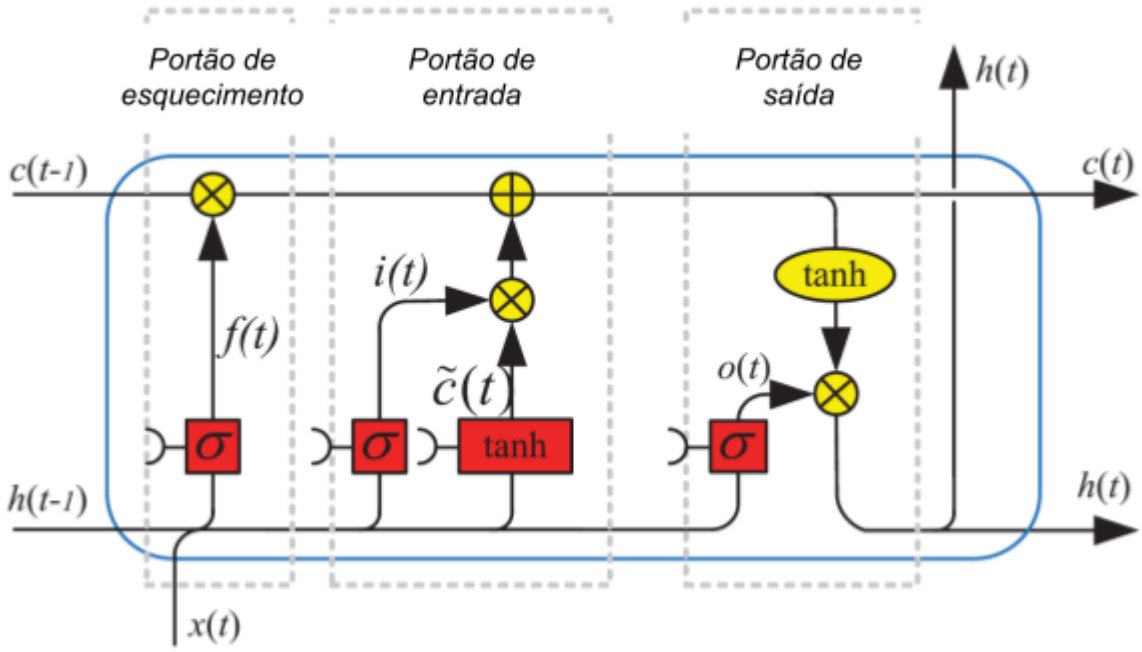
Para lidar com o problema das "dependências de longo prazo", Hochreiter e Schmidhuber (1997) propuseram a célula LSTM. Desde a sua introdução, quase todos os resultados empolgantes baseados em RNNs foram alcançados pela LSTM. A LSTM se tornou o foco do aprendizagem profunda. A capacidade de memorização da célula recorrente padrão foi aumentada ao introduzir um "portão" na célula. Desde este trabalho pioneiro, as LSTMs foram modificadas e popularizadas por muitos pesquisadores. As variações incluem LSTM sem um portão de esquecimento, LSTM com um portão de esquecimento e LSTM com uma conexão de olho mágico. Em seguida, apresentamos o modelo LSTM com um portão de esquecimento (YU, Yong *et al.*, 2019).

A Figura 8 apresenta as conexões internas de uma LSTM com portão de esquecimento. Com base nas conexões mostradas na Figura 8, a célula LSTM pode ser expressa matematicamente pela equação 2, onde c_t denota o estado da LSTM, W_i, $W_{\tilde{c}}$, e W_o são os pesos, e o operador "$\cdot$"denota a multiplicação ponto a ponto de dois vetores. Ao atualizar o estado da célula, o portão de entrada pode decidir quais novas informações podem ser armazenadas no estado da célula, o portão de saída decide quais informações podem ser enviadas com base no estado da célula e o portão de esquecimento pode decidir quais informações serão descartadas do estado da célula. Quando o valor do portão de esquecimento, f_t, é 1, ele mantém essa informação, por outro lado, um valor de 0 significa que o portão se livra de todas as informações (YU, Yong *et al.*, 2019).

$$\begin{aligned} f_t &= \sigma(W_{fh} h_{t-1} + W_{fx} x_t + b_f) \\ i_t &= \sigma(W_{ih} h_{t-1} + W_{ix} x_t + b_i) \\ \tilde{c}_t &= tanh(W_{\tilde{c}h} h_{t-1} + W_{\tilde{c}x} x_t + b_{\tilde{c})} \\ c_t &= f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \\ o_t &= \sigma(W_{oh} h_{t-1} + W_{ax} x_t + b_o) \\ h_t &= o_t \cdot tanh(c_t) \end{aligned} \tag{2}$$

Em suma, este capítulo apresentou os principais conceitos relacionados ao domínio da classificação de litofácies, assim como os conceitos sobre algoritmos de ML. A seguir, tais conceitos serão utilizado para a realização da revisão sistemática da literatura sobre classificação de litofácies utilizando algoritmos de ML.

Figura 8 – Arquitetura da LSTM com um portão de esquecimento..



Fonte: adaptado de (YU, Yong *et al.*, 2019)

3 REVISÃO SISTEMÁTICA DA LITERATURA

Neste capítulo é apresentado quatro artigos que aplicaram técnicas para a classificação de litofácies através de modelos de ML. Os trabalhos foram selecionados a fim de expandir a visão sobre o que vem sendo aplicado nessa área do conhecimento. Ao final do capítulo é realizado a comparação entre os trabalhos.

A busca por artigos foi realizada no Google Scholar utilizando as palavras chaves *facies*, *classification*, *well logs*, *machine learning*. A pesquisa retornou 1.120 artigos. Através de uma leitura dos resumos de 38 desses artigos, foram selecionados 17 artigos para uma leitura mais profunda. Por fim, foram selecionados 4 artigos para serem apresentados a seguir.

3.1 LITHOLOGICAL FACIES CLASSIFICATION USING DEEP CONVOLUTIONAL NEURAL NETWORK

No trabalho apresentado por (IMAMVERDIYEV; SUKHOSTAT, 2019) uma arquitetura baseada em CNN unidimensional (1D-CNN), que é treinada usando vários algoritmos de optimização, é proposta para a classificação de litofácies. Os algoritmos de optimização testados foram o Adagrad, Adadelta e Adamax.

A arquitetura do modelo 1D-CNN consiste em uma camada de entrada, quatro camadas convolucionais com ReLU como função de ativação não linear, duas camadas *maxpool* e três camadas totalmente conectadas. A última camada é a camada de saída, que atribui a categoria aos dados de entrada (Figura 9). Os dados de entrada foram divididos em conjuntos de treinamento e validação (20% do conjunto de dados) para conduzir os experimentos. O tamanho do lote foi determinado como 10 e a função de perda foi escolhida como a entropia cruzada categórica. Cada resultado experimental foi obtido ao longo de 4000 épocas para fornecer comparações consistentes.

No conjunto de dados considerado, existem dados de 10 poços contendo um total de 4149 amostras. O conjunto de dados contém 9 tipos de litofácies: arenito não marinho (NS), siltito grosso não marinho (NCS), siltito fino não marinho (NFS), siltito marinho e xisto (MSS), argilito (M), wackestone (W), dolomita (D), packstone-grainstone (P_G), e calcário (P_AB). Os 6 atributos do conjunto de dados são: efeito fotoelétrico, raio gama, resistividade, diferença de porosidade de neutrondensidade, porosidade de densidade média de nêutrons, e variáveis de restrição geológica.

Uma análise comparativa do modelo proposto usando otimizadores Adagrad, Adadelta e Adamax com CNN, LSTM, máquina de vetores de suporte (SVM do inglês Support Vector Machine), e k-vizinhos mais próximos KNN) com base em acurácia e métricas de medida F são mostradas na Figura 10.

3.2 FACIESNET: MACHINE LEARNING APPLICATIONS FOR FACIES CLASSIFICATION IN WELL LOGS

Nesse trabalho apresentado por (JAIKLA *et al.*, 2019) é desenvolvido um modelo de classificação de fácies usando redes neurais recorrentes bidirecionais (BRNN do inglês Bidirectional Recurrent Neural Network) que incorporam sequências de fácies na previsão. Além de BRNNs, experimentou-se outra arquitetura adicionando camadas de decodificação e codificação de redes neurais convolucionais profundas (DCNN do inglês *Deep Convolutional Neural Network*) para extrair informações latentes antes de alimentá-las nas camadas de BRNNs.

Para a arquitetura de BRNN os experimentos incluíram o treinamento do conjunto de dados em modelos com 1, 2 e 3 camadas de BRNNs com 16, 32, 64 e 128 estados ocultos de unidades recorrentes bloqueadas (GRU do inglês Gated Recurrent Unit).

Já para a arquitetura com DCNNs e BRNN, a arquitetura que tem a maior acurácia e acurácia equilibrada no conjunto de teste consiste em 5 camadas de codificação e decodificação DCNNs seguidas por 2 camadas de BRNNs com 128 estados ocultos usando a *dice loss function*. Tal arquitetura foi chamada de *FaciesNet* (Figura 11).

O conjunto de dados possui 4 poços contendo um total de 170 amostras. O conjunto de dados contém 5 tipos de litofácies: arenito cimentado, heterolítico, lamito, arenito limpo, e arenito sujo. Os 6 atributos do conjunto de dados são: raios gama (GR), fração de volume de xisto (VSH), densidade (DEN), tempo de viagem sônica compressional (DTC), tempo de viagem sônica de cisalhamento (DTS), e porosidade de nêutrons (NEU).

Foi realizado uma análise comparandos os modelos de *Naive Bayes*, árvore de decisão, floresta aleatóri, BRNN e *FaciesNet*, com base em acurácia e acurácia equilibrada (Figura 10). Além disso, uma análise comparativa entre *Naive Bayes* e *FaciesNet*, com base em precisão, *recall* e *F1-score* são mostradas na Figura 10.

3.3 CHARACTERIZING ROCK FACIES USING MACHINE LEARNING ALGORITHM BASED ON A CONVOLUTIONAL NEURAL NETWORK AND DATA PADDING STRATEGY

No trabalho apresentado por (WEI *et al.*, 2019) é proposto uma arquitetura usando CNN com estratégias de preenchimento de dados. Inspirados pelo uso de CNN em imagens multicanal, foi testado três estratégias de preenchimento para expandir os conjuntos de dados bem medidos 1-D para 2-D para melhor capturar seus recursos inerentes.

O modelo possui duas camadas convolucionais. A primeira camada convolucional tem oito filtros 3 x 3 e a segunda dezesseis camadas 3 x 3. Cada camada convolucional é seguida por uma função de ativação de ReLU e uma camada de maxpool. Finalmente, a saída é conectada por uma camada totalmente conectada (Figura 14). Os três tipos de estratégias de preenchimento de dados usados foram: preenchimento igual, preenchimento

de descolamento e preenchimento aleatório.

No conjunto de dados, existem dados de 8 poços contendo um total de 4149 amostras. O dados possuem 9 tipos de litofácies: arenito não marinho (SS), siltito grosso não marinho (CSiS), siltito fino não marinho (FSiS), siltito marinho e xisto (SiSh), lamito (calcário) (MS), wackestone (calcário) (WS), dolomita (D), packstone grainstone (calcário) (PS), e bafflestone filoidalgal (calcário) (BS). Os 5 atributos dos dados são: raio gama natural (GR), média de porosidade de nêutron e densidade (PHI), porosidade de nêutron e diferença de porosidade de densidade (DeltaPHI), efeito fotoelétrico (PE), e base de registro de resistividade verdadeira aparente (ILD_log10).

Os resultados de precisão de nossas três estratégias diferentes de preenchimento são apresentados na Figura 15. Para comparação, a Figura 16 compara diferentes algoritmos sobre o mesmo conjunto de dados.

3.4 COMPARISON OF DIFFERENT MACHINE LEARNING ALGORITHMS FOR LITHOFACIES CLASSIFICATION FROM WELL LOGS

Nesse trabalho apresentado por (DELL'AVERSANA, 2019) é realizado a comparação de 6 algoritmos de aprendizagem supervisionada. Os algoritmos utilizados foram: indução de regra CN2, naive bayes, máquina de vetor de suporte, árvore de decisão, floresta aleatória, e impulso adaptativo.

O fluxo de trabalho incluiu as seguintes etapas principais: 1) análise de dados estatísticos; 2) treinamento de 6 algoritmos de classificação; 3) avaliação quantitativa do desempenho de cada algoritmo individual; 4) classificação simultânea de litofácies usando todos os 6 algoritmos; 5) comparação e relatórios de resultados.

O conjunto de dados possui 2 poços (poço A e poço B) contendo aproximadamente 21000 amostras cada poço. O conjunto de dados contém 5 tipos de litofácies: xisto prevalente, arenitos intercalados / xisto, arenitos intercalados / siltito, hidrocarboneto de média saturação, hidrocarboneto de baixa saturação, e hidrocarboneto de alta saturação. Os 6 atributos do conjunto de dados são: sônico, resistividade (Rdep), densidade (DEN), registro de nêutrons (NEU), absorção fotoelétrica (PEF), raios gama (GR), e potenciais espontâneos (SP).

Para avaliar a performance dos 6 algoritmos, foi utilizado as métricas: área abaixo da curva ROC (AUC), precisão de classificação (CA), *F1-score*, precisão, e *recall*. A Figura 17 mostra os resultados dos diferentes modelos para o poço A, e a Figura 18 para o poço B.

3.5 COMPARAÇÃO ENTRE OS TRABALHOS

Os resultados de (IMAMVERDIYEV; SUKHOSTAT, 2019) mostraram que as melhores previsões de resposta da rede foram obtidas para o caso Adagrad com um *F1-*

score total de 76.78. Além disso, o modelo 1D-CNN mostrou resultados mais precisos em comparação com SVM, KNN, RNN e LSTM. O modelo proposto também superou o SVM em mais de 50% do resultados. O pior desempenho, é observado para RNN. A aplicação da abordagem proposta para a classificação de fácies mostrou resultados significativos para fácies de origem marinha (dolomita, bafflestone de algas filoides e packstone-grainstone) e de origem continental (siltito grosso e arenito). O modelo 1D-CNN (Adagrad) proposto apresentou uma melhora estatisticamente significativa na classificação de fácies.

Embora nosso modelo proposto por (JAIKLA *et al.*, 2019) tenha menor acurácia e acurácia equilibrada do que as abordagens de outros estudos, o *FaciesNet* pode diferenciar entre fácies reservatório e não reservatório, que são arenito limpo e arenito sujo, bem como argilito e heterolítico. Além disso, ele dá previsões geológicas significativas e não sofre ao usar dados heterogêneos e desequilibrados. A arquitetura BRNN de maior precisão de teste de 64,11% foi a de 3 camadas de BRNN com 128 estados ocultos para cada camada. Entretanto, a arquitetura que teve a maior acurácia e acurácia equilibrada no conjunto de teste foi a *FaciesNet* consistindo em 5 camadas de codificação e decodificação DCNNs seguidas por 2 camadas de BRNNs com 128 estados ocultos usando a função de perda de dados. A acurácia da FaciesNet foi de 74,85%, e a precisão balanceada foi de 40,01%.

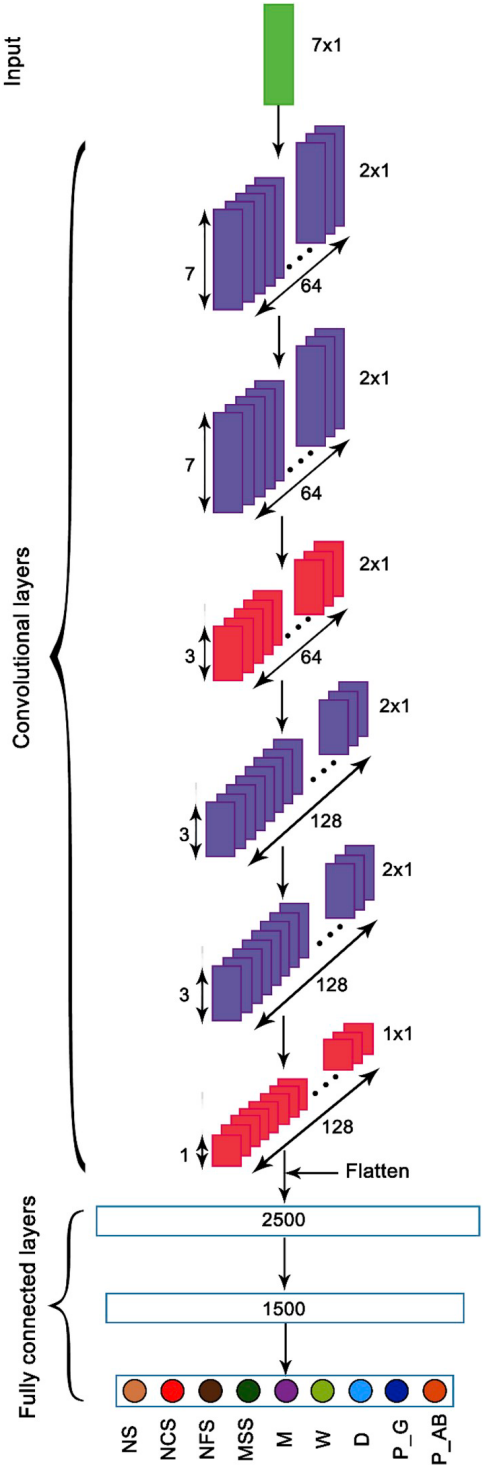
No trabalho de (WEI *et al.*, 2019) entre todos os resultados de classificação, o método de preenchimento de deslocamento igual atingiu 59,26% na pontuação F1, melhor do que os melhores resultados da CNN de 57% entre todos os concursos SEG ML 2016. Ao preencher os dados, a matriz de recursos 2D é útil para algoritmos CNN para detectar conexões entre os recursos e capturar as diferenças sutis entre fácies no processo de classificação. As estratégias de preenchimento adicionam dimensões aos dados, o que significa adicionar mais liberdade ao conjunto de dados e capturar melhor a correlação entre seus diferentes recursos. No entanto, o resultado final ainda não bate os melhores resultados alcançados pelo algoritmo de árvore impulsionada.

Por fim, no trabalho de (DELL'AVERSANA, 2019), os algoritmos de conjunto como floresta aleatória e impulso adaptativo parecem fornecer classificações/previsões ligeiramente mais confiáveis que naïve bayes, árvore de decisão, indução de regra CN2. O método SVM também demonstrou bom desempenho. No poço A a floresta aleatória atingiu o melhor resultado com 0.990 considerando o *F1-score*, enquanto a indução de regra CN2 atingiu o pior resultado com 0.868 de *F1-score*. Já no poço B o naive bayes atingiu o melhor resultado com 0.959 de *F1-score*, por outro lado, o pior resultado foi atingido pela SVM com *F1-score* de 0.744.

Três dos quatro trabalhos apresentados constroem seus modelos tendo como base arquiteturas de CNN. Em relação a RNN e LSTM, o trabalho de (IMAMVERDIYEV; SUKHOSTAT, 2019) mostrou que os piores resultados foram obtidos pela RNN e LSTM. Além disso, as arquiteturas mostradas por (JAIKLA *et al.*, 2019), reafirmam a hipótese de que incrementar uma RNN com uma CNN (culminando em uma *FaciesNet*) fez com

que a acurácia passasse de 64.11% para 74.85% e a acurácia balanceada passasse de 24.43% para 40.01%. O trabalho de (WEI *et al.*, 2019) apresenta que o modelo com maior sucesso na competição 2016 SEG ML foram árvores impulsionadas. As CNNs não ficaram entre as 10 primeiras colocadas, e não há nenhuma citação de experimento com RNN no quadro de resultados da competição. Por fim, o trabalho de (DELL'AVERSANA, 2019) fez uma apresentação de modelos que não utilizam redes neurais. Tais modelos obtiveram os melhores resultados de *F1-score*. Entretanto, como os trabalhos apresentados possuem conjuntos de dados diferentes, não é possível comparar os resultados obtidos por modelos em um determinado conjunto de dados com resultados obtidos por modelos utilizando um outro conjunto de dados.

Figura 9 – A arquitetura geral da 1D-CNN.



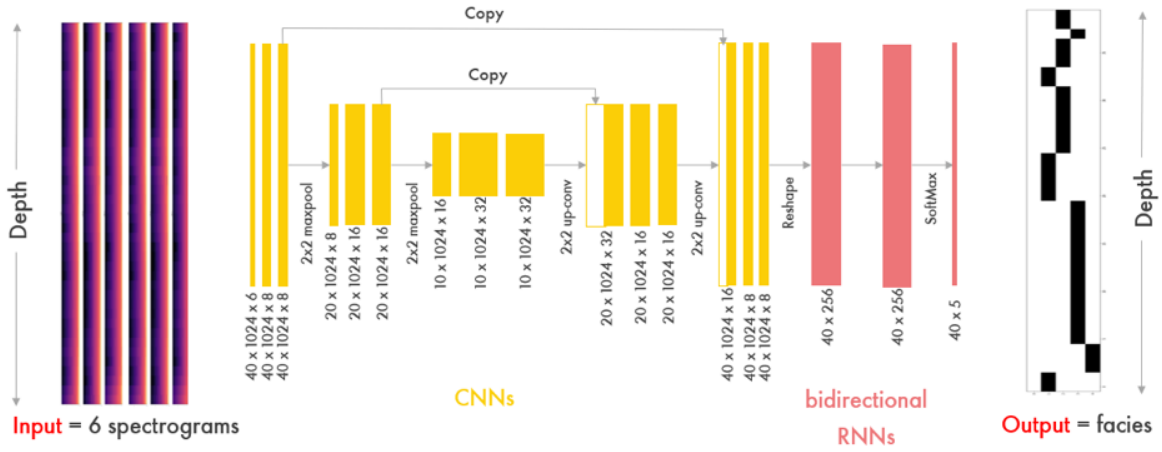
Fonte: (IMAMVERDIYEV; SUKHOSTAT, 2019)

Figura 10 – Comparação dos resultados da classificação de fácies usando a abordagem proposta com RNN, LSTM, SVM e KNN.

Method	Evaluation metrics	Facies									
		NS	NCS	NFS	MSS	M	W	D	P_G	P_AB	Total
RNN	Accuracy (%)	75.86	65.00	62.05	58.70	18.18	42.50	46.15	49.72	80.00	56.39
	F-measure (%)	53.01	67.29	64.78	55.10	10.53	43.97	43.64	57.69	52.46	55.22
LSTM	Accuracy (%)	81.25	69.36	68.92	50.85	44.00	50.39	72.73	58.13	83.87	63.13
	F-measure (%)	60.47	74.09	68.00	54.05	27.85	53.94	40.00	63.92	72.22	62.22
SVM	Accuracy (%)	78.00	75.23	76.06	73.47	58.82	62.07	83.33	76.34	**88.89**	73.73
	F-measure (%)	75.00	78.22	73.47	71.29	57.14	63.16	75.47	76.34	**93.02**	73.64
KNN	Accuracy (%)	73.08	74.31	74.66	64.62	56.25	58.87	83.33	77.27	76.74	71.33
	F-measure (%)	71.70	76.60	73.15	71.19	52.94	61.86	75.47	70.83	78.57	71.32
1D-CNN (Adagrad)	Accuracy (%)	**84.09**	**77.63**	**76.82**	**74.55**	**70.69**	**67.89**	**86.67**	**78.15**	**88.89**	**76.87**
	F-measure (%)	75.51	**80.38**	**76.57**	71.93	**70.69**	**67.58**	**80.00**	**78.81**	**93.02**	**76.78**
1D-CNN (Adadelta)	Accuracy (%)	79.59	75.77	76.47	**74.55**	61.36	62.07	84.62	77.50	85.97	74.58
	F-measure (%)	75.00	78.54	74.02	**77.36**	55.10	64.87	**80.00**	75.30	89.91	74.44
1D-CNN (Adamax)	Accuracy (%)	78.57	75.23	76.43	73.68	58.62	61.39	84.62	76.61	81.25	73.37
	F-measure (%)	**77.19**	78.10	73.29	75.00	59.13	58.77	77.19	76.92	88.64	73.20

Fonte: (IMAMVERDIYEV; SUKHOSTAT, 2019)

Figura 11 – Arquitetura *FaciesNet*.



Fonte: (JAIKLA *et al.*, 2019)

Figura 12 – Acurácia e acurácia balanceada da rede.

Model	Accuracy	Balanced accuracy
Naive Bayes	83.45%	56.97%
Decision Tree	83.69%	51.55%
Random Forest	84.88%	51.21%
BRNNs	64.11%	24.43%
FaciesNet	74.85%	40.01%

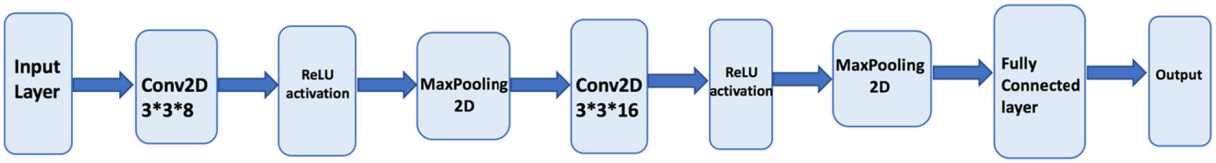
Fonte: (JAIKLA *et al.*, 2019)

Figura 13 – Comparação de precisão, *recall*, *F1-score*, da *FaciesNet* com *Naive Bayes*.

FaciesNet	Precision	Recall	F1 score
Cemented sandstone	0.8125	0.2718	0.4063
Heterolithic	0.1895	0.2000	0.1956
Mudstone	0.6209	0.5125	0.5621
Sandstone	0.8485	0.9133	0.8797
Dirty sandstone	0.1320	0.1029	0.1157
Naive Bayes	Precision	Recall	F1 score
Cemented sandstone	0.8913	0.8542	0.8723
Heterolithic	0	0	0
Mudstone	0.5745	1	0.7298
Sandstone	0.9315	0.9401	0.9358
Dirty sandstone	0	0	0

Fonte: (JAIKLA *et al.*, 2019)

Figura 14 – Arquitetura da CNN.



Fonte: (WEI *et al.*, 2019)

Figura 15 – Resultados das três diferentes estratégias de preenchimento, onde a acurácia é dada pelo *F1-score*.

	No. of filters of first convolutional layer	No. of filters of second convolutional layer	After 100 Epochs train accuracy (%)	Test set accuracy (%)
Equal padding	8	16	58.62	52.46
Shift padding	8	16	61.87	59.26
Random padding	8	16	40	22.92

Fonte: (WEI *et al.*, 2019)

Figura 16 – Resultados da competição 2016 SEG ML (https://github.com/seg/2016-ml-contest).

Team	F1 score	Algorithm	Language
LA Team (Mosser, de la Fuente)	0.641	Boosted trees	Python
Ispl (Bestagini, Tuparo, Lipari)	0.640	Boosted trees	Python
…	…	…	…
ShiangYong	0.570	ConvNet	Python
StoDIG	0.561	ConvNet	Python
…	…	…	…
BrendonHall	0.427	Support vector machine	Python

Fonte: (WEI *et al.*, 2019)

Figura 17 – Resultados da avaliação do desempenho dos modelos para o poço A.

	AUC	CA	F1	PRECISION	RECALL
METHOD					
Tree	0.971	0.943	0.943	0.944	0.943
Random Forest	1.000	0.990	0.990	0.990	0.990
CN2 Inducer	0.955	0.867	0.868	0.869	0.867
AdaBoost	0.979	0.965	0.965	0.965	0.965
SVM	0.988	0.981	0.980	0.981	0.981
Naive Bayes	0.998	0.956	0.956	0.958	0.956

Fonte: (DELL'AVERSANA, 2019)

Figura 18 – Resultados da avaliação do desempenho dos modelos para o poço B.

	AUC	CA	F1	PRECISION	RECALL
METHOD					
Tree	0.928	0.848	0.947	0.848	0.848
Random Forest	0.992	0.917	0.918	0.919	0.917
CN2 Inducer	0.909	0.767	0.768	0.769	0.767
AdaBoost	0.914	0.859	0.859	0.859	0.859
SVM	0.944	0.744	0.744	0.744	0.744
Naive Bayes	0.928	0.959	0.959	0.959	0.959

Fonte: (DELL'AVERSANA, 2019)

4 PROPOSTA

Neste capítulo, explicamos como os conceitos teóricos apresentados na seção de fundamentação teórica são colocados em prática na elaboração do método proposto. Primeiramente demonstramos alguns detalhes da implementação, tais como linguagem e bibliotecas. O código produzido até o momento pode ser encontrado em https://github.com/MatheusSchaly/TCC. Em seguida, apresentamos o conjunto de dados utilizados como entrada para o modelo proposto. Por fim, mostramos as métricas utilizadas para a avaliação dos modelos produzidos.

4.1 MÉTODO DE REDE NEURAL RECORRENTE

Dado a complexidade do formato dos dados de entrada de uma RNN, o modelo ainda não foi criado. Entretanto, temos a intenção de utilizar a linguagem Python com sua biblioteca de ML chamada TensorFlow, e a fim de facilitar a criação do modelo, utilizaremos a API chamada Keras que funciona como uma interface para o TensorFlow.

O modelo provavelmente seguirá um formato sequencial de camadas, contendo uma ou mais camadas de LSTM, *dropout* para regularização e redução de *overfitting*, e densas principalmente para a parte final da rede. A função de ativação para todas as camadas, exceto a última, provavelmente será a ReLU. Por outro lado, como queremos classificar uma rocha pertencente a uma única classe, a última camada terá a função de ativação *softmax*, visto que o resultado de tal função retorna probabilidades que possuem sua soma igual a 1.

4.2 CONJUNTO DE DADOS

Neste trabalho usaremos o conjunto de dados fornecido pela competição *Force 2020 Machine Learning Competition* (2020, s.d.). O conjunto de dados consiste em 1.170.511 exemplos de litofácies, 118 perfis de poços, sendo que 98 deles foram liberados pela competição para serem utilizados como dados de treino, os 20 perfis restantes foram liberados apenas ao final da competição a fim de realizar a avaliação dos modelos propostos. Além disso, o conjunto de dados possui 24 atributos de registro de poços, uma coluna de confiança de interpretação, e outra coluna contendo a classe das litofácies.

4.3 MÉTRICA DE AVALIAÇÃO

A competição possui sua própria métrica de avaliação. Em vez de penalizar cada previsão errada das litofácies, a competição decidiu usar uma matriz de penalidade customizada (Figura 19) derivada da entrada média de uma amostra representativa de geocientistas. Isso permite que previsões petrofisicamente irracionais sejam avaliadas por um grau de "erro"(Figura 20). A Figura 20 apresenta a métrica de avaliação, onde A é a matriz de

penalidade abaixo, N é o número de amostras, $\hat{y}_i$ é o rótulo litológico verdadeiro e y_i é o rótulo litológico previsto.

Figura 19 – Matriz de penalidade.

label \ prediction	Sandstone	Sandstone/Shale	Shale	Marl	Dolomite	Limestone	Chalk	Halite	Anhydrite	Tuff	Coal	Crystalline Basement
Sandstone	0	2	3.5	3	3.75	3.5	3.5	4	4	2.5	3.875	3.25
Sandstone/Shale	2	0	2.375	2.75	4	3.75	3.75	3.875	4	3	3.75	3
Shale	3.5	2.375	0	2	3.5	3.5	3.75	4	4	2.75	3.25	3
Marl	3	2.75	2	0	2.5	2	2.25	4	4	3.375	3.75	3.25
Dolomite	3.75	4	3.5	2.5	0	2.625	2.875	3.75	3.25	3	4	3.625
Limestone	3.5	3.75	3.5	2	2.625	0	1.375	4	3.75	3.5	4	3.625
Chalk	3.5	3.75	3.75	2.25	2.875	1.375	0	4	3.75	3.125	4	3.75
Halite	4	3.875	4	4	3.75	4	4	0	2.75	3.75	3.75	4
Anhydrite	4	4	4	4	3.25	3.75	3.75	2.75	0	4	4	3.875
Tuff	2.5	3	2.75	3.375	3	3.5	3.125	3.75	4	0	2.5	3.25
Coal	3.875	3.75	3.25	3.75	4	4	4	3.75	4	2.5	0	4
Crystalline Basement	3.25	3	3	3.25	3.625	3.625	3.75	4	3.875	3.25	4	0

Fonte: (2020, s.d.)

Figura 20 – Métrica de avaliação.

$$S = -\frac{1}{N}\sum_{i=0}^{N} \mathbf{A}_{\hat{y}_i y_i}$$

Fonte: (2020, s.d.)

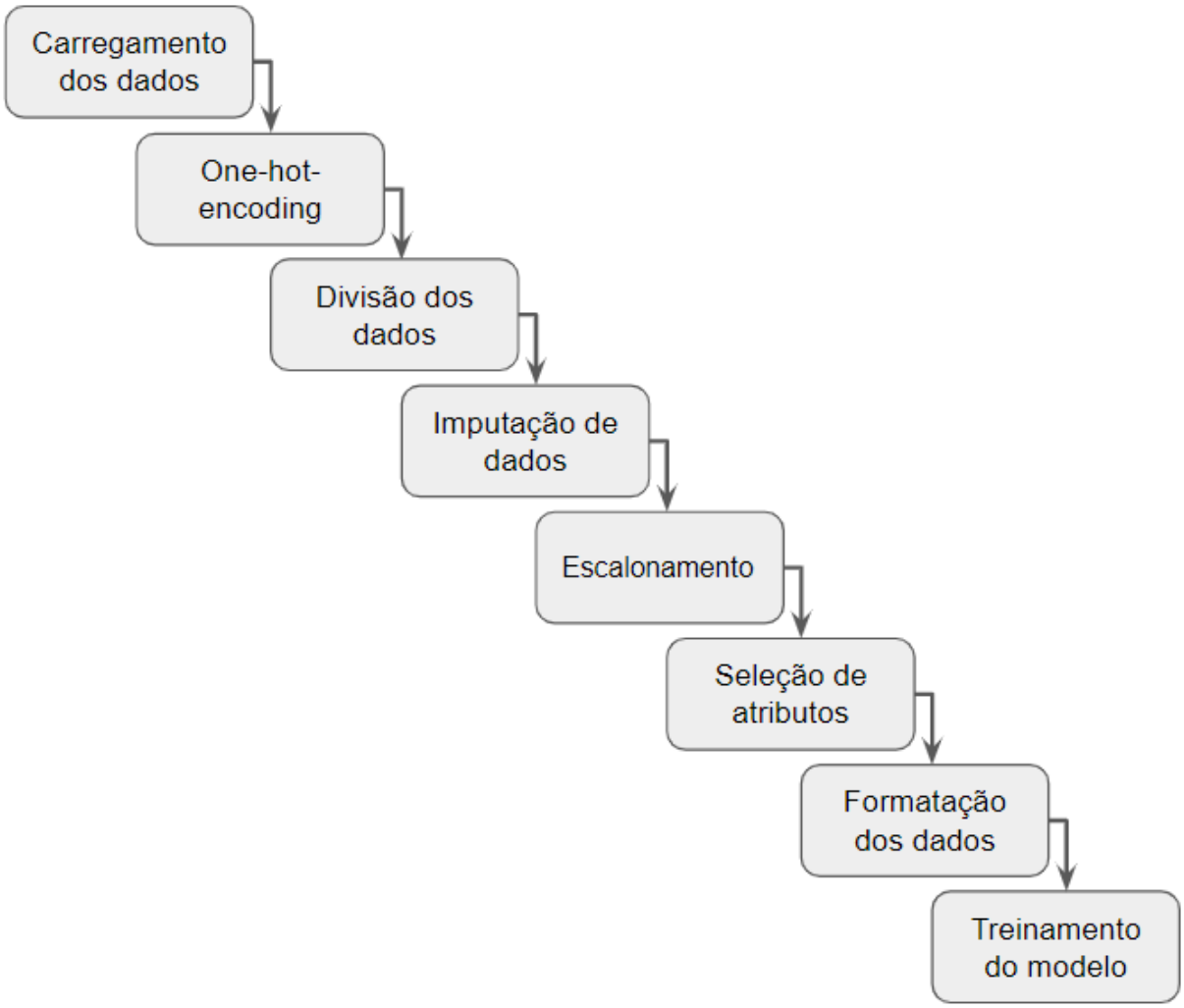
5 EXPERIMENTOS E RESULTADOS

Será apresentado apenas a parte do pré-processamento que foi parcialmente realizado.

5.1 PRÉ-PROCESSAMENTO

As etapas do pré-processamento dos dados segue a Figura 21 abaixo.

Figura 21 – Etapas pré-processamento dos dados.



Fonte: autor

O processamento dos dados foi realizado principalmente com as bibliotecas Numpy, Pandas e scikit-learn do Python. Primeiramente fazemos o carregamento dos dados, de treinamento da competição, em um pandas *dataframe*. Para tornar o pré-processamento mais rápido, reduzimos a quantidade de dados de exemplo de 1.170.511 para 1.000 exemplos aleatórios. Dado que o modelo espera atributos numéricos, usamos o *one-hot-encoding*

para converter as colunas categóricas em colunas numéricas. Em seguida, realizamos a divisão dos dados em treino e teste. Para acelerar a velocidade do modelo, assim como torná-lo mais robusto, usamos o *standard scaler* nos valores dos atributos, fazendo com que seus intervalos de valores variem entre aproximadamente -3 até aproximadamente +3. Como temos vários atributos com dados ausentes, utilizamos o modelo *Bayesian Ridge* para imputar os valores ausentes de um atributo baseado nos valores dos demais atributos. Depois, aplicamos diferentes métodos de seleção de atributos para diminuir o problema do *curse of dimensionality*, consequentemente diminuindo o risco de *overfitting*, aumentando a velocidade do método e sua acurácia. Foram aplicados os métodos de seleção de atributos chamados limite de variância, teste qui-quadrado, limiar de correlação, *backward elimination*, eliminação de recurso recursivo, e LassoCV. Para que possamos realizar o treinamento de uma RNN, precisamos formatar os dados de acordo com o que é esperado pelo modelo. Tanto a formatação dos dados quanto o treinamento do modelo ainda não foram realizados.

6 CONCLUSÕES

Neste trabalho apresentamos os conceitos básicos sobre fácies sedimentares, litofácies e uma breve descrição dos atributos utilizados pela *Force 2020 Machine Learning Competition* para a classificação de litofácies. Além disso, tivemos uma introdução aos conceitos de IA, ML, DL, RNN e LSTM. Em seguida, mostramos, e discutimos os resultados, de quatro artigos que aplicaram técnicas para a classificação de litofácies através de modelos de ML. Apresentamos também sugestões de modelos que poderão ser usados na solução do problema, o seu conjunto de dados, e a métrica de avaliação que o modelo utilizará. Por fim, falamos sucintamente sobre o processo de pré-processamento dos dados.

6.1 TRABALHOS FUTUROS

Os trabalhos futuros incluem o provável ajuste das seções 1, 2 e 3. Entretanto, o principal foco dos trabalhos futuros estará nas seções 4 e 5. Precisaremos formatar os dados a fim de servir como entrada válida a um modelo de RNN. Também, para melhorar a performance do modelo, possivelmente aprimoraremos ou mudaremos algumas das etapas de pré-processamento. Após treinar o modelo, apresentaremos os resultados obtidos e realizaremos a comparação do modelo proposto com os demais modelos da *Force 2020 Machine Learning Competition* usando a métrica de avaliação da própria competição. Por fim será feito a preparação da defesa pública, sua apresentação e ajustes finais. Cronograma completo na tabela 1 abaixo.

Tabela 1 – Planejamento das etapas do trabalho de conclusão de curso

Etapas	Meses					
	Maio	Junho	Julho	Agosto	Setembro	Outubro
Ajuste seções 1, 2 e 3	X					
Ajuste dos dados de entrada	X	X				
Experimentos		X	X			
Ajuste das seções 4, 5 e 6			X	X		
Entrega do rascunho de TCC				X	X	
Preparação da defesa pública					X	
Defesa pública					X	
Ajustes no relatório final do TCC					X	X

REFERÊNCIAS

2020, Force. **Force 2020 Machine Learning competition**. Disponível em: https://github.com/bolgebrygg/Force-2020-Machine-Learning-competition. Acessado em: 04/03/2021.

BOGGS, S. **Principles of Sedimentology and Stratigraphy**. [*S.l.*]: Prentice Hall, 2001. ISBN 9780130996961.

BRITANNICA. **Sedimentary Facies**. Disponível em: https://www.britannica.com/science/sedimentary-facies. Acessado em: 06/03/2021).

COLLEGE, Wenatchee Valley. **Depositional Environments**. Disponível em: https://commons.wvc.edu/rdawes/g101ocl/basics/depoenvirons.html#:~:text=A%5C%20depositional%5C%20environment%5C%20is%5C%20a,are%5C%20sometimes%5C%20called%5C%20sedimentary%5C%20environments. Acessado em: 06/03/2021).

DELL'AVERSANA, Paolo. Comparison of different Machine Learning algorithms for lithofacies classification from well logs. **Bollettino di Geofisica Teorica ed Applicata**, v. 60, n. 1, 2019.

DRAMSCH, Jesper Sören. Chapter One - 70 years of machine learning in geoscience in review. *In:* MOSELEY, Ben; KRISCHER, Lion (Ed.). **Machine Learning in Geosciences**. [*S.l.*]: Elsevier, 2020. v. 61. (Advances in Geophysics). P. 1–55. DOI: https://doi.org/10.1016/bs.agph.2020.08.002. Disponível em: http://www.sciencedirect.com/science/article/pii/S0065268720300054.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. [*S.l.*]: MIT Press, 2016. http://www.deeplearningbook.org.

IMAMVERDIYEV, Yadigar; SUKHOSTAT, Lyudmila. Lithological facies classification using deep convolutional neural network. **Journal of Petroleum Science and Engineering**, Elsevier, v. 174, p. 216–228, 2019.

INTERIOR BUREAU OF RECLAMATION, U.S. Department of the. **Engineering Geology Field Manual**. [*S.l.*: *s.n.*], 1998. https://www.usbr.gov/tsc/techreferences/mands/geologyfieldmanual.html.

JAIKLA, Chayawan *et al.* FaciesNet: Machine Learning Applications for Facies Classification in Well Logs. *In:* SECOND Workshop on Machine Learning and the Physical Sciences at the 33rd Conference on Neural Information Processing Systems (NeurIPS). [*S.l.*: *s.n.*], 2019. P. 10–12.

MANDAL, Partha Pratim; REZAEE, Reza. Facies classification with different machine learning algorithm–An efficient artificial intelligence technique for improved classification. **ASEG Extended Abstracts**, Taylor & Francis, v. 2019, n. 1, p. 1–6, 2019.

MITCHELL, Tom Michael. **Machine Learning**. 1. ed. [*S.l.*]: McGraw-Hill Education, 1997. ISBN 0070428077, 9780070428072.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of Machine Learning**. [*S.l.*]: MIT Press, 2012. (Adaptive Computation and Machine Learning series). ISBN 9780262018258.

MOORE, John A. Geographic variation of adaptive characters in Rana pipiens Schreber. **Evolution**, JSTOR, p. 1–24, 1949.

PETROWIKI. **Spectral gamma ray logs**. Disponível em: `https://petrowiki.spe.org/Spectral_gamma_ray_logs`. Acessado em: 21/03/2021.

RUSSELL, S.J.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 3. ed. Upper Saddle River, New Jersey 07458: Prentice Hall, 2010. ISBN 0136042597, 9780136042594.

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. **IBM Journal of Research and Development**, v. 3, n. 3, p. 210–229, 1959. DOI: `10.1147/rd.33.0210`.

SCHLUMBERGER. **The oil and gas industry's reference work**. Disponível em: `https://www.glossary.oilfield.slb.com/en/`. Acessado em: 21/03/2021.

WEI, Zhili *et al.* Characterizing rock facies using machine learning algorithm based on a convolutional neural network and data padding strategy. **Pure and Applied Geophysics**, Springer, v. 176, n. 8, p. 3593–3605, 2019.

YU, Y. *et al.* A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. **Neural Computation**, v. 31, n. 7, p. 1235–1270, 2019. DOI: `10.1162/neco_a_01199`.

YU, Yong *et al.* A review of recurrent neural networks: LSTM cells and network architectures. **Neural computation**, MIT Press, v. 31, n. 7, p. 1235–1270, 2019.