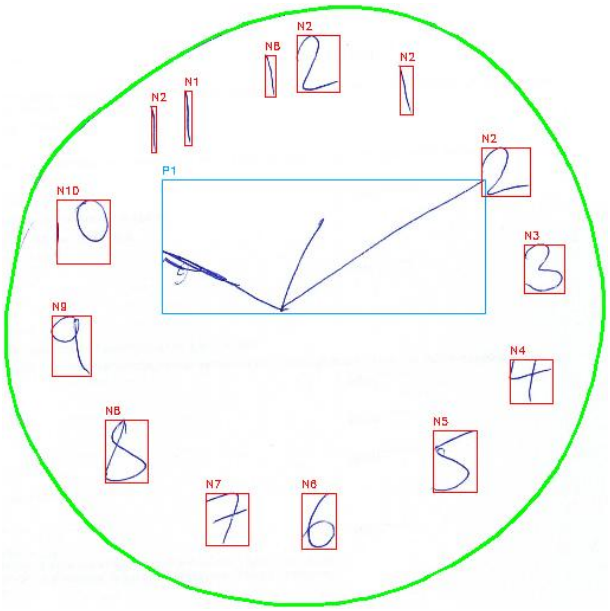
ISEL

ESCOLA SUPERIOR DE TECNOLOGIA DA SAÚDE DE LISBOA
INSTITUTO POLITÉCNICO DE LISBOA



Machine Learning no Teste do Desenho do Relógio – Segmentação e Classificação

RITA CARREIRA LOPES

(Licenciada em Engenharia Química e Biológica)

Dissertação para obtenção do grau de Mestre em Engenharia Biomédica

Orientadores:
Especialista Sérgio Rafael Reis Figueiredo
Doutor Pedro Miguel Torres Mendes Jorge

Júri:
Presidente: Doutor João Pedro Barrigana Ramos da Costa
Vogais:
Doutora Rita Gouveia Nunes
Especialista Sérgio Rafael Reis Figueiredo

Setembro 2025

Machine Learning no Teste do Desenho do Relógio – Segmentação e Classificação

RITA CARREIRA LOPES

(Licenciada em Engenharia Química e Biológica)

Dissertação para obtenção do grau de Mestre em Engenharia Biomédica

Orientadores:
Sérgio Rafael Reis Figueiredo (ESTeSL)
Pedro Miguel Torres Mendes Jorge (ISEL)

Júri:
Presidente: Doutor João Pedro Barrigana Ramos da Costa (ISEL)
Vogais:
Doutora Rita Gouveia Nunes (IST)
Especialista Sérgio Rafael Reis Figueiredo (ESTeSL)

Setembro 2025

Agradecimentos

Em primeiro lugar, gostaria de expressar o meu profundo agradecimento, ao Professor Dr. Sérgio Figueiredo, orientador desta dissertação, pela confiança depositada em mim e por me ter proporcionado a oportunidade de realizar este projeto. O seu acompanhamento e motivação ao longo de todo o processo foram fundamentais para a sua concretização.

Ao Professor Dr. Pedro Jorge, orientador desta dissertação, agradeço por toda a partilha de conhecimentos no decorrer de todas as fases deste percurso. Obrigada pela constante disponibilidade e todo o auxílio prestado, que me ajudaram a enfrentar todas as dificuldades.

Um agradecimento especial à Dra. Sofia Brissos, diretora da clínica Legismente – Psiquiatria e Psicologia Clínica e Forense, pela colaboração fundamental na disponibilização dos dados necessários para a realização deste estudo e pela confiança depositada na investigação desenvolvida. Adicionalmente, o meu sincero obrigada à Dra. Catarina Chester por toda a disponibilidade e esclarecimentos que também fizeram com que este trabalho fosse possível.

Aos meus colegas que me acompanharam ao longo de toda esta caminhada no ISEL, Inês Correia, Catarina Domingos, Raquel Figueiredo, Raul Alves, Filipa Luís, Rodrigo Ramos, Vanessa Ferrer e Bárbara Miranda, o meu sincero obrigada. Sem o vosso apoio chegar até aqui não teria sido possível e vou certamente levar para a vida todos os momentos e vivências que partilhamos ao longo dos últimos 6 anos. Em especial à Inês Correia, partilhar todas estas aventuras contigo e crescer ao teu lado, desde o jardim de infância até este momento, tem sido um verdadeiro privilégio.

Às minhas amigas de sempre, Maria, Raquel e Sara, obrigada por me ouvirem e estarem sempre presentes em todos os momentos importantes da minha vida. Tornam sempre todas as dificuldades mais leves e não consigo imaginar viver este percurso sem ser ao vosso lado.

Aos meus pais, obrigada por todo o apoio e motivação, tanto nesta como em todas as outras etapas da minha vida. Este trabalho também é vosso.

Por último, à minha irmã, obrigada por estares sempre disponível para me ouvir, sobretudo quando mais ninguém está presente. Espero, de alguma forma, conseguir inspirar-te e que acredites sempre no teu potencial tanto quanto eu.

Declaração de integridade

Declaro que esta(e) dissertação / trabalho de projeto / relatório de estágio é o resultado da minha investigação pessoal e independente. O seu conteúdo é original e todas as fontes listadas nas referências bibliográficas foram consultadas e estão devidamente mencionadas no texto. Mais declaro que todas as referências científicas e técnicas relevantes para o desenvolvimento do trabalho estão devidamente citadas e constam das referências bibliográficas.

O autor

Rita Carneira Lopes

Lisboa, 29 de Setembro de 2025

Machine Learning no Teste do Desenho do Relógio – Segmentação e Classificação

Resumo

O aumento da esperança média de vida tem vindo a agravar a prevalência de doenças neurodegenerativas, tornando crucial a adoção de métodos de rastreio objetivos e escaláveis. Esta dissertação propõe uma abordagem computacional para o Teste do Desenho do Relógio (TDR), combinando segmentação de imagem e modelos de Machine Learning (ML) para apoiar a classificação automática do desempenho cognitivo. Foi construída uma base de dados original com 117 TDR de participantes portugueses e desenvolvido um pipeline em *Python/OpenCV* para recorte, pré-processamento, deteção de contornos (Transformada de *Hough*) e segmentação dos componentes internos (números e ponteiros). O contorno principal foi identificado em todas as imagens e internamente, detetou-se pelo menos um ponteiro em 89,7% dos casos, dois ponteiros em 15,4% dos casos e um número igual ou superior a 12 dígitos em 84,6% dos casos, obtendo-se uma média de identificação de 15,3 dígitos por imagem, confirmando a viabilidade da segmentação apesar da elevada variabilidade gráfica.

Foram inicialmente extraídas 27 métricas e, após análise de correlação (com um limiar de 0,80), manteve-se um conjunto de 21. Posteriormente, definiu-se um subconjunto fixo de 9 métricas que cobre de forma equilibrada os elementos de contorno, números e ponteiros. Paralelamente, treinou-se uma *ResNet-18* no conjunto de dados *EMNIST-Digits* para a classificação de elementos candidatos a dígitos (*accuracy* de 99,65%), integrando esse classificador na *pipeline* de processamento.

Quatro modelos de ML (Regressão Logística, *Random Forest*, SVM e *Gradient Boosting*) foram avaliados recorrendo a duas estratégias de seleção de características (*Recursive Feature Elimination and Cross Validation* com *GridSearch* e conjunto fixo de *features* com *GridSearch*). Na avaliação final com 9 métricas, o *Gradient Boosting* apresentou o melhor desempenho, com uma *accuracy* de 83,33%, F1-score de 82,86% e AUC de 0,88, evidenciando boa capacidade discriminatória entre indivíduos com desempenho normal e patológico. Apesar das limitações (amostra reduzida e grafismos heterogéneos), os resultados reforçam a utilidade do método na padronização do TDR e no apoio ao rastreio precoce em contexto clínico nacional.

Palavras-chave: Teste do Desenho do Relógio; processamento de imagem; aprendizagem automática; seleção de características; classificação.

Machine Learning no Teste do Desenho do Relógio – Segmentação e Classificação

Abstract

As life expectancy increases, neurodegenerative diseases are becoming more prevalent, underscoring the need for objective and scalable screening tools. This dissertation proposes a computational approach to the Clock Drawing Test (CDT) that combines image segmentation with machine-learning (ML) models to support automatic performance classification. We assembled an original dataset of 117 Portuguese CDTs and built a Python/OpenCV pipeline for cropping, preprocessing, contour detection (Hough Transform), and segmentation of internal components (digits and hands). The main contour was detected in all images; internally, we identified ≥1 hand in 89.7% of cases, 2 hands in 15.4%, and ≥12 digits in 84.6%, with an average of 15.3 digits per image—evidence of feasible segmentation despite substantial graphic variability.

A total of 27 features were extracted and, after correlation analysis with a 0.80 threshold, retained 21. A second step defined a fixed 9-feature subset that balances contour, digits, and hands. In parallel, a ResNet-18 trained on EMNIST-Digits to label digit candidates achieved 99.65% accuracy and was integrated into the pipeline.

Four ML models—Logistic Regression, Random Forest, SVM, and Gradient Boosting—were evaluated under two feature-selection strategies (RFECV with GridSearch and a fixed-feature set with GridSearch). In the final evaluation using the 9 features, Gradient Boosting delivered the best results (accuracy 83.33%, F1-score 82.86%, AUC 0.88), showing good discrimination between Normal and Abnormal cases. Despite limitations (modest sample size and heterogeneous drawings), the findings indicate that the proposed method can help standardize CDT assessment and support early screening in national clinical settings.

Keywords: Clock Drawing Test, neurodegenerative diseases, image processing, machine learning, classification.

Lista de abreviaturas

Abn	Abnormal
AI	*Artificial Intelligence* / Inteligência Artificial
AMS	Atrofia de Múltiplos Sistemas
AUC	*Area Under the Curve*
CCL	Comprometimento Cognitivo Ligeiro
CE-ESTeSL	Comissão de Ética - Escola Superior de Tecnologia da Saúde de Lisboa
CNN	Convolutional Neural Network / Rede Neuronal Convolucional
CV	*Cross-Validation* / Validação Cruzada
DA	Doença de *Alzheimer*
DH	Doença de *Huntington*
DL	*Deep Learning* / Aprendizagem Profunda
DN	Doenças Neurodegenerativas
DP	Doença de *Parkinson*
ELA	Esclerose Lateral Amiotrófica
EMNIST	*Extended Modified National Institute of Standards and Technology*
FP16	*Floating Point 16 bits*
Grad-CAM	*Gradient-weighted Class Activation Mapping*
GPU	*Graphics Processing Unit*
KNN	*k-Nearest Neighbors*
LIME	*Local Interpretable Model-agnostic Explanations*
ML	*Machine Learning* / Aprendizagem Automática
MMSE	*Mini-Mental State Examination*
MoCA	*Montreal Cognitive Assessment*
N	*Normal*
RFECV	*Recursive Feature Elimination with Cross-Validation*
ROC	*Receiver Operating Characteristic*
SHAP	*SHapley Additive exPlanations*
SNC	*Sistema Nervoso Central*
SVM	*Support Vector Machine*
TDR	*Teste do Desenho do Relógio*
TPU	*Tensor Processing Unit*
XAI	*Explainable AI*
XGBoost	*Extreme Gradient Boosting*

Índice

Índice de Figuras

Índice de Tabelas

1. Introdução

1.1 Contextualização e Problemática

Estima-se que, em média, a cada três segundos seja diagnosticado um novo caso de demência em todo o mundo. A nível global, prevê-se que o número de indivíduos com demência duplique aproximadamente a cada 20 anos. De acordo com o *World Alzheimer Report* 2015, estima-se que a população mundial com idade igual ou superior a 60 anos tenha aumentado 51% até 2030 em comparação com o ano de 2015. O envelhecimento da população mundial é um fator crucial para o aumento do risco de desenvolvimento de demência [1].

Atualmente não existe cura para esta condição, no entanto o diagnóstico precoce e o tratamento ativo subsequente podem melhorar significativamente a qualidade de vida dos pacientes. Um dos testes mais utilizados em prática clínica para o rastreio e avaliação cognitiva, especialmente em casos de demência como a doença de Alzheimer e, em menor grau, a doença de Parkinson, é o Teste do Desenho do Relógio (TDR). O TDR é um teste não invasivo, breve e de baixo custo que consiste no simples desenho de um relógio analógico a marcar uma hora pré-determinada sem nenhum auxílio de memória. O aumento da popularidade deste teste levou ao aparecimento de uma grande variedade de sistemas de pontuação manual distintos bem-conceituados. No entanto, estes métodos tradicionais dependem frequentemente da interpretação subjetiva do profissional, uma vez que se baseiam em características qualitativas do desenho, associadas à elevada variabilidade dos grafismos. Por exemplo, pequenas variações na orientação dos ponteiros ou na posição dos algarismos podem indicar graus de declínio cognitivo diferentes e a subjetividade dos avaliadores pode levar a classificações incorretas. É necessária uma experiência clínica considerável para realizar este tipo de análise, o que torna o processo moroso pois está dependente de especialistas treinados e, consequentemente, limita a escalabilidade da aplicação do teste em grandes populações [2], [3].

Face a estas limitações, a aprendizagem automática ou *Machine Learning* (ML) surge como uma abordagem promissora para tarefas de classificação automática em contexto de diagnóstico, permitindo a segmentação da imagem e a extração automática das características relevantes para a avaliação do teste. Modelos computacionais baseados em ML têm a capacidade de identificar grafismos característicos de diferentes estágios de declínio cognitivo, minimizando a subjetividade da avaliação, padronizando os critérios da análise e, consequentemente, permitindo um aumento da precisão e fiabilidade do diagnóstico. Adicionalmente, a sua aplicação permite gerir e analisar grandes volumes de dados, possibilitando a realização de um maior número de

avaliações em menos tempo, e contribui para a deteção precoce de um maior número de casos [4].

1.2 Motivação

A relevância clínica e social das doenças neurodegenerativas e, em particular, da demência reforça a necessidade urgente de estabelecer estratégias de rastreio mais eficazes, que possibilitem não só um diagnóstico precoce, mas também uma avaliação objetiva e reprodutível em grande escala. O impacto destas doenças estende-se muito para além do indivíduo portador da doença, afetando também o quotidiano dos familiares e respetivos cuidadores assim como os sistemas de saúde, traduzindo-se, de forma direta e indireta, em elevados custos. Desta forma, qualquer abordagem que permita uma deteção precoce e precisa das alterações cognitivas associadas à demência constitui um benefício inequívoco para a saúde pública e para o bem-estar das populações.

Neste contexto, as ferramentas digitais e, em particular, a aplicação de técnicas de *Machine Learning*, surgem como uma oportunidade para melhorar significativamente a forma de avaliação do Teste do Desenho do Relógio. Ao contrário do que acontece com a análise manual, sujeita à inerente variabilidade dos avaliadores, os modelos computacionais permitem padronizar critérios, reduzir a subjetividade e aumentar a escalabilidade do teste. O potencial do ML em saúde digital tem sido amplamente evidenciado noutras áreas médicas, nomeadamente no apoio à decisão clínica através da análise de imagens médicas, e a sua aplicação ao TDR representa um passo natural na evolução dos métodos de rastreio do declínio cognitivo.

Apesar do crescente interesse nesta área a nível internacional, em Portugal verifica-se ainda uma escassez significativa neste domínio. Existem poucos estudos que explorem a utilização de técnicas de ML aplicadas à avaliação de TDR com dados de participantes portugueses, o que limita a adaptação e validação destas metodologias no contexto clínico e cultural da população nacional. Este trabalho surge, assim, motivado pela necessidade de colmatar esta lacuna, através da construção de uma base de dados própria constituída exclusivamente por participantes portugueses, da extração de características específicas dos grafismos do TDR e da avaliação do desempenho de diferentes modelos de classificação. Para além de contribuir para a literatura internacional, esta investigação pretende abrir caminho para a implementação de soluções tecnológicas adaptadas à realidade portuguesa, promovendo uma maior integração dos métodos computacionais nos processos de avaliação neuropsicológica, e tornar os rastreios de declínio cognitivo mais acessíveis, consistentes e escaláveis, com eventual impacto na prática clínica e no avanço da saúde digital no país.

1.3 Objetivos

O objetivo geral desta dissertação consiste em desenvolver um método automático baseado em ML capaz de segmentar e classificar as imagens provenientes do TDR, de forma a apoiar o rastreio e a avaliação em contexto ou suspeita de declínio cognitivo. Mais especificamente pretende-se:

- Desenvolver algoritmos de ML para processamento de imagem que tenham a capacidade de segmentar os componentes principais do TDR e identificar e extrair características relevantes dos seus grafismos.
- Treinar o modelo implementado para classificar as imagens com base na análise das características extraídas.
- Avaliar o desempenho do modelo desenvolvido em diferentes condições e cenários de aplicação.

1.4 Contribuições da dissertação

O presente trabalho apresenta quatro contributos principais para a investigação no domínio da avaliação automatizada do TDR. Em primeiro lugar, a construção de uma base de dados original composta apenas por indivíduos portugueses, que faz face à ausência de conjuntos de dados representativos da população nacional neste contexto. A disponibilização deste recurso constitui uma mais-valia científica pois permite a validação de modelos adaptados à realidade portuguesa e possibilita a sua utilização em trabalhos futuros.

Em segundo lugar, foi realizada a extração de um conjunto abrangente de características a partir dos grafismos dos desenhos, com vista à tradução da complexidade visual do TDR em métricas quantitativas. Esta etapa ajuda a estabelecer a ligação entre o processamento de imagem e a interpretação clínica, assegurando que os algoritmos de ML operam sobre atributos objetivos e clinicamente significativos.

O terceiro contributo corresponde à implementação e avaliação de quatro modelos de ML aplicados à classificação dos desenhos. Esta abordagem possibilitou a análise da aplicabilidade de diversas técnicas ao TDR, evidenciando o seu potencial enquanto ferramenta de apoio ao rastreio do declínio cognitivo. A comparação dos desempenhos obtidos permitiu identificar vantagens e limitações inerentes a cada modelo.

Por último, foi conduzido um processo de seleção de características com o objetivo de reduzir a dimensionalidade do conjunto de dados e identificar os preditores mais relevantes. Para além de melhorar a eficiência dos modelos, esta etapa aumentou a interpretabilidade dos resultados, destacando as características mais preponderantes na diferenciação entre indivíduos com desempenho normal e patológico.

Em conjunto, estes aspetos reforçam a relevância científica e prática desta dissertação, contribuindo para a avaliação automatizada do TDR, particularmente no contexto da saúde digital em Portugal.

1.5 Organização da dissertação

A presente dissertação encontra-se estruturada em cinco capítulos que contemplam os fundamentos teóricos, a metodologia adotada, os resultados obtidos e as respetivas conclusões retiradas.

No capítulo 1 é apresentada uma breve introdução ao trabalho, onde o leitor é contextualizado sobre os temas a abordar e a problemática principal, sendo também apresentadas as motivações que levaram à sua elaboração, bem como os objetivos principais e as possíveis contribuições relevantes.

No capítulo 2 é apresentada a revisão bibliográfica relacionada com os principais tópicos desta dissertação, de forma que o leitor tenha conhecimento de todos os conceitos necessários à compreensão do trabalho desenvolvido.

No capítulo 3 é descrita toda a metodologia levada a cabo para desenvolver um método em ML capaz de segmentar e classificar imagens de TDR.

No capítulo 4 são apresentados os resultados obtidos após o treino dos modelos e a avaliação do seu desempenho, paralelamente à respetiva discussão.

No capítulo 5 são apresentadas as conclusões finais em resposta aos objetivos inicialmente propostos, as limitações encontradas durante a realização deste trabalho e as perspetivas de trabalhos futuros.

2. Estado da Arte

2.1 Doenças Neurodegenerativas e Importância do Diagnóstico Precoce

As doenças neurodegenerativas (DN) são caracterizadas pela perda lenta e progressiva de neurónios no sistema nervoso central (SNC), conduzindo a défices em determinadas funções cognitivas e motoras, tais como alterações da memória, do movimento e a linguagem, podendo culminar em morte. Os sinais e sintomas característicos, que dependem da área do cérebro onde ocorre a perda neuronal, são o fator distintivo das diferentes patologias deste tipo. Algumas da principais DNs são as doenças de Alzheimer (DA), Parkinson (DP), Esclerose Lateral Amiotrófica (ELA), Huntington (DH) e Atrofia de Múltiplos Sistemas (AMS) [5], [6]. A DA é a forma mais comum de demência, tendo uma contribuição entre 60% a 70% dos casos totais. Em 2021, foram registadas cerca de 57 milhões de pessoas com demência a nível mundial, sendo registados anualmente cerca de10 milhões de novos diagnósticos [7].

O envelhecimento é considerado um dos fatores de risco mais significativos para a demência e, com a população global a envelhecer rapidamente e a esperança média de vida aumentar, é expectável que até 2030 o número de pessoas com demência atinga os 80 milhões, representando um dos maiores desafios de saúde pública [8]. Apesar do impacto significativo destas patologias, não existe ainda cura. Contudo, é amplamente reconhecido que a deteção precoce possibilita a implementação de estratégias terapêuticas e de suporte capazes de retardar a progressão clínica, atenuar sintomas e melhorar a qualidade de vida tanto dos doentes como dos cuidadores. Para além do benefício individual, o diagnóstico precoce tem um contributo importante a nível socioeconómico, permitindo otimizar a gestão de recursos de saúde e planear respostas adequadas face ao aumento projetado do número de casos [9], [10], [11].

Neste contexto, torna-se essencial identificar ferramentas de rastreio acessíveis, fiáveis e escaláveis. Entre estas, destaca-se o Teste do Desenho do Relógio (TDR), amplamente utilizado na avaliação cognitiva em contexto clínico.

2.2 Teste do Desenho do Relógio na Avaliação Cognitiva

Uma ferramenta ideal de rastreio cognitivo deve ser de administração rápida, aceitável para os pacientes, de fácil classificação, relativamente independente de fatores culturais, linguísticos e educacionais, apresentar boa fiabilidade intra e interavaliador, elevados níveis de sensibilidade e especificidade, validade preditiva e correlação com métricas de severidade e de avaliação de demência [12].

O TDR destaca-se por satisfazer grande parte destes critérios, constituindo uma das ferramentas mais amplamente utilizadas para rastreio cognitivo. A sua aplicação consiste na solicitação, a um indivíduo, de que desenhe um relógio analógico indicando uma hora pré-determinada, geralmente “11 horas e 10 minutos”, sem recurso a ajudas externas (Figura 1) Trata-se de um teste breve, de baixo custo, não invasivo e de fácil compreensão, o que o torna particularmente adequado para aplicação em contextos clínicos, hospitalares e comunitários [13]. Para além da simplicidade da tarefa, importa salientar que o TDR não se limita a avaliar uma única função cognitiva, exigindo o envolvimento simultâneo de diversas regiões corticais, nomeadamente os lobos frontal, parietal e temporal. Desta forma, o desempenho no TDR reflete uma integração complexa de múltiplos domínios cognitivos, incluindo compreensão, planeamento e organização, memória visual e reconstrutiva, habilidades visuoespaciais, programação e execução motora, conhecimento numérico, raciocínio abstrato, capacidade de inibição de respostas automáticas, atenção sustentada e até aspetos emocionais como a tolerância à frustração [13] [14].

Figura 1 – Exemplo de um Teste do Desenho do Relógio [15].

O TDR pode ser administrado em duas modalidades complementares — a de comando, onde o participante desenha um relógio a marcar uma hora específica, e a de cópia, na qual o participante reproduz um relógio previamente apresentado. A primeira enfatiza processos executivos, planeamento e organização sequencial, enquanto a segunda explora capacidades visuoconstrutivas e a atenção visuoespacial [16], [17]. Em contexto clínico, o TDR é utilizado como instrumento de rastreio global e como complemento na caracterização de défices em múltiplas condições neurológicas (demência, sequelas de acidente vascular cerebral, DP), dada a sua rapidez e excelente aceitabilidade pelos doentes [17].

A utilização do TDR remonta a mais de um século, com os primeiros registos a surgirem em 1915, embora tenha sido popularizado apenas nas décadas que se seguiram. Inicialmente, o teste foi concebido como um instrumento para avaliar distúrbios relacionados com afasia e apraxia construtiva, focando-se na capacidade de reprodução de desenhos e na organização espacial. Apesar de menções frequentes ao trabalho de MacDonald Critchley em 1953, o uso clínico sistemático do TDR só se consolidou na década de 80, quando passou a ser adotado como ferramenta de rastreio cognitivo para a demência e outras condições neurológicas [14]. O primeiro estudo publicado que fez a associação do TDR à triagem de pacientes idosos com distúrbios cognitivos, mais especificamente o rastreio e acompanhamento de demência aguda e delírio, surgiu em 1986 por Shulman et al.[18] Desde então, o interesse pelo teste tem vindo a aumentar exponencialmente, particularmente no âmbito da DA, com milhares de publicações a documentar a sua aplicação, eficácia e adaptação em diferentes contextos clínicos, confirmando o TDR como um instrumento robusto e de rápida administração na avaliação da função cognitiva global [14].

A consolidação do TDR como instrumento de rastreio originou, contudo, a necessidade de sistematizar os critérios de aplicação e interpretação. Diversos sistemas de pontuação foram desenvolvidos ao longo do tempo, variando entre metodologias qualitativas, quantitativas e híbridas. Entre os mais conhecidos encontram-se os sistemas criados por Shulman, Sunderland, Rouleau e Freedman. Cada um destes métodos procurou responder a diferentes necessidades clínicas, desde a deteção de erros subtis em fases iniciais de comprometimento cognitivo, até à avaliação global do desempenho em tarefas complexas. Esta diversidade reflete o esforço da comunidade científica em tornar a aplicação do TDR mais consistente, reduzindo variações interpretativas e aumentando o rigor na avaliação [13].

No entanto, importa reconhecer que, apesar da sua elevada utilidade clínica, o TDR não está isento de limitações. Estudos demonstram que o teste apresenta elevada sensibilidade para alterações sobretudo em funções executivas (planeamento, organização, flexibilidade cognitiva, inibição de respostas inadequadas) e visuoespaciais, refletindo o envolvimento conjunto de regiões corticais frontais e parietais. Estas características tornam o teste particularmente sensível a alterações associadas a patologias como a DA e a demência vascular. Contudo, a sua capacidade de detetar défices predominantemente mnésicos é limitada, uma vez que não avalia diretamente a memória episódica ou semântica, podendo assim falhar em identificar quadros iniciais de declínio cognitivo centrados na memória [19].

Adicionalmente, diversos estudos sublinham que o TDR apresenta baixa especificidade quando usado de forma isolada, sobretudo na distinção entre o envelhecimento normal, o comprometimento cognitivo ligeiro (CCL) e a demência. Uma revisão sistemática,

concluiu que, apesar de útil como ferramenta de rastreio, o TDR não deve ser considerado um marcador diagnóstico autónomo para CCL, dada a sua variabilidade de desempenho e vulnerabilidade a fatores educacionais e culturais.[20] Em particular, o nível de literacia pode influenciar significativamente a execução da tarefa, afetando assim a interpretação dos resultados. Para mitigar estas limitações, é recomendado que o TDR seja aplicado em conjunto com outros testes cognitivos, como por exemplo o *Mini-Mental State Examination* (MMSE) ou o *Montreal Cognitive Assessment* (MoCA). A literatura mostra que esta abordagem combinada aumenta a precisão do diagnóstico, melhora a sensibilidade para défices subtis em fases iniciais e reduz o risco de falsos negativos, proporcionando uma avaliação mais abrangente das diferentes dimensões da cognição [19].

Outro ponto crítico associado ao TDR tradicional reside na subjetividade inerente ao processo de avaliação manual. A interpretação depende fortemente da experiência clínica do avaliador, o que pode introduzir inconsistências inter e intraavaliador. Estudos clássicos evidenciam que a concordância entre diferentes profissionais é, muitas vezes, apenas moderada, refletindo divergências na interpretação de erros qualitativos e na aplicação de critérios de avaliação. Esta variabilidade compromete a fiabilidade dos resultados em contextos clínicos e limita deteção de alterações precoces [21], [22].

Não obstante as suas vantagens (brevidade, baixo custo, fácil compreensão), o TDR é influenciado por fatores como escolaridade, proficiência linguística, alterações motoras e percetivas e familiaridade com relógios analógicos. Por isso, recomenda-se a sua utilização integrada com outras medidas e a interpretação contextualizada ao perfil do doente [12], [23].

Neste enquadramento, a modernização do TDR através da digitalização e do recurso a técnicas computacionais apresenta-se como uma alternativa promissora, visando aumentar a objetividade, reduzir a variabilidade interpretativa e melhorar a precisão da avaliação. Este tema será desenvolvido no subcapítulo seguinte.

2.3 Digitalização e Processamento de Imagem no TDR

As limitações associadas à avaliação manual do TDR incentivaram o desenvolvimento de versões digitais, que permitem registar os traços com maior precisão e objetividade, utilizando dispositivos como *tablets* e *smartpens*. A digitalização permite capturar informações temporais e espaciais detalhadas, como por exemplo a velocidade de execução, pressão do traço e sequência de desenho, enriquecendo a análise quantitativa e qualitativa do desempenho do paciente [3], [24].

A disponibilidade destes dados adicionais potencia a aplicação de algoritmos de ML para a classificação de diferentes estados cognitivos, com desempenhos superiores aos

métodos tradicionais. Binaco *et al*. mostraram que modelos baseados em características extraídas dos TDR digitalizados foram capazes de diferenciar com elevada precisão subtipos de CCL e DA [24]. De forma semelhante, Souillard-Mandar *et al*. desenvolveram modelos interpretáveis de ML a partir de dados recolhidos com canetas digitalizadoras que registam a posição com alta precisão espacial e temporal. Estes modelos demonstraram ser mais precisos que os sistemas de pontuação tradicionais, sem comprometer a interpretabilidade clínica [25]. Paralelamente, Lazarova *et al*. utilizaram modelos de regressão logística aplicados a erros específicos do desenho do relógio, complementados com dados clínicos adicionais, alcançando bons resultados na deteção de DA e evidenciando o potencial da combinação entre informação gráfica e dados contextuais [26], [27].

Na literatura, observa-se também que diferentes estudos variam substancialmente na forma como operacionalizam as características extraídas do TDR. Binaco et al., por exemplo, trabalharam com um conjunto reduzido de participantes (≈50 indivíduos), centrando a análise em erros gráficos discretos — como omissões de números e perturbações da lógica do mostrador — usados como descritores primários [24]. Em contraste, os trabalhos de Lazarova e Grigorova incluíram mais de uma centena de observações e basearam-se principalmente em medidas contínuas de organização espacial, como desvios angulares e relações geométricas entre algarismos [26], [27]. Já Souillard-Mandar et al. recorreram a uma base de dados substancialmente maior, recolhida com *smartpen*, da qual foram derivadas centenas de métricas cinemáticas e espaciais que captam subtilezas do processo motor durante o desenho e não apenas o produto final [3], [25]. Estas diferenças ilustram a diversidade metodológica existente e a necessidade de contextualizar resultados segundo a natureza das amostras e dos atributos utilizados.

O desenvolvimento de bases de dados como o *Toronto Digital Clock Drawing Test Dataset* e o *USC Clock Drawing Test Dataset* tem sido determinante para treinar algoritmos de ML em larga escala, permitindo a avaliação automática de grandes volumes de dados e a identificação de défices subtis muitas vezes impercetíveis na análise manual. Deste modo, a transição para versões digitais não só aumenta a consistência entre avaliadores, como também abre caminho a aplicações de rastreio cognitivo populacional e investigação clínica de larga escala [24], [28].

Do ponto de vista técnico, a digitalização do TDR permite aplicar técnicas clássicas de processamento de imagem, capazes de transformar os desenhos em métricas objetivas. O *pipeline* comum inclui etapas de pré-processamento (conversão para escala de cinza, suavização e binarização), segmentação (separação dos elementos do desenho - círculo do relógio, ponteiros e números) e extração de contornos e características [3], [29]. Na fase de binarização, por exemplo, métodos como *Otsu thresholding* são

frequentemente adotados para determinar automaticamente um limiar que separa o fundo do traço, com base na distribuição de intensidades. Para extração de bordas e contornos, o detetor *Canny* é uma escolha clássica, reconhecido pela sua precisão em imagens com ruído moderado. Em estudos de reconhecimento de dígitos manuscritos no contexto do TDR digitalizado, assume-se implicitamente essa etapa de deteção de contornos antes da classificação de números [30], [31]. A deteção de formas geométricas como o contorno circular do relógio pode recorrer à Transformada de *Hough* para círculos (*Circle Hough Transform*). Neste método, cada ponto de borda identificado na imagem é associado a um conjunto de possíveis centros e raios compatíveis com a sua posição. Estes parâmetros são registados num espaço acumulador tridimensional, no qual as combinações mais consistentes entre diferentes pontos de borda se destacam como picos. Esses picos representam as configurações paramétricas mais prováveis de círculos presentes na imagem, permitindo identificar o mostrador do relógio mesmo em situações de ruído ou contornos incompletos [31], [32]. Em casos de imagens com ruído ou contornos imperfeitos, abordagens mais recentes propõem algoritmos adaptativos de deteção de círculo, capazes de lidar com variações e imperfeições nos traços do relógio [32].

Entre os atributos extraídos com maior frequência destacam-se a simetria do círculo, distribuição espacial dos números, ângulos e proporções entre os ponteiros, centralidade e densidade do traço. Um estudo recente de Davoudi *et al.* evidenciou que o recurso a canetas digitais em testes cognitivos permite capturar, de forma contínua e ponto a ponto, características cinemáticas e espaciais detalhadas do traço, como tempo de execução, velocidade, aceleração, pausas e irregularidade. Estas métricas, muitas vezes impercetíveis na avaliação manual, revelaram-se altamente discriminativas entre indivíduos saudáveis e com défice cognitivo. Os autores mostraram ainda que a incorporação destas variáveis em modelos de *Machine Learning* aumenta significativamente a capacidade de predição de comprometimento cognitivo ligeiro, sublinhando o valor da digitalização para detetar alterações subtis em estádios iniciais da doença. Esta abordagem reforça, assim, o potencial do TDR digital em produzir dados ricos e multifacetados que podem ser explorados por algoritmos de ML para apoiar o diagnóstico precoce [3], [33].

Em síntese, a digitalização do TDR cria condições para uma análise mais objetiva e sistemática, não apenas replicando critérios de pontuação tradicionais, mas também fornecendo dados que alimentam modelos de *Machine Learning* e *Deep Learning,* capazes de automatizar a avaliação do comprometimento cognitivo [2], [28]. A partir deste ponto, a aplicação de algoritmos de ML assume um papel central na exploração dos atributos extraídos das imagens para apoiar o diagnóstico e aumentar a precisão do rastreio cognitivo.

2.4 *Machine Learning* no Teste do Desenho do Relógio

O tipo de atributos selecionados para alimentar modelos de *Machine Learning* varia amplamente entre estudos. Masuo et al. utilizaram erros qualitativos — como omissões, desalinhamentos e posicionamento incorreto de ponteiros — codificados manualmente em variáveis discretas [34]. Em contraste, Davoudi et al. recorreram a métricas cinemáticas registadas com dispositivos digitais, incluindo velocidade do traço, flutuações de ritmo, pausas e sequências de execução [29], [30]. Já Lazarova e Grigorova adotaram uma abordagem estritamente geométrica, analisando a coerência angular dos números e a sua distribuição espacial [26], [27]. Esta variabilidade demonstra a inexistência de consenso sobre um conjunto ótimo de características, refletindo diferentes objetivos clínicos e tecnológicos presentes na literatura.

A crescente digitalização de testes cognitivos, como o TDR, gerou grandes volumes de dados multidimensionais, que vão além da avaliação visual tradicional. Neste contexto, o *Machine Learning* surge como uma abordagem particularmente promissora, capaz de processar esses dados de forma automática, identificar padrões complexos e apoiar o diagnóstico precoce de défices cognitivos. O ML distingue-se por permitir a extração de relações não lineares entre variáveis, frequentemente impercetíveis à análise humana, e pela sua capacidade de generalização, tornando-o uma ferramenta valiosa para a avaliação neuropsicológica em larga escala [2], [28].

Vários estudos já demonstraram o potencial do ML na análise do TDR. Masuo *et al.*, por exemplo, testaram a classificação baseada em erros qualitativos do desenho - como erros concetuais, omissões, deslocamento espacial dos números e posicionamento incorreto dos ponteiros – e mostraram que o modelo *Support Vector Machines* (SVM) conseguiu distinguir grupos de idosos com diferentes níveis de comprometimento cognitivo, atingindo uma *accuracy* de aproximadamente 79%. Apesar de promissor, este trabalho, tal como muitos na área, apresenta limitações importantes, nomeadamente o tamanho reduzido das amostras e a ausência de validação externa, o que restringe a generalização dos resultados para outras populações [34].

Do ponto de vista metodológico, o SVM tem a vantagem de lidar bem com espaços de alta dimensionalidade e de capturar relações complexas não lineares através de funções *kernel*, mas pode ser computacionalmente exigente e sensível à escolha dos parâmetros [35]. Já os modelos baseados em árvores de decisão, como o *Random Forest*, oferecem maior robustez a variáveis com mais ruído e interpretabilidade relativa através da análise de importância das características, mas tendem a necessitar de amostras maiores para atingir o seu desempenho máximo. A Regressão Logística, por sua vez, destaca-se pela simplicidade e interpretabilidade, sendo útil em contextos clínicos, mas assume relações lineares entre variáveis e, por isso, pode falhar em

capturar padrões mais complexos [36]. Assim, a escolha do algoritmo deve equilibrar a complexidade dos dados, a interpretabilidade clínica e a disponibilidade de amostras suficientemente grandes, destacando a necessidade de estudos com bases de dados mais extensas e validação multicêntrica para consolidar a utilização do ML no TDR [35], [36].

Na mesma linha, Souillard-Mandar *et al*. aplicaram o algoritmo *Random Forest* à análise de TDR digitalizados, explorando múltiplos atributos extraídos automaticamente dos desenhos, como a posição e distribuição dos números, a proporção e orientação dos ponteiros e métricas de simetria. Este modelo revelou-se eficaz na captura de interações complexas entre variáveis e na modelação de relações não lineares, características particularmente relevantes quando se analisam grafismos altamente heterogéneos. Uma das suas grandes vantagens é a robustez face a variáveis ruidosas e a capacidade de fornecer estimativas da importância relativa das características, o que contribui para a interpretabilidade clínica. Contudo, tal como outros modelos baseados em árvores, o *Random Forest* tende a necessitar de amostras maiores para atingir estabilidade e pode perder desempenho quando aplicado a *datasets* limitados, como é frequentemente o caso nos estudos com TDR. Além disso, a ausência de validação externa em diferentes populações limita a generalização dos resultados obtidos, reforçando a necessidade de estudos multicêntricos de maior escala [3].

Por sua vez, Lazarova e Grigorova recorreram à Regressão Logística para identificar a DA com base em erros específicos observados no TDR, como omissões de números, agrupamentos incorretos e alterações espaciais na disposição do círculo. Este modelo alcançou um valor AUC de 0,825, evidenciando que, apesar da sua simplicidade, pode produzir resultados sólidos quando as características são cuidadosamente selecionadas. A principal vantagem da Regressão Logística é a sua elevada interpretabilidade, o que facilita a integração em contextos clínicos. No entanto, este modelo assume relações lineares entre variáveis, podendo falhar em capturar padrões mais complexos presentes nos grafismos do TDR. Adicionalmente, estudos baseados nesta abordagem enfrentam os mesmos constrangimentos de pequenas amostras e de ausência de replicação em contextos externos, o que limita a confiança na sua aplicabilidade clínica generalizada [27].

Mais recentemente, Chen *et al*. desenvolveram um modelo inteligente de rastreio de impedimento cognitivo ligeiro através da combinação de dados demográficos, medidas de *eye-tracking* e métricas do TDR digitalizado, utilizando múltiplos algoritmos de ML (Regressão Logística, SVM, *Random Forest, Extreme Gradient Boostin*g (*XGBoost*), *Multilayer Perceptron* e Redes Convolucionais). Os resultados obtidos demonstraram que os modelos multimodais superaram consistentemente os unimodais, com o melhor desempenho a ser alcançado pelo modelo *Random Forest*, que atingiu um valor de AUC

de 0,947. Os modelos de *XGBoost* e SVM também apresentaram uma performance elevada, evidenciando o potencial dos métodos clássicos de ML aliados ao TDR digitalizado para aumentar a precisão da deteção precoce de défices cognitivos [37].

Em conjunto, estas investigações evidenciam que os métodos clássicos de ML aplicados ao TDR oferecem resultados promissores na modelação das relações entre atributos extraídos dos desenhos e indicadores clínicos. Além disso, constituem um marco importante na consolidação de sistemas automatizados de pontuação e diagnóstico precoce, demonstrando a viabilidade de soluções objetivas, escaláveis e replicáveis em diferentes contextos clínicos [3], [27], [37].

Apesar destes avanços, observa-se que a maioria dos estudos recorre a um conjunto relativamente restrito de algoritmos (SVM, *Random Forest* e Regressão Logística). Outros métodos, como o *k-Nearest Neighbors* (KNN), poderiam constituir alternativas interessantes em *datasets* de pequena dimensão pela sua simplicidade e intuição geométrica, embora apresentem menor escalabilidade em grandes volumes de dados. Mais recentemente, têm sido exploradas abordagens baseadas em *ensembles* híbridos, que combinam modelos distintos (e.g., *Random Forest* e SVM ou *Random Forest* e Regressão Logística), tirando partido das vantagens de cada técnica e aumentando a robustez dos resultados [38], [39]. Outro ponto crítico é a interpretabilidade dos modelos, requisito essencial em saúde digital. Ferramentas como o SHAP (*SHapley Additive exPlanations*) e o LIME (*Local Interpretable Model-agnostic Explanations*) permitem compreender a contribuição de cada variável para a decisão final, facilitando a aceitação clínica dos algoritmos. Assim, o futuro da aplicação do ML ao TDR passa, não apenas por diversificar os algoritmos utilizados, mas também por assegurar que as soluções desenvolvidas são transparentes e clinicamente interpretáveis [40].

2.5 *Deep Learning* e Redes Neuronais no TDR

O *Deep Learning* representa uma evolução significativa relativamente ao ML clássico, caracterizando-se pela capacidade de representações hierárquicas a partir dos dados, dispensando a necessidade de engenharia manual de características. Enquanto os modelos tradicionais dependem fortemente da seleção prévia de atributos, as redes neuronais profundas extraem automaticamente padrões complexos e multidimensionais, o que as torna especialmente adequadas para tarefas de reconhecimento visual e análise de imagem [41], [42].

No contexto do TDR, várias arquiteturas de Redes Neuronais Convolucionais (CNN) têm demonstrado um progresso significativo na avaliação automatizada. Estas abordagens oferecem vantagens substanciais em relação aos métodos tradicionais de ML clássico já discutidos, permitindo a identificação de padrões gráficos complexos que

dificilmente são captados por algoritmos tradicionais. Diversos estudos evidenciam a eficácia das CNNs no rastreio e na avaliação da severidade de défices cognitivos [2], [28].

Chen *et al*. aplicaram arquiteturas como *VGG16, ResNet-152* e *DenseNet-121* em imagens digitalizadas do TDR, alcançando uma precisão de 96,65% na triagem e até 98,54% na pontuação da severidade, superando as taxas de erro humano e estabelecendo um novo padrão para a avaliação automatizada do TDR [2]. Da mesma forma, Sato *et al*. treinaram uma CNN em mais de 40.000 imagens de TDR, obtendo uma precisão de 90,1% na identificação de declínio funcional executivo e 77,2% na deteção de demência provável, evidenciando o potencial das CNNs para rastreios em massa [28]. Para além destes avanços, Chen et al. integraram dados gráficos e biométricos obtidos através de *eye-tracking*, incluindo padrões de fixação, duração das explorações visuais e sequência do olhar durante a execução do teste [37]. Esta abordagem multimodal permitiu captar não apenas o produto final do desenho, mas também processos cognitivos subjacentes, enriquecendo substancialmente a informação fornecida aos modelos de classificação.

Mais recentemente, Park e Lee demonstraram as vantagens das CNNs na análise do TDR, através do desenvolvimento de uma abordagem móvel automatizada que combinou CNNs, *U-Net* e dados de sensores móveis para pontuar os desenhos do relógio. O modelo alcançou uma *accuracy* superior a 89% na avaliação de características como o contorno do círculo, posicionamento dos números, orientação dos ponteiros e centralidade, evidenciando a capacidade das CNNs para capturar padrões complexos que escapam à análise visual humana. Esta abordagem também permite a pontuação qualitativa automatizada, superando limitações existentes nos métodos tradicionais que já foram referidas anteriormente e facilitando a implementação do TDR em contextos clínicos de larga escala [43].

Apesar destes avanços, a aplicação do *DL* no âmbito da avaliação do TDR enfrenta desafios significativos. A elevada complexidade das redes neuronais profundas implica a necessidade de grandes volumes de dados anotados para evitar problemas de *overfitting*, nos quais o modelo memoriza os exemplos de treino, mas falha na generalização para novos casos [41], [44]. Tal exigência representa um obstáculo, sobretudo em saúde, onde a recolha e anotação de dados é morosa e requer avaliadores especializados. Para além disso, os modelos de DL são computacionalmente exigentes, necessitando de hardware especializado, como GPUs ou TPUs, bem como infraestruturas robustas de armazenamento e processamento, o que pode limitar a sua adoção em centros clínicos com menos recursos [42].

Outro desafio central prende-se com a interpretabilidade. Redes profundas são frequentemente caracterizadas como *"black boxes"*, dado que as suas decisões resultam de combinações complexas de pesos e ativações distribuídas por múltiplas camadas, dificultando a explicação do racional clínico subjacente a uma classificação. Esta opacidade compromete a aceitação em contexto médico, onde a transparência é um requisito essencial. Para mitigar esta limitação, têm sido desenvolvidas técnicas de *Explainable AI* (XAI), como o *Grad-CAM* (*Gradient-weighted Class Activation Mapping*), que permite visualizar as regiões do desenho mais relevantes para a decisão do modelo, ou mecanismos de atenção visual que destacam as áreas do *input* com maior impacto na saída. Estas ferramentas aumentam a confiança dos clínicos nos sistemas de DL e favorecem a sua aplicabilidade em ambiente hospitalar [45].

Para superar estes constrangimentos, vários trabalhos sugerem a adoção de estratégias híbridas, que combinam o DL com métodos clássicos de ML ou com abordagens robustas de pré-processamento e segmentação, visando melhorar a generalização, consistência e aplicabilidade clínica do TDR digitalizado. Por exemplo, a utilização de CNNs em conjunto com técnicas de segmentação permite analisar separadamente os diferentes componentes do TDR (círculo, números e ponteiros), aumentando a *accuracy* da avaliação. Da mesma forma, a implementação de técnicas de regularização, *data augmentation* e validação cruzada tem-se mostrado eficaz na redução do risco de *overfitting*, melhorando a generalização e a robustez dos modelos em amostras mais pequenas [2], [44].

Em suma, o DL apresenta claras vantagens face ao ML clássico na análise do TDR, destacando-se pela sua capacidade de extrair automaticamente características complexas e pelo desempenho superior demonstrado em diversos estudos. Contudo, a sua implementação prática depende de enfrentar desafios importantes, nomeadamente a disponibilidade de *datasets* suficientemente grandes, os requisitos computacionais e a necessidade de interpretabilidade. O futuro passa por explorar soluções híbridas e integrativas, conciliando o poder do DL com estratégias de explicabilidade e validação multicêntrica, de forma a favorecer a sua aceitação clínica e aplicação em larga escala [2], [41], [42], [45].

2.6 Desafios Atuais e Oportunidades de Investigação

Apesar dos avanços na aplicação de ML e DL na avaliação do TDR, persistem diversas lacunas na literatura, sobretudo no que diz respeito à sua aplicação em contextos clínicos reais e à generalização dos modelos desenvolvidos. Grande parte dos estudos disponíveis baseia-se em bases de dados de dimensão reduzida, frequentemente recolhidas em contextos únicos, o que limita a capacidade de generalização dos algoritmos para diferentes populações. A escassez de dados anotados de alta qualidade

constitui uma das principais barreiras ao avanço da investigação nesta área, comprometendo a fiabilidade e a reprodutibilidade dos modelos [46].

Além disso, persistem dificuldades metodológicas relacionadas com a padronização dos *pipelines* de processamento e com a ausência de validação externa multicêntrica, fatores que fragilizam a robustez dos resultados. Nos modelos de DL, acrescem os problemas do *overfitting*, os elevados requisitos computacionais e a baixa interpretabilidade, aspetos que dificultam a sua adoção em larga escala em ambiente clínico [41], [42], [45].

Outro desafio relevante prende-se com a heterogeneidade dos métodos utilizados. Diferentes estudos aplicam técnicas distintas de pré-processamento, segmentação e análise, o que dificulta a comparação de resultados e a definição de protocolos estandardizados. A anotação manual dos desenhos continua a ser demorada e sujeita a variações interavaliador, dificultando a criação de grandes *datasets* de referência consistentes [3], [24], [46].

Ainda que os resultados internacionais sejam promissores, no contexto português a investigação permanece incipiente. Até ao momento, não existem estudos publicados que explorem de forma sistemática a aplicação de ML ou DL ao TDR em populações nacionais [2], [3], [28]. A literatura limita-se maioritariamente a contributos normativos e de validação com base em sistemas de pontuação tradicionais, como os de Shulman, Rouleau e Babins, incluindo adaptações e aplicações em Portugal. Um dos contributos relevantes é o de Santana *et al.*, que apresentou dados normativos para a população portuguesa, com base em três sistemas de pontuação distintos. Embora fundamentais para estabelecer pontos de corte e enquadrar efeitos demográficos, estes dados permanecem circunscritos à avaliação manual e não abrangem a realidade digital [12], [47].

A literatura portuguesa demonstra que variáveis demográficas influenciam de forma significativa o desempenho no TDR. Santana et al. evidenciam diferenças marcantes entre níveis de escolaridade e faixas etárias, tanto na organização espacial como na precisão gráfica do desenho [47]. Assim, para que modelos de ML desenvolvidos em bases de dados nacionais apresentem validade externa, torna-se essencial avaliar se a amostra utilizada reflete adequadamente a distribuição demográfica da população portuguesa.

Investigações adicionais têm explorado a validade e fiabilidade de diferentes métodos de pontuação em amostras nacionais, confirmando que a escolaridade é um fator crítico na interpretação do TDR. Estes estudos apontam que indivíduos com menos anos de educação apresentam desempenhos significativamente inferiores, independentemente do estado cognitivo, o que pode levar a falsos positivos se não forem utilizados pontos

de corte ajustados. Foram também analisados os efeitos da idade, mostrando que o envelhecimento influencia aspetos como a organização espacial e a precisão gráfica, embora com menor impacto do que o nível educacional [12], [18].

Outro aspeto valorizado pela literatura portuguesa foi a comparação entre sistemas de pontuação — qualitativos, quantitativos e híbridos — com vista a determinar quais oferecem maior sensibilidade para alterações precoces. Estudos de validação nacionais sugerem que os sistemas híbridos ou quantitativos, ao atribuírem pontuações graduais a diferentes tipos de erros, aumentam a fiabilidade interavaliador e permitem uma distinção mais clara entre envelhecimento normal, comprometimento cognitivo ligeiro e demência. Estes resultados reforçam a utilidade do TDR em Portugal, mas também demonstram as limitações inerentes à avaliação manual, dependente da experiência do avaliador e vulnerável a vieses interpretativos [21], [47].

Assim, a realidade nacional confirma a relevância do TDR como instrumento de rastreio, mas também expõe lacunas importantes: os estudos disponíveis permanecem restritos à versão manual do teste, não existindo ainda bases de dados digitais portuguesas nem investigações que integrem o TDR com métodos computacionais. Esta ausência de transição para o digital impede que os avanços internacionais sejam validados em populações portuguesas e adaptados às especificidades linguísticas, culturais e educacionais do país [12], [18].

O *World Alzheimer Report 2023* destaca a necessidade urgente de estratégias de rastreio precoce adaptadas a diferentes realidades populacionais, enquanto dados recentes da Alzheimer Portugal indicam que mais de 200 mil pessoas vivem atualmente com demência no país, número que tende a crescer com o envelhecimento demográfico [48], [49].

Estas limitações constituem também oportunidades de investigação. A criação de bases de dados digitais nacionais do TDR, associadas a informação clínica relevante, representa um passo fundamental para validar modelos no contexto português e assegurar a sua aplicabilidade. A integração de abordagens multimodais, combinando métricas gráficas, cinemáticas, demográficas e biomarcadores, poderá aumentar a precisão preditiva [37]. O recurso a ensembles híbridos que conciliem ML e DL, aliado a técnicas de *Explainable* AI (XAI) como SHAP, LIME e *Grad-CAM*, permitirá desenvolver modelos mais robustos e transparentes, facilitando a sua aceitação clínica [40], [45]. Por fim, a realização de estudos multicêntricos e colaborativos será determinante para garantir validação externa, reprodutibilidade e escalabilidade, consolidando o papel do TDR digital como ferramenta central no rastreio precoce de défices cognitivos [25].

3. Metodologia

3.1 Amostragem e Recolha de Dados

O estudo, de natureza retrospetiva, foi realizado na clínica Legismente – Psiquiatria e Psicologia Clínica e Forense, com recurso a uma base de dados disponibilizada pela instituição, com parecer favorável de referência Nº.03-2025 emitido pela CE-ESTeSL a 3 de fevereiro de 2025. Todos os utentes, ou os seus representantes legais, tinham previamente consentido a utilização dos dados para fins de investigação, através da assinatura de consentimento informado.

Com base na literatura [50], [51], a amostra deve ser constituída no mínimo por 100 participantes de forma a garantir validade estatística, reduzir o risco de *overfitting* e viabilizar a utilização dos algoritmos de *Machine Learning*. Por esta razão, e tendo disponíveis um número limitado de dados, a amostra utilizada neste estudo é composta por 117 indivíduos maiores de idade e com indicação clínica para avaliação neuropsicológica, selecionados através de amostragem não probabilística por conveniência. Cada participante desenhou um relógio analógico em papel, seguindo instruções padronizadas, que incluíam a representação do círculo, dos números e dos ponteiros. Neste caso, não foi definida uma hora específica a ser desenhada pelos participantes. A classificação diagnóstica foi estabelecida de forma dicotómica, categorizando os registos como normal (N) e anormal (Abn). A Tabela 1 apresenta as características demográficas incluídas na análise — sexo, idade e escolaridade — selecionadas por serem os fatores sociodemográficos mais consensualmente associados ao desempenho no TDR e indicados como essenciais pelo neuropsicólogo responsável pela avaliação clínica.

Tabela 1 - Características da amostra recolhida.

		n	%	Min	Max	Média ± DP
Sexo	Masculino	43	36.8			
	Feminino	74	63.2			
Idade				18	99	62.4 ± 16.2
Escolaridade	Sem Escolaridade	3	2.6			
	Ensino Básico[1]	19	16.2			
	2º e 3º Ciclos[2]	22	18.8			
	Ensino Secundário[3]	31	26.5			
	Bacharelato / Licenciatura	36	30.8			
	Mestrado	5	4.3			
	Doutoramento	1	0.9			
Diagnóstico	N[4]	60	51.3			
	Abn[5]	57	48.7			

[1]Ensino Básico: 1º - 4º ano; [2]2º e 3º Ciclos: 5º - 9º ano; [3]Ensino Secundário: 10º - 12º ano; [4]N: Normal; [5] Abn: *Abnormal*

A adequação da amostra ao perfil demográfico português constitui um aspeto relevante, uma vez que o desempenho no TDR é sensível a fatores como idade e escolaridade. Estudos normativos nacionais demonstram que participantes com menor escolaridade apresentam maiores desvios espaciais e mais erros estruturais, independentemente do estado cognitivo [47]. Assim, verificar a correspondência entre a composição demográfica da amostra e a população portuguesa é fundamental para garantir a generalização dos modelos desenvolvidos.

3.2 Preparação da Base de Dados

Os grafismos de cada paciente foram digitalizados individualmente, tendo sido incluídos apenas os que se mantinham legíveis após a digitalização. Imagens com artefactos externos ou que prejudicassem a perceção da estrutura do relógio foram excluídas.

Cada grafismo digitalizado foi posteriormente armazenado e codificado com um identificador sequencial único, permitindo a sua associação controlada às variáveis demográficas recolhidas*, i.e.*, género, idade e escolaridade, em conformidade com a garantia de confidencialidade dos participantes.

3.3 Análise da Base de Dados

Para a realização da análise da base de dados composta pelos TDR recolhidos foi necessário desenvolver scripts em *Python*, com recurso à biblioteca *OpenCV*. Para esta análise foram desenvolvidos quatro *scripts* cuja metodologia será explicada ao longo deste capítulo.

3.3.1 Configuração, Processamento e Análise da Base de Dados

3.3.1.1 Configuração e Pré-Processamento

Durante a fase de visualização e digitalização dos testes recolhidos para construir a base de dados verificou-se uma grande variabilidade nas folhas onde os desenhos foram realizados, existindo, por vezes, em diversas posições, outros desenhos/escrita que, para esta análise, não tinham interesse. Por esta razão, foi necessário estabelecer configurações de recorte para cada imagem para focar a análise na área relevante – o teste do desenho do relógio. Foram estabelecidas ao todo treze configurações de recorte diferentes para todo o *dataset* tendo em conta as áreas que eram necessárias remover em cada folha. Na figura 2 estão representados três exemplos de folhas utilizadas para a realização dos testes.

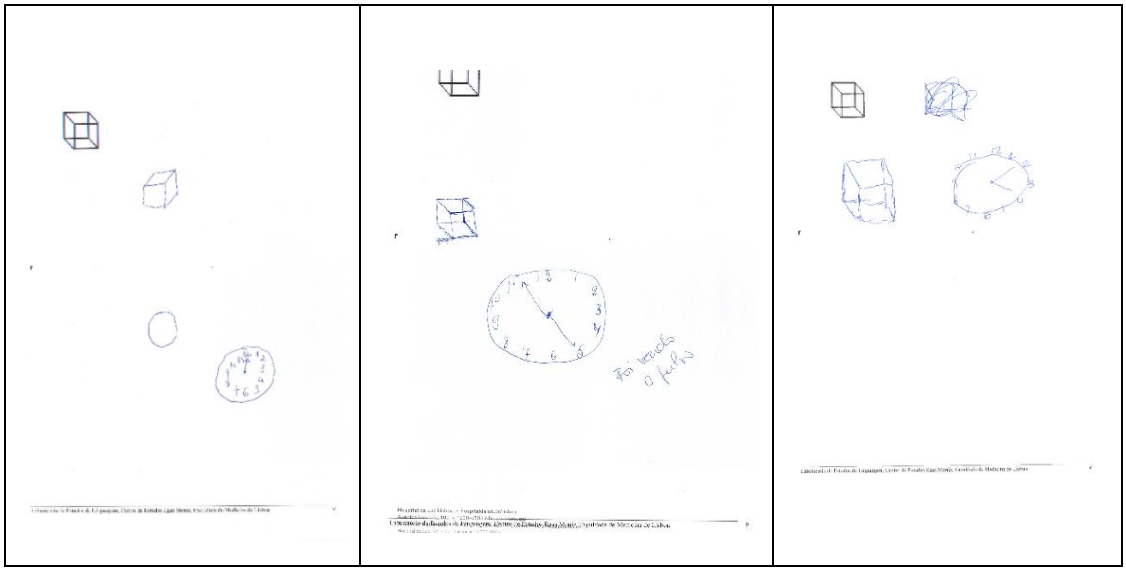


Figura 2 – Exemplos de folhas utilizadas para a realização dos TDR da base de dados.

De seguida, aplicou-se às imagens, já recortadas, uma série de etapas de pré processamento. Na primeira etapa, as imagens a cores foram convertidas em tons de cinzento para simplificar os dados que são retirados e assim diminuir as necessidades computacionais e, de seguida, submetidas a um filtro de suavização Gaussiano, ou seja, um desfoque que tem a função de suavizar e uniformizar os contornos e os detalhes das imagens, reduzindo assim o ruído presente. Posteriormente, aplicaram-se as duas estratégias complementares de binarização apresentadas na Figura 3:

1. **Binarização (*'Thresholding'*) com um limiar fixo** associado a uma operação morfológica de fecho ('*close'*) com uma máscara de *kernel* de dimensão 3x3, para realçar os traços principais do relógio;
2. **Binarização adaptativa**.

A primeira abordagem, baseada em *thresholding* fixo e operação morfológica de fecho, revelou-se eficaz na maioria dos casos; contudo, em imagens com variações de iluminação, traços muito ténues ou presença de ruído, produzia contornos incompletos ou perda de detalhes relevantes. Nesses casos, considerou-se que os resultados não eram satisfatórios e, como alternativa, aplicou-se a segunda estratégia — o *thresholding* adaptativo — que se mostrou mais robusta para realçar os traços do relógio de forma consistente. Na Figura 3 é possível observar a sequência das etapas de pré-processamento descritas acima.

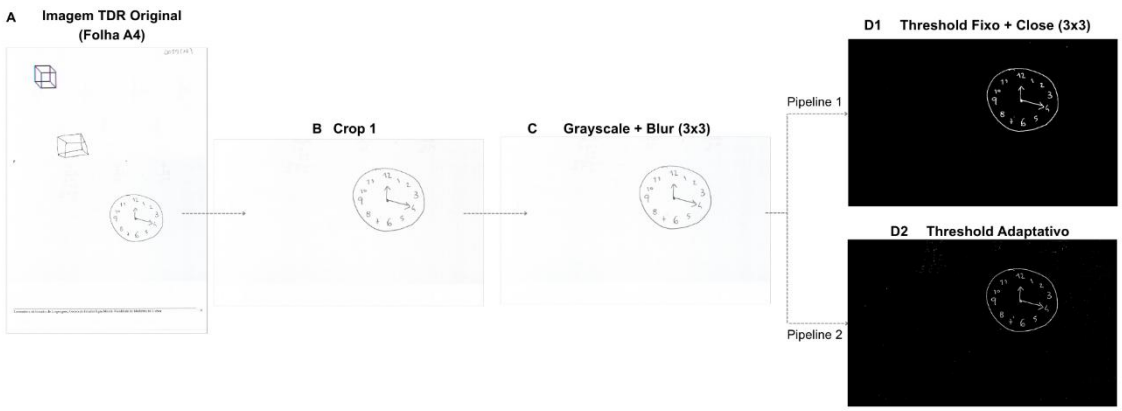


Figura 3 - Etapas de pré processamento aplicadas ao *dataset* de TDR.

3.3.1.2 Deteção do Contorno Principal do Relógio

Para a identificação do contorno principal correspondente ao mostrador do relógio foi necessário estabelecer uma função que combina a deteção de contornos, através da aplicação do algoritmo *cv2.findCountours (RETR_CCOMP, CHAIN_APPROX_SIMPLE)* – que implementa o algoritmo de Suzuki-Abe para seguimento de contornos – sobre a imagem binarizada com a Transformada de *Hough* [52], [53]. Em cada imagem são identificados todos os contornos externos e cada contorno candidato é avaliado com base, primeiramente, com base na sua área, sendo eliminados os considerados demasiado pequenos para representar a face do relógio - foi definido um limiar fixo de área com base nas áreas obtidas para todos os contornos correspondentes ao relógio encontrados. Para os restantes candidatos após esta primeira fase de eliminação foram calculados o centroide e a distribuição das distâncias dos pontos do contorno em questão a este centroide. Desta forma, calcula-se, para cada contorno candidato, uma métrica (‘*score*’) de circularidade baseada na variância das distâncias dos pontos do contorno ao seu centroide: quanto menor for essa variância, mais circular é o contorno e maior a sua pontuação. Esta pontuação é combinada com a área do contorno para formar uma pontuação composta que prioriza contornos simultaneamente grandes e circulares. Adicionalmente, aplica-se a Transformada de *Hough* para a deteção de círculos e compara-se cada contorno com o círculo detetado: quando o centroide do contorno se encontra a uma distância inferior a metade do raio desse círculo e o seu raio médio difere do respetivo raio em menos de 20%, a pontuação do contorno é duplicada, aumentando a probabilidade de ser selecionado. O contorno com maior pontuação final é designado como o contorno principal do relógio e deverá, idealmente, corresponder ao mostrador do relógio desenhado.

As equações utilizadas para estas determinações estão apresentadas abaixo (equações 1 - 3).

Seja C o conjunto de pixels do contorno de um objeto numa imagem, isto é:

$$\boldsymbol{C = \{(x_i, y_i)\, \epsilon\, \mathbb{Z}^2 \,|i = 1, 2, \ldots, N\}} \qquad (1)$$

onde cada (x_i, y_i) é a coordenada de um pixel do contorno.
Seja $p = (x_i, y_i)\, \epsilon\, \mathbb{R}^2$ o centroide da região correspondente. Definimos a distância euclidiana entre dois pontos $(x_1, y_1)\, e\, (x_2, y_2)$ como:

$$d\big((x_1, y_1), (x_2, y_2)\big) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad (2)$$

A distância mínima do ponto p a todos os pixels do contorno C é então:

$$D_{min}(p, C) = \min_{(x_i, y_i)\, \epsilon\, C} d\big((x_C, y_C), (x_i, y_i)\big) = \min_{(x_i, y_i)\, \epsilon\, C} \sqrt{(x_C - x_i)^2 + (y_C - y_i)^2} \qquad (3)$$

onde $D_{min}(p, C)$ representa a menor distância do ponto ao contorno C.

Após a identificação do contorno principal é calculado o menor polígono convexo que envolve os seus pixéis - *Convex Hull*. Este polígono é menos suscetível a pequenas irregularidades e fornece uma base mais estável para o cálculo do centroide da face do relógio e métricas espaciais que vão ser mencionadas mais adiante. Esta determinação é necessária principalmente em casos de relógios cujo contorno principal não está completamente fechado. Na Figura 4 está representado um exemplo da identificação do contorno principal após a deteção do *ConvexHull* (a azul) e o respetivo centroide (a verde).

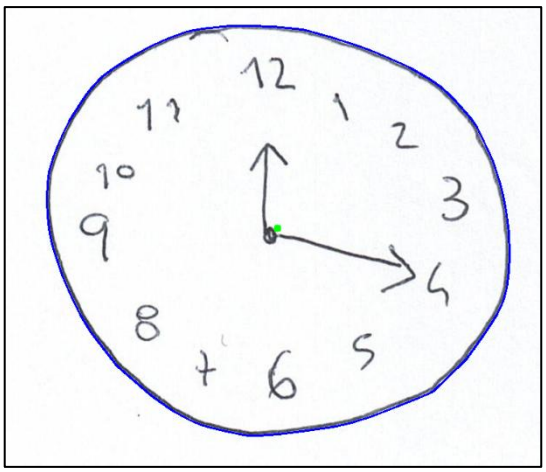


Figura 4 – Exemplo de identificação do contorno principal.

3.3.1.3 Normalização e Recorte do Mostrador do Relógio

Após a identificação do contorno principal, foi aplicado um novo recorte à imagem do relógio, com uma margem de 10 pixels, para ser analisada somente a região de interesse correspondente ao seu interior (números e ponteiros). Além disso, é realizado

o redimensionamento desta região para uma dimensão padrão (512×512 *pixels*). Este processo garantiu a normalização espacial das amostras de forma ser possível a sua aplicação consistente nas etapas seguintes de segmentação e análise, independentemente do tamanho original da imagem. O centroide e o contorno do relógio determinados anteriormente na imagem foram também normalizados para as novas coordenadas do recorte redimensionado.

3.3.1.4 Segmentação e Classificação de Componentes Internos

Os componentes internos, isto é, contornos localizados dentro do contorno principal são segmentados e classificados como números, ponteiros ou descartados. Para esta análise dos componentes foi criada uma nova função que começa por identificar todos os contornos que se encontram dentro do contorno principal, sendo que contornos muito pequenos foram eliminados (área inferior a 15 *pixels*). Para cada contorno candidato foram calculadas características, nomeadamente a área, caixa delimitadora ('*bounding box*'), centroide, distância mínima ao centro do relógio e a razão entre a largura e a altura ('*aspect ratio*') da caixa delimitadora.

Os ponteiros são identificados combinando a proximidade ao centro do relógio, *aspect ratio* e área. São considerados candidatos os contornos que integrem o conjunto dos maiores componentes e/ou apresentem forma muito alongada, medida pela razão entre o lado maior e o lado menor do respetivo retângulo envolvente alinhado aos eixos (*bounding box*); considera-se "muito alongado" quando esta razão é igual ou superior a 5. Um candidato só é aceite como ponteiro se estiver próximo do centro do relógio, isto é, se alguma parte do contorno ficar a menos de cerca de 30% do raio do mostrador. Quando são identificados dois ou mais ponteiros, estes são ordenados pelo comprimento (lado maior do retângulo envolvente): o mais curto é assumido como ponteiro das horas e o mais longo como ponteiro dos minutos.

Os componentes que não foram classificados como ponteiros são considerados candidatos a dígitos. A janela de cada candidato a dígito é extraída (matriz correspondente à *bounding box* de cada contorno) e redimensionada para 28x28 pixels. A classificação destes componentes candidatos a dígitos é realizada por uma função que utiliza um modelo de rede neuronal (*ResNet-18*) pré-treinada [54]. Este modelo foi treinado com o conjunto de dados *EMNIST-Digits* num script descrito mais adiante na secção 2.3.3 [55].

Os candidatos são finalmente submetidos à rede de classificação para serem classificados como dígitos caso apresentem uma área superior a um limiar mínimo definido com base na área que todos os componentes apresentaram (25 pixels). Para

além da área, é necessário apresentarem também um nível de confiança da classificação feita pelo modelo superior a 90% ou terem uma distância ao centro superior a pelo menos metade do raio do círculo do relógio.

Após a classificação inicial, é aplicada uma lógica de agrupamento para identificar e combinar os dígitos de dois algarismos, como é o caso dos números "10", "11" e "12", necessária para a correta interpretação dos números do relógio. Na Figura 5 está representado um exemplo de segmentação de todos os componentes do relógio, estando o contorno principal identificado a verde, os números assinalados a vermelho com a respetiva classificação indicada por cima e o(s) ponteiro(s) assinalados a azul.

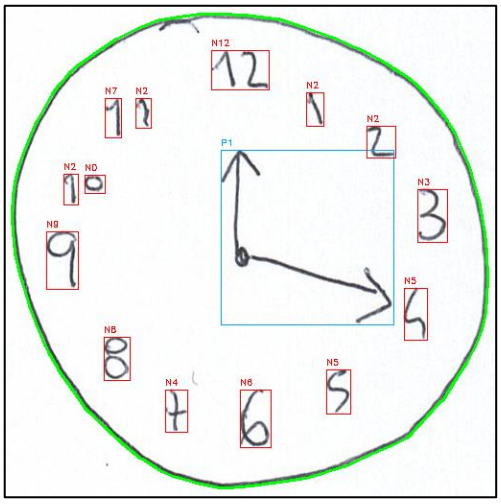


Figura 5 – Exemplo da segmentação final dos componentes encontrados num TDR.

Caso o número total de dígitos e ponteiros detetados nesta primeira passagem baseada na sequência de pré-processamento 1, referida na secção 3.3.1.1, seja inferior a um limiar definido como 12, a análise de componentes é repetida utilizando a imagem binarizada obtida pela pipeline de pré-processamento 2. A lista final de componentes é escolhida a partir da passagem que detetou o maior número de componentes. Na Figura 6 está um exemplo de um caso onde foi necessário recorrer à *pipeline* 2 (imagem apresentada à direita), após a sequência 1 ter resultado na segmentação de um número de componentes inferior ao limiar definido (imagem apresentada à esquerda).

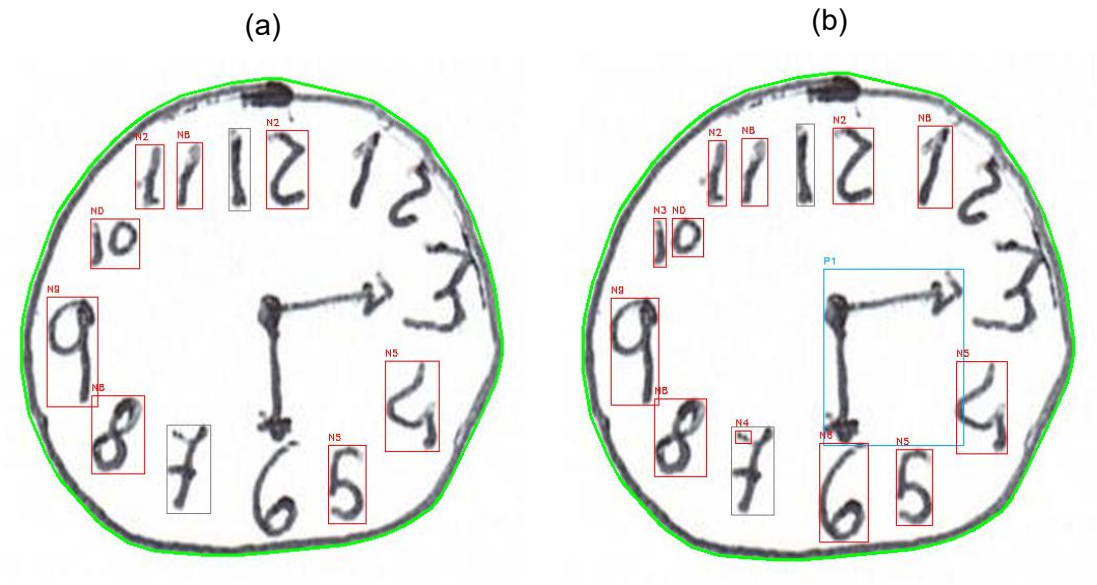


Figura 6 – Exemplo de TDR onde se aplicou a primeira (a) e a segunda (b) sequências de pré-processamento.

Depois de obtidos os elementos do TDR descritos anteriormente foram extraídas as características, apresentadas na Tabela 2. A seleção destas características foi motivada pela relevância clínica e pelo que os sistemas clássicos de pontuação captam no TDR (por exemplo, Shulman, Sunderland, Rouleau, Freedman): a geometria do mostrador (circularidade e área) resume assimetria do círculo; a organização numérica — contagem de dígitos detetados, erro médio do ângulo de cada número, sequência 1–12 e distância média dos números ao contorno principal — operacionaliza critérios usados na avaliação manual (colocação, ordem e espaçamento dos algarismos); as características do traço — espessura média e variabilidade, conectividade e extensão — refletem tremor, hesitações e correções frequentemente descritas em versões digitais do teste; os ponteiros — número de ponteiros, proximidade ao centro e razão de comprimentos (minutos e horas) — traduzem regras clínicas de configuração e orientação; métricas globais do traço do relógio (espessura/conectividade) sintetizam consistência motora. Estas dimensões são convergentes com o que a literatura reporta para validade do TDR e para registos digitais com caneta/tablet, onde o processo de desenho (pausas, revisões, cinemática) é informativo para o rastreio cognitivo [3], [12], [21].

Tabela 2 – Métricas calculadas para todas as características extraídas das imagens de TDR

Componente	Características	Métricas
Contorno Principal	A. Regularidade/Deformação da forma da face (Circularidade)	1. Variância das distâncias dos pontos do contorno original ao centroide 2. Variância das distâncias dos pontos do convex hull do contorno ao seu centroide
	B. Propriedades do traço do contorno do relógio	3. Espessura média do traço do contorno 4. Desvio padrão da espessura do traço do contorno 5. Número de componentes conectados do contorno
Números	C. Contagem de números identificados	6. Número total de componentes identificados e classificados como números
	D. Precisão angular das posições dos números	7. Erro angular médio entre a posição angular detetada de cada número e a sua posição angular ideal num relógio analógico
	E. Ordem angular dos números	8. Proporção de transições angulares corretas entre números consecutivos
	F. Dispersão angular das posições dos números	9. Desvio padrão circular dos ângulos das posições dos centroides dos números em relação ao centro do relógio
	G. Proximidade dos números à fronteira do relógio	10. Distância média dos centroides dos números à fronteira do convex hull
	H. Tamanho e área dos números	11. Área total dos componentes classificados como números 12. Área média dos componentes classificados como números 13. Desvio padrão dos componentes classificados como números
	I. Propriedades do traço dos números	14. Espessura média dos traços dos números 15. Desvio padrão entre as espessuras médias dos diferentes números 16. Média dos desvios padrão individuais da espessura de cada número 17. Média da conectividade (componentes) dos traços dos números 18. Média da extensão (área/componente) dos traços dos números
Ponteiros	J. Contagem de ponteiros identificados	19. Número total de componentes identificados como ponteiros
	K. Relação de comprimento entre ponteiros (Minuto/Hora)	20. Razão entre o comprimento do ponteiro mais longo (minutos) e o do ponteiro mais curto (horas), se pelo menos dois ponteiros forem detetados
	L. Tamanho e área dos ponteiros	21. Área total dos componentes identificados como ponteiros 22. Área média dos componentes identificados como ponteiros 23. Desvio padrão dos componentes identificados como ponteiros
	M. Propriedades do traço dos ponteiros	24. Espessura média dos traços dos ponteiros 25. Média dos desvios padrão individuais da espessura de cada ponteiro 26. Média da conectividade (componentes) dos traços dos ponteiros 27. Média da extensão (área/componente) dos traços dos ponteiros

À medida que cada imagem é processada, os valores das suas métricas são guardados como um dicionário numa lista *Python*. No fim do ciclo, esta lista é convertida num *DataFrame* da biblioteca *pandas* (tabela em que cada linha corresponde a uma imagem

e cada coluna a uma métrica) e é guardada num ficheiro .csv com as métricas originais. Em paralelo, os valores numéricos de cada imagem são adicionados a uma lista de "vetores de métricas", que é convertida para um *array NumPy* e guardada num ficheiro .npy. Estes vetores são depois normalizados e tanto o *array* normalizado (.npy) como a tabela com as métricas normalizadas (.csv) são também guardados, juntamente com o normalizador utilizado.

Finalmente, aplicou-se padronização (*z-score*) com recurso à classe *StandardScaler* da biblioteca *scikit-learn*, garantindo que todas as métricas tivessem peso comparável na análise. O normalizador foi ajustado aos vetores de métricas originais (*fit*) e aplicado aos dados (*transform*), sendo guardado para assegurar reprodutibilidade. A Equação (4), apresentada posteriormente (3.3.4.1), formaliza de forma analítica esta técnica de normalização.

Os vetores de métricas normalizados são armazenados num ficheiro no formato .npy e, adicionalmente, o objeto *StandardScaler* treinado que encapsula os parâmetros de normalização é guardado num ficheiro através da biblioteca *joblib*. Este passo é fundamental para assegurar a consistência na normalização de quaisquer novos dados que venham a ser processados, permitindo que sejam transformados utilizando a mesma escala definida pelos dados originais. Finalmente, as métricas normalizadas são combinadas com as informações de identificação da imagem num *DataFrame pandas* e guardadas num ficheiro .csv.

3.3.2 Treino do modelo *ResNet-18*

Para a tarefa de classificação de dígitos utilizou-se o *dataset EMNIST Digits*, uma extensão do MNIST com dígitos manuscritos [55]. O *dataset* foi dividido nos conjuntos de treino e teste e as imagens foram pré-processadas através das seguintes transformações: conversão para escala de cinza (garantindo um único canal), redimensionamento para 28x28 *pixels* e normalização dos valores dos *pixels*.

A arquitetura de rede neuronal escolhida foi uma versão modificada do modelo já existente *ResNet-18*. As modificações foram realizadas para adaptar o modelo à entrada da imagem 28x28 e de um só canal, uma vez que a arquitetura *ResNet-18* original é tipicamente utilizada para imagens RGB maiores (224x224 *pixels*). Especificamente, a primeira camada convolucional foi ajustada para aceitar um canal de entrada e o *maxpool* inicial foi removido para preservar a resolução espacial em imagens de menor dimensão. A camada totalmente ligada (*fully-connected*, FC) final da rede *ResNet*-18 original (treinada para o ImageNet, com 1000 unidades de saída) foi substituída por uma

nova camada com 10 unidades de saída, correspondentes às 10 classes de dígitos existentes (0 a 9) [54].

O modelo foi treinado com a função de perda *Cross-Entropy*, apropriada à classificação multiclasse. Após a retro propagação, os pesos foram ajustados pela função de otimização Adam, configurado com taxa de aprendizagem inicial de 0,002 e regularização L2 (‘*weight decay*’) de 1×10^{-4}. Para controlar a aprendizagem ao longo do treino, implementou-se um *scheduler* StepLR, que reduz a taxa em 0,1 a cada 5 épocas, favorecendo um ajuste fino nas fases finais. Para acelerar o treino e reduzir o uso de memória, recorreu-se a treino com a técnica de *Mixed Precision Training*, utilizando os utilitários *autocast* e *GradScaler*, que executam operações em meia precisão (FP16 - *Floating Point* de 16 *bits*), quando apropriado, no processador.

O treino decorreu ao longo de 10 épocas. Ao longo das iterações, monitorizou-se apenas a perda no conjunto de treino para acompanhar a convergência. No final, os pesos do modelo foram guardados para posterior utilização na inferência e para avaliação em dados não vistos (conjunto de teste), apresentada adiante no capítulo 4.

3.3.3 Análise de Correlação entre Características

Nesta etapa avaliou-se a correlação entre as características previamente extraídas, com o objetivo de identificar relações de dependência e redundância entre variáveis.

Inicialmente, os dados resultantes do processamento de imagem foram carregados a partir de um ficheiro consolidado em formato .csv, contendo todas as métricas calculadas e dados de identificação de cada teste de TDR recolhidos na fase inicial deste trabalho. Apenas as variáveis numéricas foram selecionadas para a análise estatística. De seguida calculou-se então a matriz de correlação *Pearson* , isto é, o coeficiente que quantifica a associação linear entre pares de variáveis, obtido pela covariância normalizada pelas dispersões (varia entre −1 e 1, sendo que valores próximos de 1 indicam correlação positiva forte, próximos de −1 correlação negativa forte e próximos de 0, ausência de relação linear). Este coeficiente é sensível a *outliers* e pressupõe uma relação aproximadamente linear. A matriz foi calculada com *pandas* e representada num mapa de calor com recurso à biblioteca *Seaborn*, para destacar pares com correlação elevada em valor absoluto e suportar a identificação de métricas redundantes.

Posteriormente, foram selecionados pares específicos de características que apresentavam elevada correlação e foi construído o respetivo gráfico de dispersão, permitindo avaliar visualmente a relação entre as métricas e confirmar possíveis sobreposições de informação.

Com base nesta análise das correlações foi realizada a primeira redução do conjunto inicial de características, eliminando-se aquelas consideradas redundantes (descrita em maior detalhe no capítulo 4.3). Esta análise serviu de base para a seleção final das características a serem utilizadas nos modelos de classificação subsequentes.

3.3.4 Treino e Avaliação dos Modelos de Classificação

Foi desenvolvido um script com quatro *pipelines* comparativas, com dois objetivos principais: (i) identificar um subconjunto de características efetivamente informativas para a predição do diagnóstico e (ii), com base nesse subconjunto, treinar um classificador capaz de estimar se o teste é Normal (N) ou Anormal (Abn). A seleção de características revelou-se necessária dado que nem todas as variáveis fornecem informação discriminativa relevante, existindo métricas altamente correlacionadas entre si, presença de ruído e um número reduzido de exemplos – fatores que, em conjunto, podem conduzir a *overfitting*, aumentar a variância dos modelos e diminuir a interpretabilidade.

Os quatro *pipelines* partilharam etapas comuns, incluindo o carregamento, o pré-processamento, a divisão e a normalização dos dados, diferenciando-se apenas na estratégia de seleção de características. O método com melhor desempenho é escolhido segundo métricas pré-definidas (*Area Under the Curve* - AUC, *F1-score*, sensibilidade e especificidade), sendo o respetivo modelo adotado para a fase de inferência. Todas as abordagens compartilham as etapas iniciais: carregamento e pré-processamento dos dados, divisão e normalização dos dados.

3.3.4.1 Carregamento, Pré-Processamento, Divisão e Normalização dos Dados

Para o carregamento e pré-processamento dos dados, foram lidas a partir de um arquivo .csv as características extraídas do TDR (variáveis descritivas geradas no processamento de imagem – métricas do mostrador, dos algarismos e dos ponteiros, espessura/variabilidade do traço, conectividade, etc. Do ficheiro, manteve-se a coluna-alvo (‘diagnóstico’) e as colunas de características, excluindo os números identificadores de cada TDR. Em seguida, filtraram-se valores ausentes nas colunas das características, aplicou-se padronização (*z-score*) e eliminaram-se as variáveis altamente correlacionadas com base na matriz de correlação de *Pearson* para reduzir redundância antes da modelação. A escolha da padronização pelo *z-score* justifica-se pelo facto de colocar todas as variáveis numa escala comparável, com média zero e desvio padrão unitário, sem restringir os valores a um intervalo fixo. Esta abordagem é particularmente adequada quando as variáveis apresentam diferentes ordens de magnitude e distribuições aproximadamente gaussianas, permitindo que algoritmos

sensíveis à escala, como SVM ou regressão logística, funcionem de forma mais estável e equilibrada. O *z-score* (z) é definido pela seguinte expressão:

$$z = \frac{x - \mu}{\sigma} \quad (4)$$

onde, x é o valor observado, μ o valor da média dos dados e σ o desvio padrão.

Posteriormente, o conjunto de dados, 21 características selecionadas e a variável alvo, foi submetido à divisão em conjuntos de treino e teste, sendo 75% dos dados para treino e 25% para teste. Esta divisão foi realizada de forma estratificada com o intuito de assegurar que a proporção das classes (Normal/Abnormal) fosse preservada em ambos os conjuntos.

3.3.4.2 Seleção de Características

Numa primeira abordagem de seleção de características após a primeira fase de eliminação pela correlação foi aplicada a técnica de *Recursive Feature Elimination with Cross-Validation* (RFECV) individualmente para cada modelo de classificação a ser avaliado. O RFECV utiliza um estimador base para remover, de forma repetida, as características menos importantes, avaliando o desempenho (*accuracy*) em validação cruzada (CV) a cada passo.

Optou-se por avaliar quatro famílias de modelos complementares para dados tabulares: Regressão Logística, *Support Vector Machine* (SVM), *Random Forest* e *Gradient Boosting*. A Regressão Logística funciona como base linear, é interpretável e produz probabilidades calibráveis, aspeto relevante em contexto clínico. O SVM é adequado a amostras pequenas/médias e a relações não lineares, beneficiando da padronização previamente aplicada às características. Os métodos por árvores — *Random Forest* e *Gradient Boosting* — são não paramétricos, capturam interações e não linearidades sem engenharia de características, são robustos à colinearidade e são úteis para análise explicativa. A inclusão de ambos permite comparar duas estratégias distintas: *bagging* (Random Forest), mais estável e menos sensível ao ruído, e *boosting* (Gradient Boosting), frequentemente superior em problemas tabulares pela sua capacidade de modelar padrões mais subtis. [56], [57], [58], [59]. Em conjunto, estes modelos cobrem um espectro de viés–variância e de interpretabilidade–desempenho, permitindo selecionar a melhor abordagem para predizer o diagnóstico (Normal vs. Anormal) com validação cruzada e métricas de desempenho. Além disso, privilegiaram-se modelos clássicos de *Machine Learning* em vez de redes profundas devido à dimensão da amostra e à necessidade de garantir interpretabilidade. Numa fase inicial de investigação e com dados limitados, soluções mais leves e transparentes oferecem

maior controlo metodológico, melhor explicabilidade e menor risco de overfitting, justificando a escolha destes modelos. Cada um dos quatro modelos foi sujeito a um processo de otimização de hiperparâmetros através *GridSearchCV*. Para tal, definiu-se uma grelha de variação dos valores dos hiperparâmetros mais influentes de cada modelo e o *GridSearchCV* procedeu à avaliação de cada combinação. Para este processo é utilizada a validação cruzada estratificada (com 5 *folds*), tendo como métrica de otimização a *accuracy*. A métrica *accuracy*, ou exatidão em português, reflete a proporção de observações corretamente classificadas, isto é, o número total de predições corretas dividido pelo número total de amostras, calculada em cada *fold* e depois feita a média entre todas *foldas*. A combinação de hiperparâmetros que resultou no melhor score médio de *accuracy* na validação cruzada foi selecionada como a configuração ideal para o modelo em questão.

Para avaliar o desempenho dos modelos, os 117 TDR foram divididos de forma estratificada, preservando a proporção entre casos Normal (N) e Abnormal (Abn). Em cada iteração da validação cruzada utilizada, 80% dos dados foram destinados ao treino/validação interna e 20% ao teste final. Assim, em média, cerca de 93 TDR foram usados em treino/validação e 24 em teste por *fold*. Esta estratégia permitiu maximizar o uso da amostra disponível, reduzindo o risco de *overfitting* e garantindo que cada modelo fosse avaliado em dados não utilizados durante o treino.

Finalmente, procedeu-se à avaliação da performance final deste modelo no conjunto de teste, que compreende dados não utilizados durante as fases de treino e otimização. Foram realizadas a comparação e visualização dos resultados de cada modelo, nomeadamente o número de características selecionadas, os melhores hiperparâmetros e o valor da *accuracy* antes e depois da aplicação do *Gridsearch*, de forma a facilitar a análise. Adicionalmente, foram obtidas as matrizes de confusão (contagens de verdadeiros/falsos positivos e verdadeiros/falsos negativos) para cada modelo e calculadas as seguintes métricas: sensibilidade (‘*recall’*) — capacidade de detetar casos positivos/anormais; especificidade — capacidade de identificar corretamente os negativos/normais; precisão — fração de casos que realmente são positivos entre os classificados como positivos; *F1-score* — média entre precisão e *recall*, útil com classes não balanceadas; curva ROC — relação entre taxa de verdadeiros positivos e taxa de falsos positivos com variação do limiar; e AUC — área sob a curva ROC, medida global de discriminação. As métricas referidas acima estão definidas nas seguintes expressões:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5) \qquad Precisão = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7) \qquad F1 - score = \frac{2 \,.\, Precisão \,.\, Recall}{Precisão + Recall} \quad (8)$$

$$Especificidade = \frac{TN}{TN + FP} \quad (9)$$

onde:

- TP (*True Positives*) – número de positivos corretamente classificados;
- TN (*True Negatives*) – número de negativos corretamente classificados;
- FP (*False Positives*) – número de negativos incorretamente classificados como positivos;
- FN (*False Negatives*) – número de positivos incorretamente classificados como negativos.

Após esta primeira abordagem, verificou-se que existia um conjunto pequeno de características selecionadas em comum por todos os métodos de classificação. Depois de uma análise crítica das restantes características, definiu-se um conjunto final fixo de características para a segunda abordagem de avaliação dos métodos de classificação. Esta segunda abordagem de avaliação realizada com o conjunto fixo de características foi aplicada para os mesmos quatro modelos, tendo sido feita uma otimização para a métrica *accuracy*. Todos os resultados descritos anteriormente estão apresentados no capítulo 4.

4. Resultados e Discussão

Nesta secção vão ser apresentados e discutidos os resultados obtidos após a aplicação do método proposto para a segmentação, extração e seleção de características e classificação das imagens de TDR, descrito anteriormente no capítulo 3, subdividindo-se em cinco vertentes principais. A primeira vertente prende-se com a avaliação do desempenho das etapas de pré processamento e segmentação dos TDR. Em segundo lugar são apresentados os resultados relativos à extração e análise das características obtidas e, em terceiro, os resultados obtidos após o treino do modelo *ResNet-18* para classificação de dígitos. A quarta vertente apresentada é a seleção das características mais relevantes e, por último, é apresentada uma avaliação comparativa dos diferentes modelos de classificação aplicados, destacando-se os respetivos desempenhos. No final é realizada uma discussão geral, onde os resultados são interpretados de forma crítica, considerando-se as suas implicações, limitações e potenciais desenvolvimentos futuros.

4.1 Processamento e Segmentação das Imagens de TDR

A aplicação do pipeline de processamento e segmentação identificou um contorno principal em todas as 117 imagens. A circularidade desse contorno foi melhor do que a mediana do conjunto em 59 das 117 imagens e particularmente elevada em 30 imagens, isto é, estão entre os 25% melhores casos. A mediana corresponde ao valor central de um conjunto de dados após serem ordenados. Por não ser afetada por valores extremos, constitui uma medida robusta de tendência central, especialmente útil quando existem distribuições assimétricas ou valores atípicos. A combinação da transformada de *Hough* com a aplicação do *convex hull* permitiu, na grande maioria dos casos, contornar irregularidades apresentadas nos desenhos e garantiu uma deteção robusta, mesmo em casos de traços incompletos. Na Figura 7 é possível observar alguns exemplos que corroboram a eficácia do método utilizado para a identificação do contorno principal, apesar da existência de uma elevada variabilidade de grafismos. No entanto, na imagem (d) é possível verificar que em casos mais complexos existe margem para melhorar este método, visto que é visível que o contorno identificado não corresponde exatamente ao mostrador do relógio.

Relativamente à segmentação dos componentes internos do relógio (números e ponteiros), procedeu-se a uma análise quantitativa do conjunto completo de 117 imagens. Assim, observou-se que foi possível detetar pelo menos um ponteiro em 105 das 117 imagens (89,7%), enquanto dois ponteiros foram identificados em apenas 18 de 117 (15,4%); em 12 de 117 (10,3%) não foram detetados ponteiros e, pontualmente,

surgiram 3 casos com três ponteiros (falsos positivos). No que respeita aos algarismos foram detetados em média 15,3 dígitos por imagem e um número igual ou superior a 12 dígitos em 99 de 117 imagens (84,6%). Apesar destes resultados encorajadores, a extração interna revelou-se naturalmente mais difícil do que a do contorno principal, devido à elevada proximidade e sobreposição de traços, à presença de múltiplos ponteiros e a variações de escrita, fatores que explicam os casos com ausência de ponteiros ou contagens de dígitos mais baixas.

Uma das principais dificuldades foi a identificação de múltiplos ponteiros pois, na maioria dos casos, não existia uma separação clara entre eles, o que dificultou a sua identificação individual. Esta dificuldade foi ultrapassada em alguns casos com a utilização da *pipeline* 2, como é possível observar nas imagens (g) e (h) da Figura 7, que correspondem à aplicação das *pipelines* 1 e 2, respetivamente. Além disso, revelou-se desafiante reconhecer números de dois algarismos como um único componente pois o processo baseia-se em componentes conexos e, nesses números (10, 11, 12), os dois algarismos surgem quase sempre como objetos separados, exigindo um agrupamento posterior. Esse agrupamento falha com frequência devido à grande variabilidade de espaçamento e rotação entre algarismos e, no caso particular do “11”, à semelhança com marcas finas do mostrador. Por outro lado, alargar demasiado os critérios de união aumenta o risco de fusões indevidas com elementos vizinhos.

Importa ainda referir que, na imagem (c) da Figura 7, não foi possível detetar os ponteiros porque a etapa de binarização/segmentação não conseguiu separar os traços dos ponteiros dos dígitos adjacentes e até do contorno exterior do relógio, fazendo com que vários elementos se fundissem num único componente e inviabilizando a sua identificação.

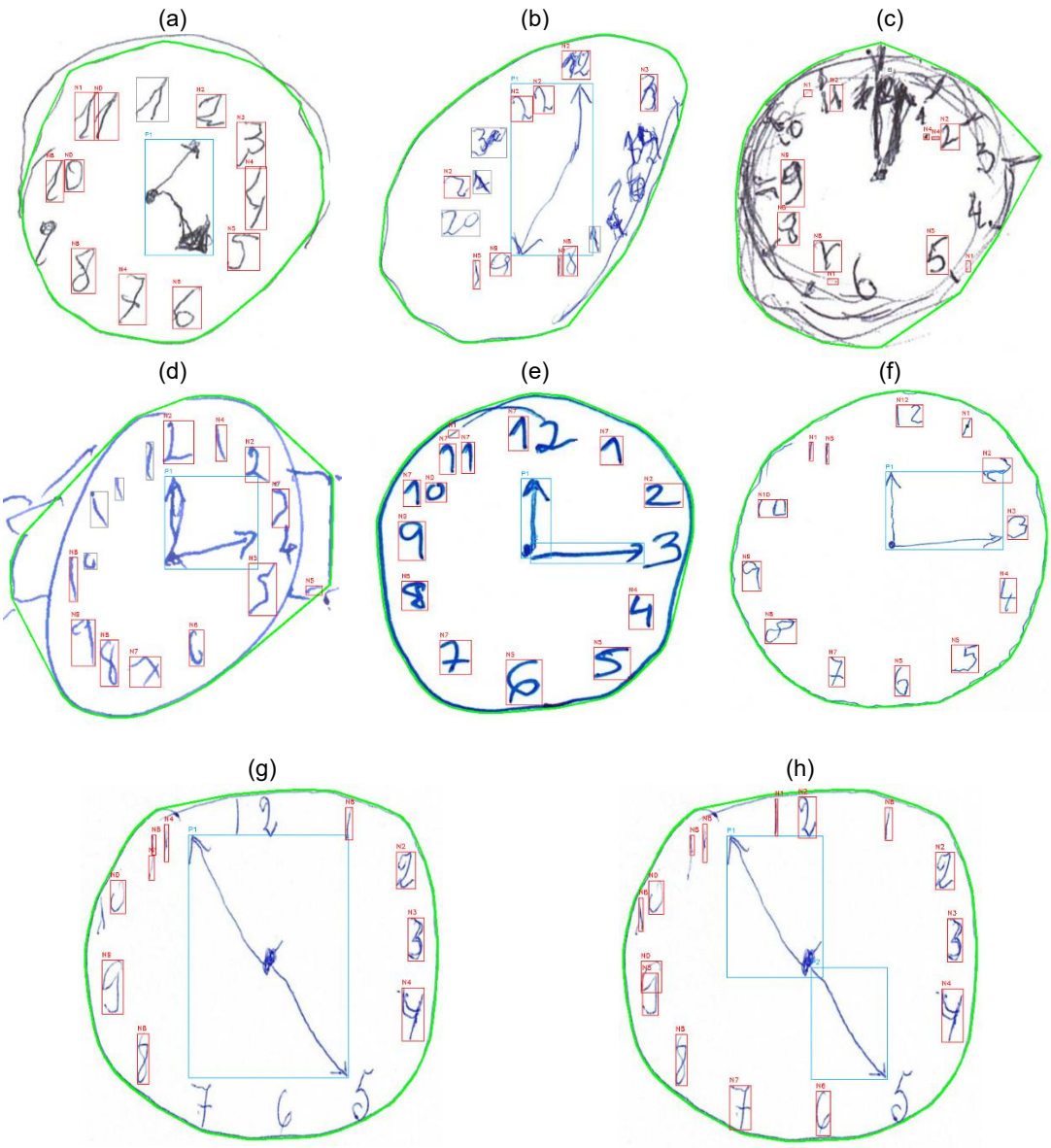


Figura 7 – Exemplos da segmentação obtida em imagens de TDR com elevada variabilidade de grafismos.

Desta forma, a utilização das duas *pipelines* foi bem-sucedida na sua globalidade e os resultados desta etapa demonstrarem a viabilidade da abordagem proposta para segmentação automática de TDR. Contudo, é visível nos exemplos da Figura 7 que o pré-processamento poderia ser ainda mais refinado com vista à obtenção de resultados ainda mais satisfatórios nos desenhos mais complexos em termos de grafismos.

4.2 Avaliação do Modelo de Classificação de Dígitos

Após o treino do modelo de classificação de dígitos utilizado, descrito anteriormente na secção 3.3.2, obteve-se uma *accuracy* de 99.65% e na Tabela 3 é possível verificar os resultados obtidos do treino do modelo para as métricas de precisão, *recall* e *F1-score* para cada dígito.

Tabela 3 – Resultados obtidos para as métricas de precisão, *recall* e *F1-score* do modelo de classificação de dígitos *ResNet-18* treinado com o *dataset EMNIST-Digits.*

Dígito	Precisão	Recall	F1-score
0	0.9985	0.9968	0.9976
1	0.9975	0.9968	0.9971
2	0.9938	0.9978	0.9958
3	0.9955	0.9962	0.9959
4	0.9967	0.9960	0.9964
5	0.9960	0.9972	0.9966
6	0.9980	0.9970	0.9975
7	0.9977	0.9930	0.9954
8	0.9970	0.9968	0.9969
9	0.9943	0.9975	0.9959

Apesar de o treino do modelo ter evidenciado um desempenho muito elevado, quando este modelo foi aplicado na *pipeline* de segmentação (avaliação dos resultados obtidos apresentada na secção 4.1), verificou-se que a classificação dos dígitos nas imagens de TDR foi apenas parcialmente bem-sucedida. Observou-se, após análise das imagens posterior à classificação, que se sucederam ainda uma quantidade classificações erradas de dígitos, ocorrendo principalmente a troca da classificação entre os números '1' e '7' e '4' e '7'. Esta troca pode dever-se ao facto de a área de contorno destes números ser mais reduzida e semelhante entre si ou mesmo pela deformação elevada que alguns dígitos apresentaram. A literatura evidencia que o reconhecimento de dígitos manuscritos é um problema complexo devido à elevada variabilidade das suas representações gráficas. Entre os principais fatores apontados estão a deformação, a variação de escala e a inconsistência na espessura do traço, que dificultam a diferenciação fiável entre dígitos morfologicamente semelhantes e, consequentemente, aumentam a probabilidade de classificações incorretas [60].

4.3 Análise das Características Extraídas

Inicialmente, como é possível verificar na Tabela 2, foram extraídas 27 características. Em seguida foi calculada a matriz de correlação de Pearson e consideraram-se altamente correlacionadas as variáveis com coeficiente no valor absoluto igual ou superior a 0,80. Esse valor foi escolhido porque é um patamar de referência amplamente usado para classificar correlação 'muito elevada' e, nos dados, esse corte permitiu ter um equilíbrio entre a redução de colinearidade e a preservação de informação, removendo apenas os pares claramente redundantes. Aplicando este critério (Figura 8), identificaram-se seis pares de características: característica 12 vs característica 18; característica 12 vs característica 11; característica 13 vs. característica 14; característica 22 vs. característica 27; característica 22 vs característica 21; característica 25 vs. característica 24 (ver descrição das métricas na Tabela 2 - pág.35).

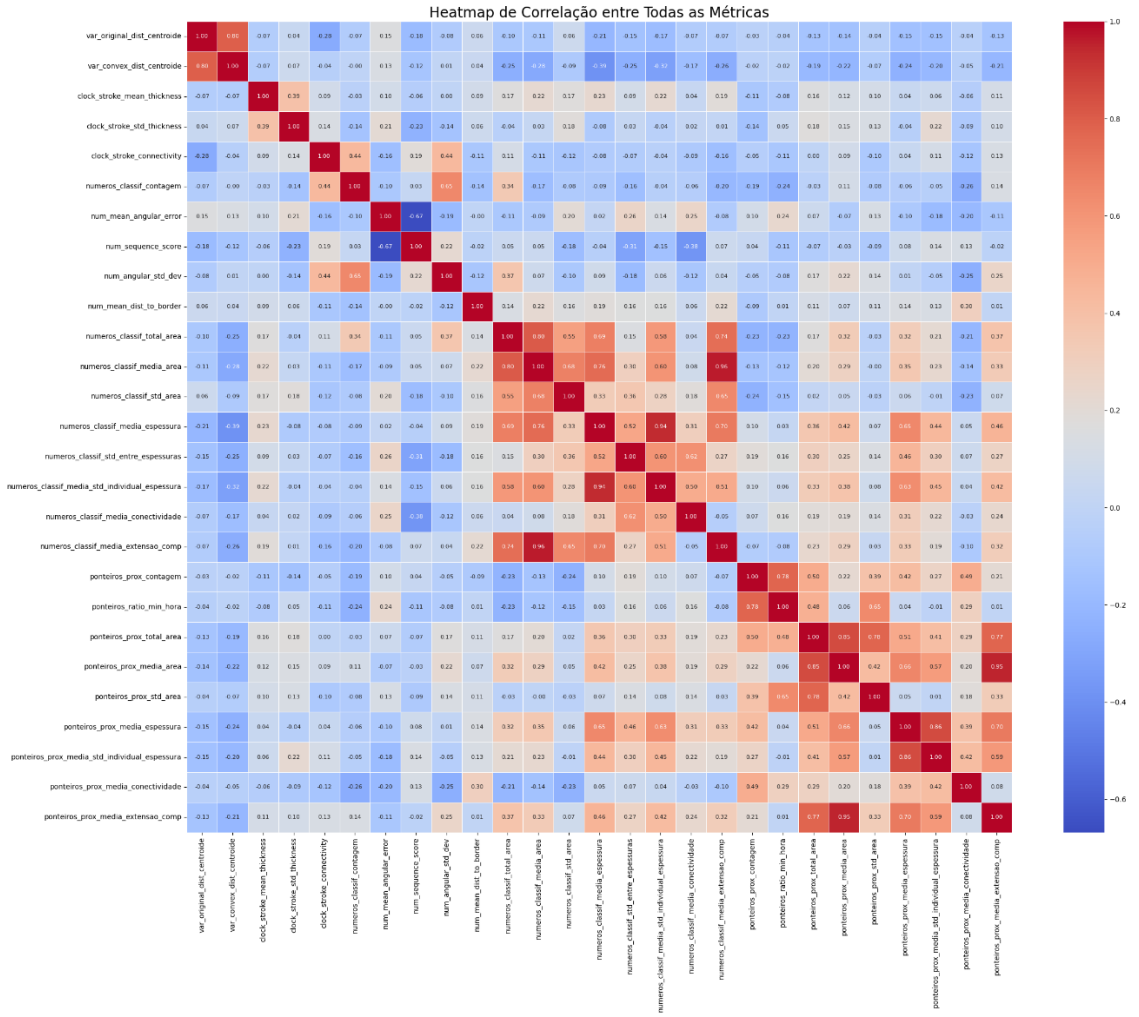


Figura 8 – Mapa de correlação do conjunto inicial de 27 características extraídas dos TDR.

No caso do primeiro par, ambas as métricas estão relacionadas ao tamanho dos números, mas a característica 12 é uma métrica mais direta e frequentemente utilizada para descrever o tamanho de objetos, enquanto que a característica 18 pode ser mais suscetível a variações na forma ou fragmentação dos números. Assim, manter a área média fornece uma medida mais robusta e intuitiva do tamanho típico dos números. Este resultado está em consonância com o reportado na literatura, onde a área é descrita como uma métrica recorrente e eficaz para a distinção entre dígitos manuscritos [61].

No caso do segundo par, a característica 11 pode ser fortemente influenciada pelo número de componentes detetados (por exemplo, se um número como o '8' for dividido em dois componentes). A característica 12 fornece uma medida mais consistente do tamanho *típico* de um número, independentemente de como ele foi segmentado, tornando-a mais útil para comparar o tamanho dos números entre diferentes imagens. Este tipo de abordagem é suportado pela literatura de reconhecimento de dígitos manuscritos, onde métricas geométricas estáveis (como área ou largura/altura do

contorno) são escolhidas devido à sua robustez em relação a variações de fragmentação ou ruído [29].

Para o par 3, a característica 14 está associada à espessura dos traços dos números, no entanto, a característica 13 indica a média da variabilidade da espessura dentro de cada número, podendo ser mais útil para identificar problemas na uniformidade do traço dos números. Estudos de reconhecimento de dígitos manuscritos indicam que incluir métricas de variabilidade, em vez de apenas médias, pode aumentar a capacidade de diferenciação do modelo no que diz respeito a variações na escrita [30].

Para o quarto par, semelhante ao par 1, a característica 22 é uma métrica mais direta e comum para o tamanho dos ponteiros do que a característica 27, que pode estar sujeita a ruído ou fragmentação.

No caso do quinto par, seguindo a mesma lógica do par 2, característica 21 pode ser distorcida pela contagem de componentes (se um ponteiro for segmentado), logo a característica 22 fornece uma medida mais estável e representativa do tamanho típico de um ponteiro.

Por último, para o par 6 e de forma semelhante à escolha realizada no par 3, a métrica que considera o desvio padrão individual da espessura dos ponteiros (característica 25) oferece uma medida mais detalhada da uniformidade da espessura dos ponteiros do que apenas a espessura media (característica 24). A ênfase na variabilidade interna está em consonância com abordagens em reconhecimento de manuscritos que destacam a relevância de métricas de dispersão ou estrutura interna para capturar inconsistências subtis [62].

Os padrões observados entre os grupos Normal e Abn são coerentes com descrições clássicas da literatura clínica. Estudos como os de Rouleau, Sunderland e Freedman [21], [22] documentam maior dispersão angular dos números, irregularidades no contorno do círculo e desvios na orientação dos ponteiros em indivíduos com défice cognitivo. As métricas que no presente estudo mostraram diferenças mais pronunciadas — simetria do mostrador, precisão angular e variabilidade do traço — corresponderam precisamente aos aspetos referidos como mais sensíveis à deterioração cognitiva, reforçando a pertinência das características selecionadas.

4.4 Seleção Final de Características e Avaliação dos Modelos de Classificação

Após a primeira fase de eliminação de características tendo em conta as correlações obtidas entre si, procedeu-se então à aplicação das duas abordagens para identificar

um novo subconjunto de características mais relevantes em cada modelo de classificação testado. A primeira abordagem consistiu na seleção de características através da técnica *Recursive Feature Elimination with Cross-Validation* (RFECV), que remove iterativamente as variáveis menos relevantes até determinar o número ótimo de características. Em paralelo, recorreu-se ao GridSearch para o ajuste simultâneo dos hiperparâmetros. E uma segunda abordagem com um conjunto de características fixo, que se manteve inalterado após a remoção por correlação (21 variáveis), com afinação de hiperparâmetros através do *GridSearchCV*. Estas abordagens estão descritas em maior detalhe na secção 3.3.4.

Na Tabela 4 estão apresentados então os resultados comparativos entre os quatro modelos avaliados utilizando a primeira abordagem de seleção de características.

Tabela 4 - Comparação dos resultados obtidos para os quatro métodos de classificação com a primeira abordagem de seleção de características por RFECV e otimização com *Gridsearch.*

Modelo	Nº de Características	Accuracy antes de Gridsearch (%)	Accuracy após Gridsearch (%)	Melhores Parâmetros
SVM	18	63.33	63.33	'C': 1, 'gamma': 'scale', 'kernel': 'linear'
Logistic Regression	12	70.00	70.00	'C': 0.1, 'penalty': 'l1'
Random Forest	19	76.67	76.67	'max_depth': None, 'min_samples_leaf': 1, 'n_estimators': 50
Gradient Boosting	15	80.00	73.33	'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 100

Para os modelos *Random Forest*, SVM e *Logistic Regression*, a *accuracy* no conjunto de teste manteve-se em 76.67%, 63.33% e 70.00%, respetivamente, ou seja, apesar de terem sido encontrados parâmetros diferentes do padrão, o desempenho foi o mesmo. No caso do *Gradient Boosting*, a *accuracy* diminuiu de 80.00% para 73.33%. Este decréscimo pode ser um sinal de *overfitting* nos dados de treino ou pode dever-se à elevada variabilidade existente num conjunto de teste de pequena dimensão como o utilizado, ou por escolha de hiperparâmetros demasiado conservadora (*learning rate* baixa, lr=0.01). Estudos prévios já demonstraram que a otimização excessiva de modelos de *boosting,* em conjuntos de dados de pequena dimensão, pode conduzir à perda de capacidade de generalização, em vez de a melhorar [63], [64].

Para além disso, observou-se a existência de uma grande diferença no número de características selecionadas para cada modelo. Na Tabela 5 estão presentes as características selecionadas por cada modelo que levaram à obtenção dos resultados apresentados na Tabela 4.

Tabela 5 – Características selecionadas por cada modelo para a abordagem 1 (RFECV e *GridSearch*)

Métricas	SVM	Logistic Regression	Random Forest	Gradient Boosting
1. Variância das distâncias dos pontos do contorno original ao centroide	X	X	X	
2. Variância das distâncias dos pontos do convex hull do contorno ao seu centroide	X		X	
3. Espessura média do traço do contorno			X	X
4. Desvio padrão da espessura do traço do contorno	**X**	**X**	**X**	**X**
5. Número de componentes conectados do contorno	**X**	**X**	**X**	**X**
6. Número total de componentes identificados e classificados como números	**X**	**X**	**X**	**X**
7. Erro angular médio entre a posição angular detetada de cada número e a sua posição angular ideal num relógio analógico	X	X	X	X
8. Proporção de transições angulares corretas entre números consecutivos			X	X
9. Desvio padrão circular dos ângulos das posições dos centroides dos números em relação ao centro do relógio	**X**	**X**	**X**	**X**
10. Distância média dos centroides dos números à fronteira do convex hull	**X**	**X**	**X**	**X**
11. Área total dos componentes classificados como números				
12. Área média dos componentes classificados como números			X	X
13. Desvio padrão dos componentes classificados como números	X		X	X
14. Espessura média dos traços dos números				
15. Desvio padrão entre as espessuras médias dos diferentes números	**X**	**X**	**X**	**X**
16. Média dos desvios padrão individuais da espessura de cada número	X		X	X
17. Média da conectividade (componentes) dos traços dos números	X		X	X
18. Média da extensão (área/componente) dos traços dos números				
19. Número total de componentes identificados como ponteiros	X	X	X	
20. Razão entre o comprimento do ponteiro mais longo (minutos) e o do ponteiro mais curto (horas), se pelo menos dois ponteiros forem detetados	X			
21. Área total dos componentes identificados como ponteiros				
22. Área média dos componentes identificados como ponteiros	X		X	X
23. Desvio padrão dos componentes identificados como ponteiros	X	X		
24. Espessura média dos traços dos ponteiros				
25. Média dos desvios padrão individuais da espessura de cada ponteiro	**X**	**X**	**X**	**X**
26. Média da conectividade (componentes) dos traços dos ponteiros	X	X	X	
27. Média da extensão (área/componente) dos traços dos ponteiros				

Ao comparar as que foram selecionadas simultaneamente pelos quatro modelos, verificou-se que existiam apenas oito em comum, as características 4, 5, 6, 7, 9, 10, 15 e 25 (destacadas a negrito na Tabela 5), sugerindo que são estas que capturam os

aspetos cruciais e mais robustos para esta análise e assim avaliar os TDR. As características 4 e 5 refletem a variabilidade na espessura da linha e a existência de fragmentação do contorno principal do relógio. As características 6, 7, 9, 10 e 15 estão intrinsecamente ligadas à capacidade de escrita, posicionamento e variabilidade do traço dos números. Já a característica 25 reflete a variabilidade da espessura do traço dos ponteiros.

O conjunto inclui, pelo menos, uma variável associada a cada um dos três componentes principais do relógio: círculo, ponteiros e números. Além disso, são medidas relevantes para quantificar aspetos visuais e estruturais importantes para o diagnóstico. Estas métricas exprimem capacidades que frequentemente ficam comprometidas quando há défices de memória, de movimento e de cognição — sintomas típicos em doentes com DN [5], [6].

Por outro lado, algumas características apresentaram uma seleção mais variável entre os modelos, refletindo as diferentes naturezas e sensibilidades dos algoritmos aos padrões nos dados. A característica 1 foi selecionada por três dos quatro modelos (SVM, Regressão Logística e *Random Forest*), enquanto a *característica* 2 foi selecionada apenas pelo SVM e *Random Forest*, apesar de expressarem informações semelhantes. Esta discrepância pode ser explicada pela forma diferenciada como cada algoritmo avalia a relevância das características durante o processo de seleção — através dos coeficientes na Regressão Logística, das margens no SVM ou dos critérios de divisão hierárquica nas árvores de decisão do *Random Forest* [36], [65].

Características ligadas à presença e ao traço dos ponteiros, como é o caso das características 19 e 26, mostraram alguma variabilidade na seleção entre modelos. Isto não contraria a sua relevância clínica, todavia, cada modelo pondera as variáveis de forma diferente e, quando existem outras características que explicam de modo mais geral o desempenho do desenho, o próprio modelo acaba por lhes dar mais peso. Nesses casos, diz-se que a contribuição das características dos ponteiros fica atenuada pelo processo de seleção do modelo, que privilegia preditores mais robustos ou redundantes.

Já a característica 8 foi selecionada apenas por modelos baseados em árvores de decisão (por exemplo, *Random Forest* e *Gradient Boosting*), uma vez que estes algoritmos são capazes de modelar interações complexas e não lineares entre variáveis através de divisões hierárquicas sucessivas. Em contraste, modelos lineares, como a Regressão Logística ou o SVM com *kernel* linear, tendem a não atribuir relevância a este tipo de variáveis, dado que assumem predominantemente relações lineares entre as características e a variável-alvo. Este resultado está em consonância com o

reportado na literatura, que destaca a capacidade superior dos métodos baseados em árvores para captar efeitos não lineares e interações de ordem superior [58], [66].

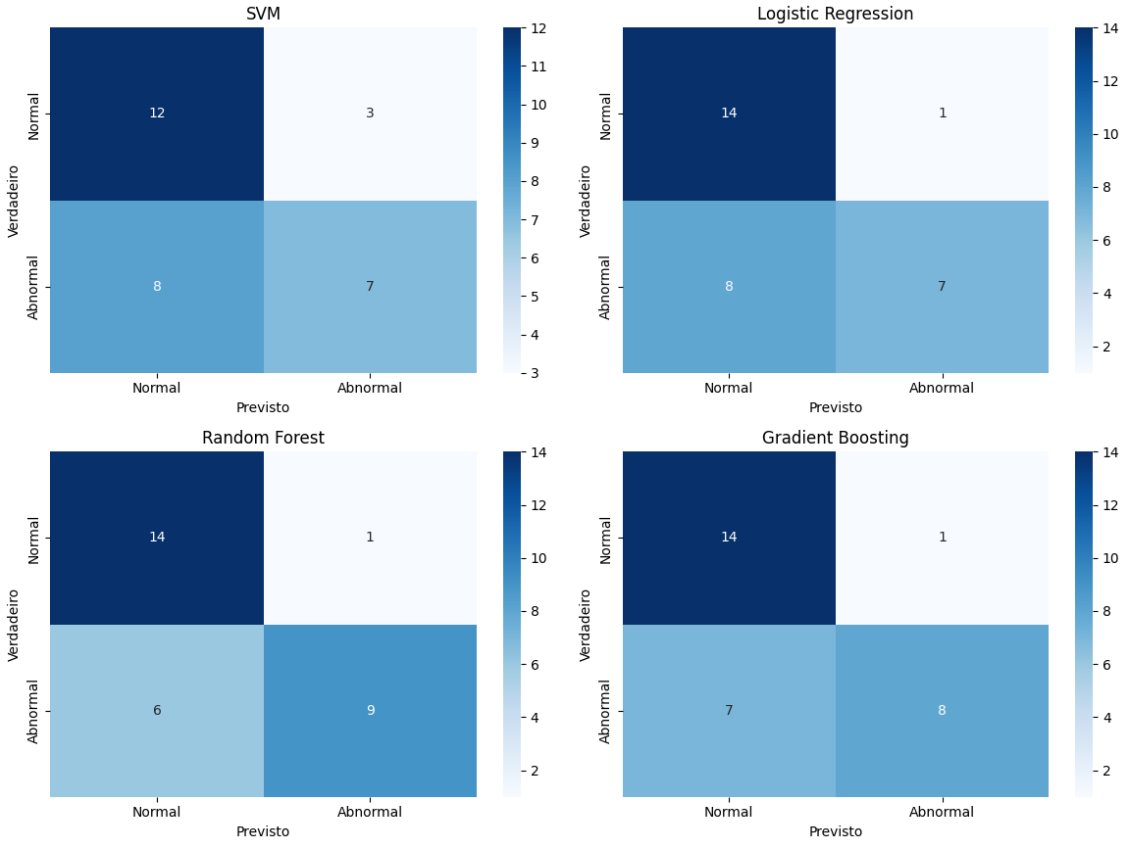


Figura 9 - Matrizes de confusão obtidas para os quatro métodos a partir da primeira abordagem de seleção de características por RFECV e otimização com *Gridsearch.*

Pela análise das matrizes de confusão apresentadas na Figura 9, verificou-se que os métodos *Logistic Regression, Random Forest* e *Gradient Boosting* são melhores a evitar falsos positivos (1 em 15 casos), o que espelha uma elevada precisão para a classe 'Abnormal'. O modelo *Random Forest* apresenta o melhor equilíbrio nesta primeira abordagem de todos os modelos, tendo também o menor número de falsos negativos (6 em 15 casos), isto é, é o modelo mais exato a identificar corretamente os casos 'Abnormal'. O SVM, para a presente abordagem, foi o modelo com menor *accuracy* e precisão, principalmente na identificação da classe *Abnormal*, apresentando o maior número de falsos negativos (8 em 15 casos). Este resultado pode ser explicado pelo facto de que, conforme descrito na literatura, os modelos SVM tendem a apresentar desempenho inferior em cenários com conjuntos de dados reduzidos e classes não balanceadas, situações em que o hiperplano de separação privilegia frequentemente a classe maioritária, resultando num maior número de falsos negativos na classe minoritária[67].

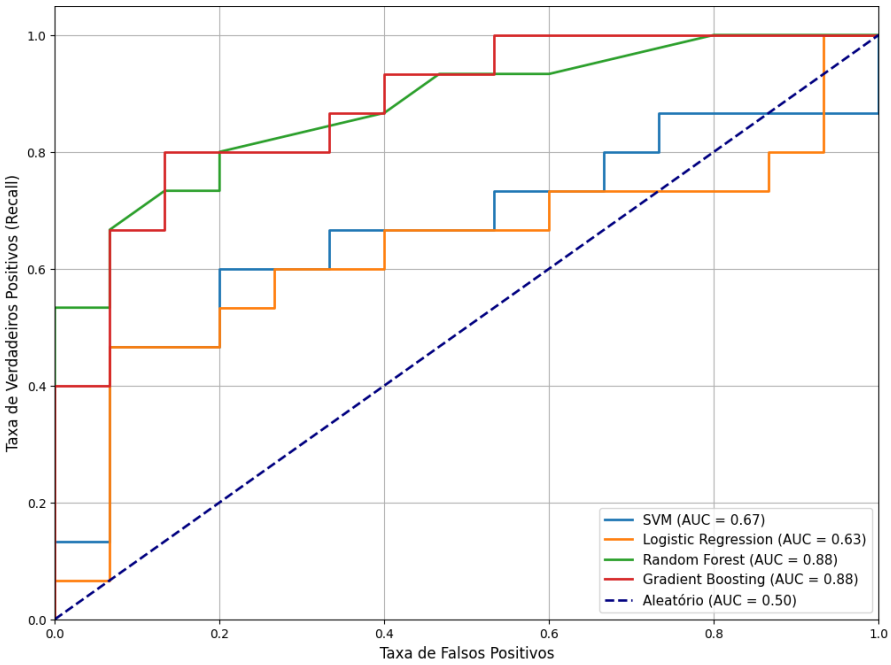


Figura 10 – Curvas ROC obtidas para os quatro métodos a partir da primeira abordagem de seleção de características por RFECV e otimização com *Gridsearch.*

As curvas ROC apresentada na Figura 10 vêm corroborar os que já tinham sido verificados anteriormente nas matrizes de confusão para esta primeira abordagem. Os modelos *Random Forest* e *Gradient Boosting* são os modelos com melhor desempenho discriminatório, ambos com um valor de AUC de 0.88, indicando uma boa capacidade de distinção entre casos 'Normal' e 'Abnormal'. Visualmente, as curvas ROC destes modelos localizam-se mais próximas do canto superior esquerdo do gráfico (TP=1 e FP=0), demonstrando uma maior taxa de verdadeiros positivos para uma dada taxa de falsos positivos em comparação com os outros modelos. Por outro lado, o SVM e o *Logistic Regression* apresentam um desempenho inferior, com um valor de AUC de 0.67 e 0.63, respetivamente, sugerindo que têm menor *accuracy* a separar as duas classes em comparação com os outros dois modelos.

Após esta análise dos resultados da primeira abordagem, decidiu-se acrescentar a característica 19 ao conjunto de 8 características consensuais entre os quatro modelos já referidas anteriormente para aplicar a segunda abordagem. Apesar de esta característica ter sido selecionada apenas por três dos quatro modelos e existirem outras cinco nas mesmas condições de seleção, esta escolha é justificada pelo facto de se ter considerado que extraía informação relevante dos TDR para a tarefa do diagnóstico para os especialistas e tornava o conjunto escolhido mais completo visto que este só continha uma outra característica relacionada com os ponteiros do relógio.

Na Tabela 6 estão apresentados os resultados obtidos para os mesmos quatro modelos a partir da segunda abordagem de seleção fixa das 9 características, realizada através

da análise dos resultados da primeira abordagem anteriormente discutida, e respetiva otimização para a métrica *accuracy*.

Tabela 6 - Comparação dos resultados obtidos para os quatro métodos com a segunda abordagem de seleção fixa de 9 características e otimização para a métrica accuracy.

Modelo	Accuracy (%)	F1-Score (%)	Hiperparâmetros
SVM	76.67	76.00	'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 100
Logistic Regression	70.00	68.27	'C': 1, 'gamma': 'scale', 'kernel': 'rbf'
Random Forest	70.00	68.27	'C': 0.1, 'penalty': 'l1'
Gradient Boosting	83.33	82.86	'max_depth': None, 'n_estimators': 50

Pela análise dos resultados obtidos na Tabela 6, o método de *Gradient Boosting* obteve o melhor resultado para ambas as métricas, *accuracy* e *F1-Score*, seguido do SVM. A Regressão Logística e o *Random Forest* apresentaram desempenho similar, mas inferior aos dois primeiros. Esta superioridade do modelo *Gradient Boosting* pode ser justificada pela sua capacidade intrínseca de modelar melhor relações complexas entre características, mesmo num conjunto limitado de dados. Por outro lado, o desempenho inferior dos restantes modelos pode ser atribuído às suas limitações em lidar com não linearidades, no caso da Regressão Logística, ou à necessidade de um conjunto de características mais extenso para atingir o seu potencial máximo, no caso do *Random Forest*.

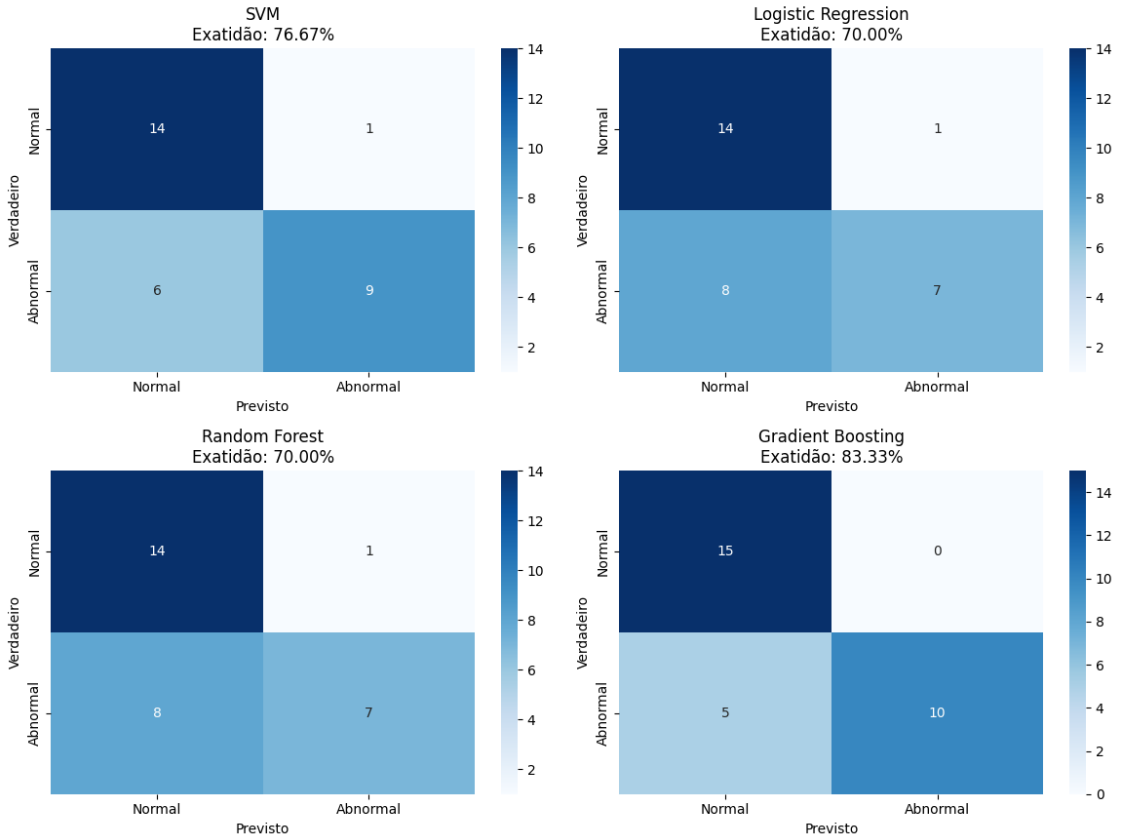


Figura 11 - Matrizes de confusão obtidas para os quatro métodos com a segunda abordagem de seleção fixa de 9 características e otimização para a métrica *accuracy*.

As matrizes de confusão apresentadas na Figura 11 confirmam o melhor desempenho do modelo *Gradient Boosting* nesta abordagem, sobretudo pela sua capacidade de reduzir falsos positivos (maior precisão). Já o SVM apresentou melhor equilíbrio entre falsos positivos e falsos negativos porque obteve, no conjunto de teste, a maior *accuracy* (*accuracy*) e o maior *F1-score* entre os modelos comparados. Em contraste, os modelos de Regressão Logística e *Random Forest* registaram valores de *F1-score* e *accuracy* inferiores, evidenciando maior assimetria entre os dois tipos de erro.

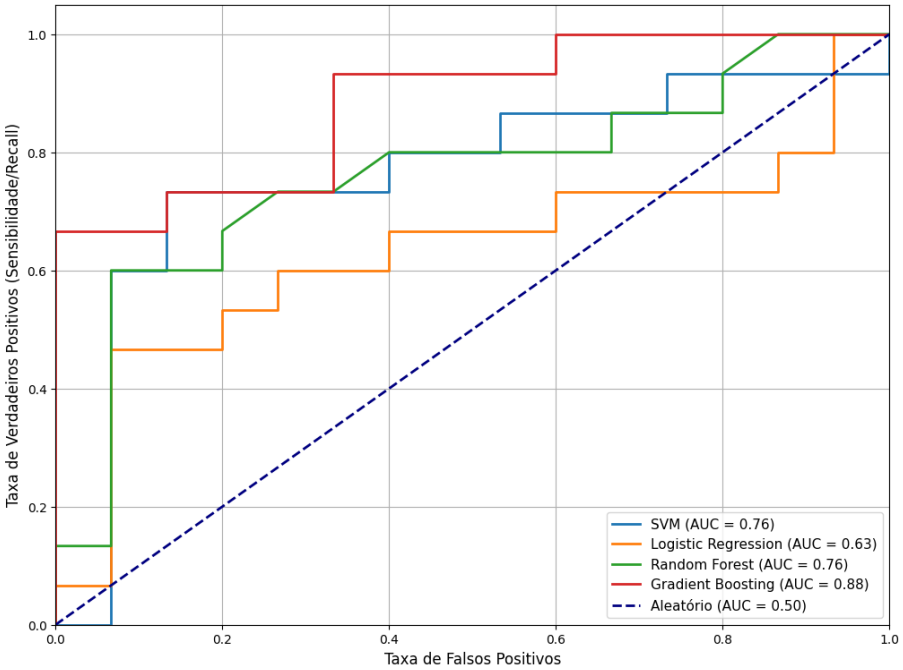


Figura 12 - Curvas ROC obtidas para os quatro métodos com a segunda abordagem de seleção fixa de 9 características e otimização para a métrica *accuracy*.

Pela Figura 12 verifica-se que o *Gradient Boosting* se destaca com um valor de AUC de 0.88, o que significa que este modelo tem uma probabilidade de 88% de classificar corretamente um caso 'Abnormal' selecionado aleatoriamente com uma pontuação de probabilidade mais alta do que um caso 'Normal' selecionado aleatoriamente. Este resultado vem confirmar a sua superioridade na capacidade de separar as duas classes nesta abordagem. Para além disso, a curva ROC do *Gradient Boosting*, ao estar mais próxima do canto superior esquerdo, demonstra que este modelo é capaz de atingir uma alta taxa de verdadeiros positivos, mantendo uma baixa taxa de falsos positivos ao longo de vários limiares de decisão. Os modelos SVM e o *Random Forest* têm ambos o segundo melhor valor de AUC (0.76), indicando uma capacidade discriminatória razoável. Quanta às respetivas curvas ROC, embora acima da linha aleatória, não se aproximam tanto do canto superior esquerdo quanto a do *Gradient Boosting*, refletindo um compromisso maior entre as taxas de verdadeiros positivos e falsos positivos. Por último, a Regressão Logística apresenta o valor mais baixo de AUC (0.63), sugerindo uma capacidade discriminatória mais limitada, e a sua curva ROC está mais próxima da linha diagonal aleatória, confirmando que a sua capacidade de separar as duas classes é limitada em comparação com os outros modelos nesta abordagem.

4.5 Limitações

Apesar dos resultados promissores, este estudo apresenta algumas limitações. A variabilidade dos desenhos, incluindo a sobreposição de traços, a presença de dígitos compostos e múltiplos ponteiros, dificultou a segmentação automática. A capacidade de generalização dos modelos foi também restringida pelo tamanho reduzido da amostra e

pela heterogeneidade gráfica. Embora o *Gradient Boosting* tenha alcançado melhor desempenho, outros modelos revelaram dificuldades em lidar com padrões não lineares, refletindo limitações no espaço de características disponível. Além disso, verificou-se menor desempenho na classe *Abnormal*, com *recall* moderado e um número significativo de falsos negativos, fenómeno comum em cenários com não balanceamento de classes e sobreposição de padrões.

4.6 Perspetivas Futuras

Como perspetivas de continuidade, destaca-se a necessidade de aumentar a dimensão e diversidade da base de dados, incorporando participantes de diferentes idades e níveis de comprometimento cognitivo, de modo a melhorar a robustez e a generalização dos modelos. A utilização de Redes Neuronais Convolucionais constitui uma via promissora para a extração automática de padrões complexos nos TDR, reduzindo a dependência de seleção manual de características e potencialmente elevando a *accuracy* e adaptabilidade do sistema. Adicionalmente, a realização de testes cegos (‘*blind test*’) com profissionais clínicos poderá validar a aplicabilidade prática do método proposto e identificar áreas de melhoria. Por fim, a integração de pipelines híbridos que combinem métricas clássicas com aprendizagem profunda poderá resultar numa ferramenta mais robusta, interpretável e escalável para apoio ao diagnóstico de doenças.

5. Conclusões

O trabalho desenvolvido teve como objetivo a conceção, implementação e avaliação de uma metodologia automática para análise do Teste do Desenho do Relógio, visando a identificação de padrões nos grafismos associados ao declínio cognitivo em Doenças Neurodenerativas. A abordagem incluiu etapas de segmentação de imagens, extração e seleção de características e classificação automática, com o intuito de determinar que métricas são mais relevantes e que modelos de aprendizagem têm melhor desempenho nesta tarefa de diagnóstico.

O *pipeline* criado demonstrou robustez na identificação do contorno principal do relógio, mesmo perante traços incompletos ou irregulares, enquanto a segmentação de componentes internos (números e ponteiros) se revelou mais desafiante devido à proximidade e sobreposição dos traços. A utilização de duas sequências de segmentação distintas permitiu melhorar os resultados, confirmando a viabilidade da abordagem, embora existam oportunidades de otimização, sobretudo no pré-processamento de imagens mais complexas.

A extração de características inicialmente contemplou 27 métricas, posteriormente reduzidas a 9, tendo em conta a sua relevância clínica e robustez computacional. Estas métricas cobrem de forma equilibrada os principais componentes do grafismo do Teste do Desenho do Relógio (contorno, números e ponteiros), refletindo aspetos críticos do desempenho cognitivo e motor, tais como a variabilidade da espessura do traço, fragmentação do contorno, irregularidade no posicionamento angular dos números e proporções entre ponteiros.

Na comparação entre modelos de classificação, o algoritmo *Gradient Boosting* foi o que apresentou melhores resultados, atingindo 83,33% de *accuracy*, 82,86% de *F1-score* e AUC de 0,88, demonstrando elevada capacidade discriminatória entre casos “Normal” de casos “Abnormal”. Estes resultados são consistentes com a literatura [58], [66], que aponta os métodos baseados em árvores como particularmente eficazes na modelação de relações complexas e não lineares entre variáveis. Para além deste desempenho, importa contextualizar o enquadramento científico do estudo no panorama nacional. A literatura portuguesa dedicada à avaliação cognitiva inclui diversos trabalhos que analisam o TDR em contexto clínico e psicométrico tradicional (Sousa et al., 2010; Apóstolo et al., 2017). Contudo, não foram identificadas investigações que recorram a processamento de imagem, extração automática de métricas ou modelos de *Machine Learning* aplicados especificamente ao TDR na população portuguesa. Esta lacuna evidencia que a análise deste teste tem permanecido essencialmente ancorada em métodos clínicos convencionais, sem exploração de abordagens computacionais que

permitam quantificação objetiva do grafismo. Assim, a metodologia desenvolvida neste estudo representa um contributo inovador, oferecendo uma alternativa automatizada e ajustada às especificidades demográficas e clínicas da população portuguesa.

Apesar de existirem abordagens recentes baseadas em *deep learning* aplicadas ao TDR, a dimensão reduzida da base de dados justificou a opção por métodos clássicos de *Machine Learning.* Trabalhos futuros poderão explorar modelos de *deep learning*, nomeadamente CNNs, desde que suportados por bases de dados de maior escala.

Em síntese, os resultados obtidos reforçam o potencial da abordagem desenvolvida para apoiar a padronização e a automatização da análise do Teste do Desenho do Relógio, constituindo um contributo relevante para o rastreio precoce de alterações cognitivas em contexto clínico. Apesar das limitações decorrentes da dimensão da amostra utilizada, este trabalho abre caminho para futuras investigações com bases de dados mais extensas e diversificadas, bem como para o desenvolvimento de ferramentas clínicas escaláveis e integráveis em sistemas de apoio à decisão médica.

Referências bibliográficas

[1] A. Martin Prince *et al.*, «World Alzheimer Report 2015 The Global Impact of Dementia an Analysis of Prevalence, Incidence, Cost and Trends». [Em linha]. Disponível em: www.alz.co.uk/worldreport2015corrections

[2] S. Chen, D. Stromer, H. A. Alabdalrahim, S. Schwab, M. Weih, e A. Maier, «Automatic dementia screening and scoring by applying deep learning on clock-drawing tests», *Sci Rep*, vol. 10, n. 1, Dez. 2020, doi: 10.1038/s41598-020-74710-9.

[3] W. Souillard-Mandar *et al.*, «Learning classification models of cognitive conditions from subtle behaviors in the digital Clock Drawing Test», *Mach Learn*, vol. 102, n. 3, pp. 393–441, Mar. 2016, doi: 10.1007/s10994-015-5529-5.

[4] M. Shah, A. Shandilya, K. Patel, M. Mehta, J. Sanghavi, e A. Pandya, «Neuropsychological detection and prediction using machine learning algorithms: a comprehensive review», *Intelligent Medicine*, vol. 4, n. 3, pp. 177–187, Ago. 2024, doi: 10.1016/J.IMED.2023.04.003.

[5] H. M. Gao e J. S. Hong, «Why neurodegenerative diseases are progressive: uncontrolled inflammation drives disease progression», *Trends Immunol*, vol. 29, n. 8, pp. 357–365, Ago. 2008, doi: 10.1016/j.it.2008.05.002.

[6] D. G. Gadhave *et al.*, «Neurodegenerative disorders: Mechanisms of degeneration and therapeutic approaches with their clinical relevance», *Ageing Res Rev*, vol. 99, p. 102357, Ago. 2024, doi: 10.1016/J.ARR.2024.102357.

[7] «Dementia». Acedido: 14 de Setembro de 2025. [Em linha]. Disponível em: https://www.who.int/news-room/fact-sheets/detail/dementia

[8] Q. Behfar, A. Ramirez Zuniga, e P. V. Martino-Adami, «Aging, Senescence, and Dementia», *Journal of Prevention of Alzheimer's Disease*, vol. 9, n. 3, pp. 523–531, Jul. 2022, doi: 10.14283/jpad.2022.42.

[9] G. Livingston *et al.*, «Dementia prevention, intervention, and care: 2020 report of the Lancet Commission», *The Lancet*, vol. 396, n. 10248, pp. 413–446, Ago. 2020, doi: 10.1016/S0140-6736(20)30367-6.

[10] «2023 Alzheimer's disease facts and figures», *Alzheimers Dement*, vol. 19, n. 4, pp. 1598–1695, Abr. 2023, doi: 10.1002/ALZ.13016.

[11] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, e M. Karagiannidou, «World Alzheimer Report 2016 Improving healthcare for people living with dementia coverage, Quallty and costs now and In the future», Acedido: 14 de Setembro de 2025. [Em linha]. Disponível em: www.daviddesigns.co.uk

[12] K. I. Shulman, «CLOCK-DRAWING: IS IT THE IDEAL COGNITIVE SCREENING TEST?», *Int J Geriatr Psychiatry*, pp. 548–561, 2000, doi: 10.1002/1099-1166(200006)15:6<548::AID-GPS242>3.0.CO;2-U.

[13] B. Spenciere, H. Alves, e H. Charchat-Fichman, «Scoring systems for the Clock Drawing Test: A historical review», *Dement Neuropsychol*, vol. 11, n. 1, p. 6, 2017, doi: 10.1590/1980-57642016DN11-010003.

[14] E. Hazan, F. Frankenburg, M. Brenkel, K. Shulman, e W. O. Library, «The test of time: a history of clock drawing», 2017, doi: 10.1002/gps.4731.

[15] «CDT-API-Network/README.md at main · cccnlab/CDT-API-Network · GitHub». Acedido: 29 de Setembro de 2025. [Em linha]. Disponível em: https://github.com/cccnlab/CDT-API-Network/blob/main/README.md

[16] «History of the Clock Drawing Test and the Linus Health Platform». Acedido: 28 de Setembro de 2025. [Em linha]. Disponível em: https://linushealth.com/learn/history-of-the-clock-drawing-test?utm_source=chatgpt.com

[17] L. Schmitt, «Clock Drawing», em *Encyclopedia of Autism Spectrum Disorders*, I. B. Harris e F. R. Volkmar, Eds., Springer, New York, NY, 2013, pp. 665–666. doi: 10.1007/978-1-4419-1698-3_337.

[18] I. Aprahamian, J. Eduardo Martinelli, A. Liberalesso Neri, e M. Sanches Yassuda, «The Clock Drawing Test: A review of its accuracy in screening for dementia», *Dement Neuropsychol*, vol. 3, n. 2, pp. 74–80, 2009, doi: 10.1590/S1980-57642009DN30200002.

[19] E. Pinto e R. Peters, «Literature review of the Clock Drawing Test as a tool for cognitive screening», *Dement Geriatr Cogn Disord*, vol. 27, n. 3, pp. 201–213, Mar. 2009, doi: 10.1159/000203344.

[20] L. Ehreke, M. Luppa, H. H. König, e S. G. Riedel-Heller, «Is the clock drawing test a screening tool for the diagnosis of mild cognitive impairment? A systematic review», *Int Psychogeriatr*, vol. 22, n. 1, pp. 56–63, Fev. 2010, doi: 10.1017/S1041610209990676.

[21] M. Freedman, L. Leach, E. Kaplan, G. Winocur, K. Shulman, e D. C. Delis, *Clock drawing : a neuropsychological analysis*, 1.ª ed. Oxford University Press, 1994.

[22] E. Seigerschmidt, E. Mösch, M. Siemen, H. Förstl, e H. Bickel, «The clock drawing test and questionable dementia: reliability and validity», *Int J Geriatr Psychiatry*, vol. 17, n. 11, pp. 1048–1054, Nov. 2002, doi: 10.1002/GPS.747.

[23] M. L. Cera, T. S. C. Minett, e K. Z. Ortiz, «Analysis of error type and frequency in apraxia of speech among Portuguese speakers», *Dement Neuropsychol*, vol. 4, n. 2, p. 98, 2010, doi: 10.1590/S1980-57642010DN40200004.

[24] R. Binaco *et al.*, «Machine Learning Analysis of Digital Clock Drawing Test Performance for Differential Classification of Mild Cognitive Impairment Subtypes Versus Alzheimer's Disease», *J Int Neuropsychol Soc*, vol. 26, n. 7, pp. 690–700, Ago. 2020, doi: 10.1017/S1355617720000144.

[25] W. Souillard-Mandar, R. Davis, C. Rudin, R. Au, D. L. Penney, e D. L. P. Org, «Interpretable Machine Learning Models for the Digital Clock Drawing Test», Jun. 2016, Acedido: 27 de Setembro de 2025. [Em linha]. Disponível em: https://arxiv.org/pdf/1606.07163

[26] S. Lazarova, D. Grigorova, e D. Petrova-Antonova, «Detection of Alzheimer's Disease Using Logistic Regression and Clock Drawing Errors», *Brain Sciences 2023, Vol. 13, Page 1139*, vol. 13, n. 8, p. 1139, Jul. 2023, doi: 10.3390/BRAINSCI13081139.

[27] S. Lazarova, D. Grigorova, e D. Petrova-Antonova, «Detection of Alzheimer's Disease Using Logistic Regression and Clock Drawing Errors», *Brain Sci*, vol. 13, n. 8, Ago. 2023, doi: 10.3390/BRAINSCI13081139.

[28] K. Sato, Y. Niimi, T. Mano, A. Iwata, e T. Iwatsubo, «Automated Evaluation of Conventional Clock-Drawing Test Using Deep Neural Network: Potential as a Mass Screening Tool to Detect Individuals With Cognitive Decline», *Front Neurol*, vol. 13, p. 896403, Mai. 2022, doi: 10.3389/FNEUR.2022.896403/BIBTEX.

[29] Z. Harbi, Y. Hicks, R. Setchi, e A. Bayer, «Segmentation of Clock Drawings Based on Spatial and Temporal Features», *Procedia Comput Sci*, vol. 60, n. 1, pp. 1640–1648, Jan. 2015, doi: 10.1016/J.PROCS.2015.08.274.

[30] Z. Harbi, Y. Hicks, e R. Setchi, «Clock Drawing Test Digit Recognition Using Static and Dynamic Features», *Procedia Comput Sci*, vol. 96, pp. 1221–1230, 2016, doi: 10.1016/J.PROCS.2016.08.166.

[31] A. Lapušinskij, I. Suzdalev, N. Goranin, J. Janulevičius, S. Ramanauskaitė, e G. Stankūnavičius, «The Application of Hough Transform and Canny Edge Detector Methods for the Visual Detection of Cumuliform Clouds», *Sensors (Basel)*, vol. 21, n. 17, p. 5821, Set. 2021, doi: 10.3390/S21175821.

[32] L. Han *et al.*, «Circle Detection with Adaptive Parameterization: A Bottom-Up Approach», *Sensors (Basel)*, vol. 25, n. 8, p. 2552, Abr. 2025, doi: 10.3390/S25082552.

[33] E. F. Matusz *et al.*, «Dissociating Statistically Determined Normal Cognitive Abilities and Mild Cognitive Impairment Subtypes with DCTclock», *J Int Neuropsychol Soc*, vol. 29, n. 2, p. 148, Fev. 2022, doi: 10.1017/S1355617722000091.

[34] A. Masuo, Y. Ito, T. Kanaiwa, K. Naito, T. Sakuma, e S. Kato, «Dementia Screening Based on SVM Using Qualitative Drawing Error of Clock Drawing Test», *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2022, pp. 4484–4487, 2022, doi: 10.1109/EMBC48229.2022.9871889.

[35] C. M. Bishop, *Pattern Recognition and Machine Learning | SpringerLink*, 1.ª ed. New York: Springer New York, 2006.

[36] M. Kuhn e K. Johnson, *Applied predictive modeling*, 1.ª ed. New York: Springer New York, 2013. doi: 10.1007/978-1-4614-6849-3/COVER.

[37] J. Chen *et al.*, «An intelligent screener for mild cognitive impairment via integrated eye-tracking and the digital clock drawing test», *J Alzheimers Dis*, Jun. 2025, doi: 10.1177/13872877251350101.

[38] R. Qasrawi *et al.*, «Hybrid ensemble deep learning model for advancing breast cancer detection and classification in clinical applications», *Heliyon*, vol. 10, n. 19, p. e38374, Out. 2024, doi: 10.1016/J.HELIYON.2024.E38374.

[39] Y. ; Zhang, J. ; Liu, W. Shen, Y. Zhang, J. Liu, e W. Shen, «A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications», *Applied Sciences 2022, Vol. 12, Page 8654*, vol. 12, n. 17, p. 8654, Ago. 2022, doi: 10.3390/APP12178654.

[40] S. M. Lundberg e S. I. Lee, «A Unified Approach to Interpreting Model Predictions», *Adv Neural Inf Process Syst*, vol. 2017-December, pp. 4766–4775, Mai. 2017, Acedido: 27 de Setembro de 2025. [Em linha]. Disponível em: https://arxiv.org/pdf/1705.07874

[41] Ian Goodfellow, Aaron Courville, e Yoshua Bengio, *Deep Learning*. The MIT Press, 2016.

[42] Y. Lecun, Y. Bengio, e G. Hinton, «Deep learning», *Nature*, vol. 521, n. 7553, pp. 436–444, Mai. 2015, doi: 10.1038/NATURE14539;SUBJMETA.

[43] I. Park e U. Lee, «Automatic, Qualitative Scoring of the Clock Drawing Test (CDT) Based on U-Net, CNN and Mobile Sensor Data», *Sensors (Basel)*, vol. 21, n. 15, p. 5239, Ago. 2021, doi: 10.3390/S21155239.

[44] S. Bandyopadhyay, J. Wittmayer, D. J. Libon, P. Tighe, C. Price, e P. Rashidi, «Explainable semi-supervised deep learning shows that dementia is associated with small, avocado-shaped clocks with irregularly placed hands», *Sci Rep*, vol. 13, n. 1, pp. 1–12, Dez. 2023, doi: 10.1038/S41598-023-34518-9;SUBJMETA.

[45] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, e D. Batra, «Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization», *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 618–626, Dez. 2017, doi: 10.1109/ICCV.2017.74.

[46] R. Heyrani *et al.*, «Limits on using the clock drawing test as a measure to evaluate patients with neurological disorders», *BMC Neurol*, vol. 22, n. 1, p. 509, Dez. 2022, doi: 10.1186/S12883-022-03035-Z.

[47] I. Santana, D. Duro, S. Freitas, L. Alves, e M. R. Simões, «The Clock Drawing Test: Portuguese norms, by age and education, for three different scoring systems», *Arch Clin Neuropsychol*, vol. 28, n. 4, pp. 375–387, 2013, doi: 10.1093/ARCLIN/ACT016.

[48] A. D. International, «World Alzheimer Report 2023: Reducing Dementia Risk: Never too early, never too late», Set. 2023. Acedido: 28 de Setembro de 2025. [Em linha]. Disponível em: https://www.alzint.org/resource/world-alzheimer-report-2023/

[49] «Fact-Sheet - Associação Alzheimer Portugal». Acedido: 28 de Setembro de 2025. [Em linha]. Disponível em: https://alzheimerportugal.org/fact-sheet/?utm_source=chatgpt.com

[50] C. Flint *et al.*, «Systematic misestimation of machine learning performance in neuroimaging studies of depression», *Neuropsychopharmacology*, vol. 46, n. 8, pp. 1510–1517, Jul. 2021, doi: 10.1038/S41386-021-01020-7;TECHMETA.

[51] D. Rajput, W. J. Wang, e C. C. Chen, «Evaluation of a decided sample size in machine learning applications», *BMC Bioinformatics*, vol. 24, n. 1, pp. 1–17, Dez. 2023, doi: 10.1186/S12859-023-05156-9/FIGURES/5.

[52] S. Suzuki e K. A. be, «Topological structural analysis of digitized binary images by border following», *Comput Vis Graph Image Process*, vol. 30, n. 1, pp. 32–46, 1985, doi: 10.1016/0734-189X(85)90016-7.

[53] R. O. Duda e P. E. Hart, «Use of the Hough transformation to detect lines and curves in pictures», *Commun ACM*, vol. 15, n. 1, pp. 11–15, Jan. 1972, doi: 10.1145/361237.361242.

[54] K. He, X. Zhang, S. Ren, e J. Sun, «Deep Residual Learning for Image Recognition», *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dez. 2015, doi: 10.1109/CVPR.2016.90.

[55] G. K. Cohen, S. Afshar, J. Tapson, e A. van Schaik, «EMNIST: an extension of MNIST to handwritten letters», *Arxiv preprint*, Fev. 2017, Acedido: 28 de Setembro de 2025. [Em linha]. Disponível em: https://arxiv.org/pdf/1702.05373

[56] D. R. Cox, «The Regression Analysis of Binary Sequences», *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, n. 2, pp. 215–232, Jul. 1958, doi: 10.1111/J.2517-6161.1958.TB00292.X.

[57] J. H. Friedman, «Greedy function approximation: A gradient boosting machine.», *https://doi.org/10.1214/aos/1013203451*, vol. 29, n. 5, pp. 1189–1232, Out. 2001, doi: 10.1214/AOS/1013203451.

[58] L. Breiman, «Random forests», em *Machine Learning*, vol. 45, n. 1, Springer, 2001, pp. 5–32. doi: 10.1023/A:1010933404324/METRICS.

[59] C. Cortes, V. Vapnik, e L. Saitta, «Support-vector networks», em *Machine Learning 1995 20:3*, vol. 20, n. 3, Springer, 1995, pp. 273–297. doi: 10.1007/BF00994018.

[60] K. Gupta, «Digit Recognition Using Convolution Neural Network», n. June, pp. 2601–2603, Abr. 2020, Acedido: 29 de Setembro de 2025. [Em linha]. Disponível em: https://arxiv.org/pdf/2004.00331

[61] R. M. O. Cruz, G. D. C. Cavalcanti, e T. I. Ren, «Handwritten Digit Recognition Using Multiple Feature Extraction Techniques and Classifier Ensemble», 2010.

[62] A. Boukharouba e A. Bennia, «Novel feature extraction technique for the recognition of handwritten digits», *Applied Computing and Informatics*, vol. 13, n. 1, pp. 19–26, Jan. 2017, doi: 10.1016/J.ACI.2015.05.001.

[63] Y. Zhang e A. Haghani, «A gradient boosting method to improve travel time prediction», *Transp Res Part C Emerg Technol*, vol. 58, pp. 308–324, Set. 2015, doi: 10.1016/J.TRC.2015.02.019.

[64] T. Chen e C. Guestrin, «XGBoost: A scalable tree boosting system», *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Ago. 2016, doi: 10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4.

[65] GuyonIsabelle e ElisseeffAndré, «An introduction to variable and feature selection», *The Journal of Machine Learning Research*, Mar. 2003, doi: 10.5555/944919.944968.

[66] T. Hastie, R. Tibshirani, e J. Friedman, «The Elements of Statistical Learning», 2009, doi: 10.1007/978-0-387-84858-7.

[67] R. Akbani, S. Kwek, e N. Japkowicz, «Applying support vector machines to imbalanced datasets», *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, vol. 3201, pp. 39–50, 2004, doi: 10.1007/978-3-540-30115-8_7.