INSTITUTO POLITÉCNICO DE SANTARÉM
Escola Superior de Gestão e Tecnologia de Santarém

Classificação de Perigos Rodoviários com Recurso a Dados de Acelerómetros: Um Estudo Comparativo entre Abordagens de Aprendizagem Automática Clássica e Aprendizagem Profunda

Dissertação

Mestrado em Informática Aplicada

Joana Barreira Pardal Moutinho

Orientação:

Prof. Filipe Madeira
Prof. Ivan Miguel Serrano Pires
Prof. Fernando José da Silva Velez

Junho, 2026

POLITÉCNICO DE SANTARÉM

Declaração de Autoria e Originalidade

Eu, Joana Barreira Pardal Moutinho, estudante nº 230001650 declaro que sou a única autora da dissertação, intitulada 'Classificação de Perigos Rodoviários com Recurso a Dados de Acelerómetros: Um Estudo Comparativo entre Abordagens de aprendizagem Automática Clássica e Aprendizagem Profunda' apresentado(a) para obtenção do grau de Mestre em Informática Aplicada pela Escola Superior de Gestão e Tecnologias de Santarém.

Declaro ainda que identifiquei de uma forma clara e citei corretamente trabalhos de outros autores que tenham sido utilizados neste trabalho; no caso de ter utilizado frases retiradas de trabalhos de outros autores, referenciei-as devidamente ou, se as redigi com palavras diferentes, indiquei o trabalho original de onde foram adaptadas. Assim, declaro que não há qualquer plágio (apropriação indevida da obra intelectual de outra pessoa, assumindo, implícita ou explicitamente, a autoria da mesma, ainda que por omissão) no documento entregue e que reconheço que tal prática poderia resultar em sanções disciplinares e legais.

Por fim, declaro que este trabalho, em parte ou no todo, não foi previamente submetido para a mesma finalidade.

Tenho consciência que o plágio ou a cópia pode determinar a anulação do grau atribuído, bem como, originar responsabilidade criminal, civil e disciplinar. Constitui igualmente uma grave violação da ética académica.

Declaração sobre Utilização de Ferramentas de Inteligência Artificial

No âmbito da elaboração desta dissertação, foram utilizadas ferramentas de inteligência artificial generativa como apoio a tarefas específicas, nomeadamente: revisão linguística do inglês do resumo (*Abstract*); apoio na depuração de código Python utilizado nas análises; e organização e formatação das referências bibliográficas. A autora declara assumir total responsabilidade pelo conteúdo científico, pelas análises realizadas e pelas conclusões apresentadas neste trabalho, independentemente do recurso a estas ferramentas.

Dedicatória

À minha família, pelo amor, apoio e compreensão ao longo deste percurso.
Aos que estiveram sempre presentes nos momentos mais difíceis, que me incentivaram a continuar e acreditaram em mim quando mais precisei.
Esta conquista é também vossa.

Agradecimentos

A realização deste trabalho de projeto representa o culminar de um percurso académico exigente, marcado por aprendizagem, dedicação e superação. Assim, gostaria de expressar o meu sincero agradecimento a todos os que contribuíram, direta ou indiretamente, para a concretização deste trabalho.

Em primeiro lugar, agradeço aos meus orientadores, Professor Filipe Madeira, Professor Ivan Miguel Serrano Pires e Professor Fernando José da Silva Velez, pela orientação, disponibilidade, acompanhamento e contributos prestados ao longo do desenvolvimento deste projeto. As suas sugestões e recomendações foram fundamentais para a estruturação e melhoria deste trabalho.

Agradeço também ao Instituto Politécnico de Santarém e à Escola Superior de Gestão e Tecnologia de Santarém pela formação proporcionada ao longo do mestrado, bem como pelas condições oferecidas para o desenvolvimento académico e científico.

À minha família, agradeço pelo apoio incondicional, pela paciência e pela motivação constante, especialmente nos momentos de maior exigência. O seu incentivo foi essencial para concluir esta etapa.

Por fim, agradeço aos colegas e amigos que, de alguma forma, acompanharam este percurso, partilhando conhecimento, apoio e palavras de encorajamento.

Acrónimos/Siglas

CNN – Convolutional Neural Networks
DBBE – Data Balanced Bagging Ensemble
GAN – Generative Adversarial Network
GPS – Global Positioning System
IC – Intervalos de Confiança
IoT – Internet of Things
KNN – K-Nearest Neighbours
LSTM – Long Short-Term Memory
MLP – Multi-Layer Perceptron
PVS – Passive Vehicular Sensors
RBF – Radial Basis Function
RMS – Root Mean Square
SSD – Solid-State Drive
SVM – Support Vector Machine
YOLO – You Only Look Once

Resumo

A classificação automática de perigos rodoviários, como buracos e lombas, com base em dados de acelerómetro constitui uma solução de baixo custo para a monitorização de infraestruturas viárias. Comparam-se abordagens de aprendizagem automática clássica e de aprendizagem profunda, recorrendo a três conjuntos de dados públicos integrados num fluxo unificado de pré-processamento e extração de características. Foram avaliados nove modelos clássicos e três arquiteturas profundas, com métricas sensíveis ao desequilíbrio entre classes. A 1D CNN alcançou o melhor desempenho global (F1 macro = 0.5108; exatidão balanceada = 0.6143), revelando maior capacidade na deteção de buracos, enquanto os modelos clássicos baseados em ensembles de árvores se mostraram competitivos na deteção de lombas, com custo computacional inferior. O desequilíbrio entre classes manteve-se um desafio significativo. A aprendizagem profunda demonstra vantagens na extração automática de padrões, mas os modelos clássicos continuam relevantes onde a interpretabilidade e a eficiência computacional são prioritárias.

Palavras-chave: perigos rodoviários, acelerómetro, aprendizagem automática, aprendizagem profunda, classificação multiclasse, buracos, lombas.

Road Hazard Classification Using Accelerometer Data: A Comparative Study between Classical Machine Learning and Deep Learning Approaches

Abstract

The automatic classification of road hazards, such as potholes and speed bumps, based on accelerometer data offers a low-cost solution for monitoring road infrastructure. Classical machine learning and deep learning approaches are compared, using three public datasets integrated through a unified pipeline of pre-processing, segmentation and feature extraction. Nine classical models and three deep learning architectures were evaluated, using metrics suited to class imbalance. The 1D CNN achieved the best overall performance (macro F1-score = 0.5108; balanced accuracy = 0.6143), showing greater ability to detect potholes, while classical tree-based ensemble models proved competitive in speed bump detection, with lower computational cost. Class imbalance remained a significant challenge throughout. Deep learning thus demonstrates advantages in automatically extracting patterns from raw data, while classical models remain relevant in scenarios where interpretability and computational efficiency are priorities.

***Key-words**: road hazards, accelerometer, machine learning, deep learning, multiclass classification, potholes, speed bumps.*

Índice

Índice

Lista de figuras

Lista de tabelas

Capítulo 1 – Introdução

1.1. Enquadramento do Estudo

A infraestrutura rodoviária desempenha um papel central na segurança dos transportes, na mobilidade e no desenvolvimento económico [1], [2]. Estradas bem conservadas promovem a circulação eficiente de pessoas e bens, enquanto superfícies rodoviárias danificadas ou irregulares podem aumentar os custos de manutenção dos veículos, reduzir o conforto da condução, contribuir para o congestionamento do tráfego e criar riscos de segurança para os utilizadores da via [1], [2]. Entre as irregularidades mais comuns da superfície rodoviária encontram-se os buracos e as lombas [2], [3]. Embora as lombas sejam estruturas intencionais de acalmia de tráfego e os buracos sejam defeitos não planeados do pavimento, ambos geram vibrações significativas no veículo, que podem ser detetadas através de sensores de movimento [4].

Os métodos tradicionais de inspeção rodoviária dependem, geralmente, de levantamentos manuais, veículos especializados de inspeção ou relatórios periódicos de manutenção [5], [6]. Embora estas abordagens possam fornecer informação fiável, são frequentemente dispendiosas, intensivas em mão de obra e difíceis de escalar para grandes redes rodoviárias. Consequentemente, tem-se verificado um interesse crescente em sistemas automatizados e baseados em sensores para a monitorização rodoviária. Os avanços recentes nas tecnologias de deteção móvel tornaram possível a utilização de acelerómetros integrados em *smartphones* ou veículos para recolher dados sobre as condições rodoviárias durante a condução normal [7], [8], [9]. Esta abordagem permite uma monitorização da infraestrutura rodoviária de baixo custo e potencialmente colaborativa [5], [6].

A deteção de perigos rodoviários baseada em acelerómetros assenta no pressuposto de que diferentes condições da estrada produzem padrões de vibração distinguíveis [3], [4], [10]. Por exemplo, os buracos podem gerar alterações abruptas na aceleração vertical, seguidas de efeitos de ressalto, enquanto as lombas podem produzir padrões de pico mais regulares e característicos [2], [11]. Contudo, a classificação precisa continua a ser desafiante, uma vez que os sinais de acelerómetros são afetados por diversos fatores externos, incluindo a velocidade do veículo, as características da suspensão, o material da estrada, a posição do sensor, a orientação do smartphone e as vibrações de fundo da condução [1], [12].

A aprendizagem automática constitui uma solução promissora para este problema, permitindo que os modelos aprendam padrões a partir de dados de acelerómetros etiquetados [3], [10], [13]. Os métodos clássicos de aprendizagem automática, como *Random Forest*, Máquinas de Vetores de Suporte, *Gradient Boosting*, Regressão Logística e K-Vizinhos Mais Próximos, dependem normalmente de características estatísticas manualmente definidas, extraídas de janelas segmentadas de aceleração [14], [15], [16], [17]. Mais recentemente, modelos de aprendizagem profunda, como Redes Neuronais Convolucionais unidimensionais e redes de Memória de Curto e Longo Prazo, têm sido utilizados para aprender representações discriminativas diretamente a partir de dados brutos de séries temporais [12], [18], [19], [20].

Apesar destes avanços, a classificação de perigos rodoviários continua a ser um problema de investigação em aberto, particularmente quando os modelos são avaliados em condições realistas [1], [3], [4], [10]. Os conjuntos de dados reais de monitorização rodoviária são, geralmente, altamente desequilibrados, uma vez que os segmentos rodoviários normais ocorrem com muito maior frequência do que os eventos perigosos [21], [22]. Nestes casos, um modelo pode alcançar uma elevada exatidão global, falhando, contudo, na deteção de eventos raros, mas importantes, como os buracos [23]. Assim, a avaliação de sistemas de classificação de perigos rodoviários exige métricas que vão além da exatidão, incluindo a pontuação F1 macro, a exatidão balanceada, a precisão e a revocação por classe, bem como a análise da matriz de confusão.

1.2. Problema de Investigação

Embora a deteção de perigos rodoviários baseada em acelerómetros tenha sido amplamente estudada, persistem várias limitações na literatura existente [1], [3], [4], [10]. Em primeiro lugar, muitos estudos centram-se na classificação binária, como a distinção entre condições rodoviárias normais e perigosas, em vez de abordarem a classificação multiclasse envolvendo diferentes tipos de perigos [19]. Na monitorização rodoviária prática, é importante não apenas detetar a existência de uma anomalia na estrada, mas também distinguir entre diferentes tipos de eventos, como buracos e lombas [2], [11].

Em segundo lugar, muitos estudos existentes avaliam apenas um número limitado de modelos ou utilizam fluxos de pré-processamento, conjuntos de dados e métricas de

avaliação inconsistentes [3], [12], [13], [18]. Isto dificulta a determinação de qual abordagem, aprendizagem automática clássica ou aprendizagem profunda, proporciona melhor desempenho na classificação de perigos rodoviários. Uma comparação justa exige que os diferentes modelos sejam treinados e testados sob as mesmas condições de preparação dos dados, segmentação, representação de características e avaliação [15], [24], [25].

Em terceiro lugar, o acentuado desequilíbrio entre classes permanece um desafio significativo [21], [22], [23]. No conjunto de dados utilizado nesta dissertação, as janelas correspondentes à estrada normal representam a classe dominante, enquanto as janelas de buracos e lombas são muito menos frequentes. O conjunto de teste combinado contém 47 706 janelas, das quais 45 911 correspondem à estrada normal, 980 a buracos e 815 a lombas. Esta distribuição reflete condições realistas de condução, mas torna difícil a deteção das classes minoritárias de perigo.

Por fim, é necessário considerar não apenas o desempenho de classificação, mas também a eficiência computacional e a viabilidade de implementação prática [5], [9], [26]. Os modelos destinados à monitorização rodoviária podem ter de operar em *smartphones*, sistemas embebidos ou dispositivos de computação periférica com recursos computacionais limitados. Assim, o melhor modelo não é necessariamente aquele que apresenta a maior exatidão, mas aquele que proporciona um equilíbrio adequado entre desempenho na deteção de perigos, robustez, interpretabilidade e custo computacional [6], [15], [17].

1.3. Objetivo do Estudo

O objetivo desta dissertação é investigar a classificação de perigos rodoviários com recurso a dados de acelerómetro, comparando abordagens de aprendizagem automática clássica e de aprendizagem profunda sob um enquadramento experimental unificado [3], [13], [12], [18]. O estudo centra-se na classificação das condições rodoviárias em três categorias: superfície rodoviária normal, buraco e lomba [2], [11].

1.4. Objetivos de Investigação

Os objetivos específicos deste estudo são os seguintes:

- Integrar e pré-processar múltiplos conjuntos de dados de acelerómetros publicamente disponíveis para a classificação de perigos rodoviários [27], [28], [29].
- Transformar sinais contínuos de acelerómetros em janelas de comprimento fixo adequadas à aprendizagem supervisionada [24], [25].
- Extrair características estatísticas e baseadas no sinal a partir dos dados de acelerómetros para modelos clássicos de aprendizagem automática [3], [14], [30].
- Treinar e avaliar um conjunto de modelos clássicos de aprendizagem automática para classificação multiclasse de perigos rodoviários [15], [16], [17].
- Treinar e avaliar modelos de aprendizagem profunda utilizando sequências brutas de acelerómetros segmentadas em janelas [18], [19], [20].
- Comparar modelos clássicos e modelos de aprendizagem profunda utilizando métricas sensíveis ao desequilíbrio, como a pontuação F1 macro, a exatidão balanceada, a revocação por classe e matrizes de confusão [21], [22], [23].
- Analisar a importância das características nos modelos clássicos, de modo a identificar as características derivadas dos acelerómetros mais informativas [3], [14], [30].
- Avaliar o compromisso entre desempenho preditivo e eficiência computacional para aplicações práticas de monitorização rodoviária [5], [9], [26].

1.5. Questões de Investigação

Esta dissertação é orientada pelas seguintes questões de investigação:

- **Q1:** Com que eficácia podem os dados de acelerómetros ser utilizados para classificar condições rodoviárias nas categorias de superfície rodoviária normal, buraco e lomba [3], [2], [4]?
- **Q2:** Qual modelo clássico de aprendizagem automática proporciona o melhor desempenho na classificação de perigos rodoviários utilizando características de acelerómetros manualmente definidas [14], [15], [17]?

- **Q3**: Os modelos de aprendizagem profunda superam os modelos clássicos de aprendizagem automática na classificação de perigos rodoviários a partir de sequências brutas de acelerómetros [12], [18], [19]?
- **Q4:** De que forma o acentuado desequilíbrio entre classes afeta a deteção de classes minoritárias de perigo, como buracos e lombas [21], [22], [23]?
- **Q5:** Quais características derivadas dos acelerómetros são mais importantes para distinguir perigos rodoviários de condições rodoviárias normais [3], [14], [30]?
- **Q6:** Que compromissos existem entre desempenho de classificação, eficiência computacional e adequação à implementação prática [5], [9], [26]?

1.6. Relevância do Estudo

Este estudo contribui para a investigação em sistemas de transporte inteligentes ao proporcionar uma comparação sistemática de abordagens de aprendizagem automática para a classificação de perigos rodoviários baseada em acelerómetros [3], [5], [6]. A sua relevância manifesta-se em várias dimensões.

Em primeiro lugar, o estudo aborda um problema prático de segurança rodoviária e manutenção da infraestrutura [2], [5], [6]. A deteção automatizada de perigos rodoviários pode apoiar a identificação mais rápida de segmentos rodoviários danificados e auxiliar as autoridades de transporte na priorização das atividades de manutenção [2], [5], [6].

Em segundo lugar, o estudo utiliza múltiplos conjuntos de dados públicos, o que melhora a reprodutibilidade e proporciona uma avaliação mais realista do que estudos baseados num único conjunto de dados controlado [27], [28], [29]. Os conjuntos de dados incluem diferentes configurações de sensores, condições rodoviárias e tipos de perigos, tornando a tarefa de classificação mais representativa da monitorização rodoviária em contexto real [1], [12], [10].

Em terceiro lugar, a dissertação compara modelos clássicos de aprendizagem automática e modelos de aprendizagem profunda sob um enquadramento experimental consistente [3], [12], [13], [18]. Isto permite uma avaliação mais justa sobre se a aprendizagem profunda oferece vantagens significativas relativamente aos modelos clássicos, especialmente quando o custo computacional é considerado [15].

Em quarto lugar, o estudo aborda explicitamente o desequilíbrio entre classes, que constitui um dos desafios mais importantes na deteção de perigos rodoviários [21], [22],

[23]. Ao utilizar a pontuação F1 macro, a exatidão balanceada, métricas por classe e análise da matriz de confusão, a investigação avalia os modelos de acordo com a sua capacidade de detetar perigos raros, em vez de depender apenas da exatidão global.
Por fim, a análise da importância das características fornece informação sobre quais características dos acelerómetros são mais úteis para a classificação [3], [14], [30]. Isto pode apoiar a conceção de sistemas de monitorização rodoviária mais eficientes e interpretáveis [3], [14], [30].

1.7. Âmbito do Estudo

O âmbito desta dissertação limita-se à classificação de perigos rodoviários baseada em acelerómetros, utilizando aprendizagem automática supervisionada [3], [4], [10]. O estudo centra-se em três classes: superfície rodoviária normal, buraco e lomba [2], [11]. Os dados são obtidos a partir de três conjuntos de dados publicamente disponíveis e processados através de um fluxo unificado que envolve normalização de etiquetas, segmentação por janelas deslizantes, extração de características, divisão entre treino e teste, reequilíbrio do conjunto de treino e avaliação dos modelos [24], [25], [28], [29].
O estudo compara nove modelos clássicos de aprendizagem automática e três arquiteturas de aprendizagem profunda [3], [12], [13], [18]. Os modelos clássicos utilizam características estatísticas extraídas, enquanto os modelos de aprendizagem profunda utilizam sequências brutas de acelerómetros segmentadas em janelas [14], [15], [17], [18], [19], [20]. A avaliação centra-se no desempenho preditivo, na deteção de classes minoritárias, na importância das características e na eficiência computacional [20], [21], [22].
Este estudo não inclui implementação em campo em tempo real, deteção rodoviária baseada em câmaras, agrupamento espacial de perigos baseado em GPS ou fusão multimodal de sensores [31], [32]. Estas áreas são identificadas como possíveis direções para investigação futura [31], [32].

1.8. Estrutura da Dissertação

A presente dissertação encontra-se organizada em seis capítulos. O primeiro capítulo introduz o enquadramento da investigação, apresentando a formulação do problema, o objetivo geral, os objetivos específicos, as questões de investigação, a relevância do estudo, o seu âmbito e a estrutura da dissertação. O segundo capítulo é dedicado à

revisão da literatura, abordando a deteção de perigos rodoviários baseada em acelerómetros, a deteção móvel, os métodos clássicos de aprendizagem automática, a aprendizagem profunda, o problema do desequilíbrio entre classes, a extração de características e as lacunas existentes na investigação [3], [4], [5], [18], [21].

O terceiro capítulo apresenta a metodologia de investigação adotada. Neste capítulo são descritos os conjuntos de dados utilizados, o fluxo de pré-processamento, a normalização das etiquetas, a segmentação por janelas deslizantes, a extração de características, a seleção dos modelos, os procedimentos de treino, a otimização de hiperparâmetros e as métricas de avaliação [15], [24], [25], [27], [28], [29]. O quarto capítulo apresenta os resultados experimentais, reportando o desempenho dos modelos clássicos de aprendizagem automática e dos modelos de aprendizagem profunda, bem como os resultados obtidos com o *Random Forest* otimizado, as matrizes de confusão, a análise da importância das características e o desempenho computacional [3], [18], [19].

O quinto capítulo discute os resultados obtidos à luz das questões de investigação e de estudos anteriores. São analisados o desempenho dos modelos, os efeitos do desequilíbrio entre classes, as implicações práticas, as limitações do estudo e as direções para investigação futura [21], [22], [23], [31], [32]. Por fim, o sexto capítulo conclui a dissertação, sintetizando os principais resultados, os contributos alcançados, as limitações identificadas e as recomendações para trabalhos futuros.

1.9. Síntese do Capítulo

Este capítulo introduziu o contexto e a motivação da investigação sobre classificação de perigos rodoviários baseada em acelerómetros [3], [4], [5]. Explicou a importância da monitorização rodoviária automatizada, as limitações dos métodos tradicionais de inspeção e o potencial das abordagens de aprendizagem automática e aprendizagem profunda [6], [12], [13], [18]. O capítulo identificou também o principal problema de investigação, nomeadamente a dificuldade de detetar com precisão perigos rodoviários raros em conjuntos de dados de acelerómetros heterogéneos e desequilibrados [21], [22], [23]. Em seguida, foram apresentados o objetivo, os objetivos específicos, as questões de investigação, a relevância, o âmbito e a estrutura da dissertação. O próximo capítulo revê a literatura relevante e posiciona o presente estudo no contexto da

investigação existente sobre deteção de perigos rodoviários e sistemas de transporte inteligentes.

Capítulo 2 – Revisão de literatura

A monitorização da infraestrutura rodoviária tornou-se uma área de investigação relevante no âmbito dos sistemas de transporte inteligentes, uma vez que os defeitos na superfície rodoviária podem afetar a segurança dos veículos, o conforto da condução, os custos de manutenção e a eficiência do tráfego. Buracos, superfícies irregulares e lombas produzem padrões de vibração distintivos quando os veículos os atravessam, tornando os sensores inerciais montados em veículos ou integrados em *smartphones* uma fonte prática de dados para a avaliação automatizada das condições rodoviárias. Em comparação com a inspeção manual convencional, a monitorização baseada em acelerómetros oferece a possibilidade de uma vigilância rodoviária contínua, de baixo custo e escalável, particularmente quando os dados são recolhidos através de *smartphones* ou veículos de frota. Esta tendência é igualmente confirmada por uma revisão sistemática conduzida pela autora desta dissertação, no âmbito do mesmo projeto de investigação, que identifica a deteção inercial (acelerómetro e giroscópio) combinada com GNSS/GPS como a pilha de deteção dominante na literatura de monitorização de perigos rodoviários baseada em IoT, presente em cerca de 62% e 57% dos 21 estudos analisados, respetivamente [33].

A literatura sobre deteção de perigos rodoviários evoluiu de métodos iniciais baseados em limiares e regras para abordagens de aprendizagem automática e, mais recentemente, de aprendizagem profunda. Os sistemas iniciais demonstraram que os sensores de *smartphones* podiam detetar anomalias rodoviárias, mas o seu desempenho dependia frequentemente da calibração, das características do veículo, da posição do sensor e do contexto rodoviário. Estudos posteriores introduziram métodos de aprendizagem supervisionada que utilizavam características estatísticas ou no domínio da frequência extraídas de janelas de acelerómetro. Trabalhos mais recentes investigaram redes neuronais convolucionais e recorrentes, capazes de aprender representações diretamente a partir de sequências brutas de aceleração.

Apesar destes avanços, persistem vários problemas por resolver. Em primeiro lugar, os conjuntos de dados de perigos rodoviários são, tipicamente, altamente desequilibrados: os segmentos de estrada normais superam largamente os eventos perigosos. Isto significa que os modelos podem apresentar elevada exatidão, enquanto falham na

deteção de eventos raros, mas críticos para a segurança. Em segundo lugar, grande parte do trabalho existente centra-se na classificação binária, geralmente distinguindo segmentos rodoviários perigosos de segmentos não perigosos, em vez de diferenciar entre tipos de perigos, como buracos e lombas. Em terceiro lugar, as comparações diretas entre abordagens clássicas de aprendizagem automática e abordagens de aprendizagem profunda são limitadas, uma vez que estudos anteriores utilizam frequentemente diferentes conjuntos de dados, etapas de pré-processamento, representações de características e métricas de avaliação. Por fim, muitos estudos dependem de conjuntos de dados privados ou controlados, limitando a reprodutibilidade e dificultando a avaliação da generalização entre conjuntos de dados.

Este capítulo revê a literatura relevante para a classificação de perigos rodoviários baseada em acelerómetros. Começa por definir os fundamentos conceptuais da deteção de perigos rodoviários e da classificação supervisionada de séries temporais. Em seguida, apresenta o enquadramento teórico que sustenta o estudo, seguido de uma síntese temática das principais áreas de investigação: deteção móvel, representação de características, aprendizagem automática clássica, aprendizagem profunda, desequilíbrio entre classes, conjuntos de dados e questões de implementação. O capítulo conclui com a identificação das lacunas na literatura que justificam o presente estudo.

2.1. Enquadramento Conceptual

2.1.1. Perigos rodoviários e sistemas de transporte inteligentes

Os perigos rodoviários referem-se a condições ou estruturas da superfície rodoviária que perturbam o movimento normal do veículo e podem afetar a segurança, o conforto ou o planeamento da manutenção. No contexto desta dissertação, as classes relevantes são superfície rodoviária normal, buracos e lombas. Estas classes são conceptualmente distintas. Os buracos são, tipicamente, defeitos não planeados causados pela deterioração do pavimento, enquanto as lombas são estruturas intencionais de acalmia de tráfego. Contudo, ambas produzem vibrações verticais no veículo e podem, por isso, apresentar semelhanças nos dados de acelerómetros se não forem cuidadosamente distinguidas.

A deteção automatizada de perigos rodoviários está estreitamente associada aos sistemas de transporte inteligentes, uma vez que permite que autoridades rodoviárias,

condutores e entidades responsáveis pela manutenção identifiquem segmentos rodoviários problemáticos sem dependerem exclusivamente de inspeções programadas. Os sistemas de deteção baseados em *smartphones* e sensores montados em veículos são particularmente atrativos, pois permitem recolher dados continuamente durante deslocações comuns. Isto cria um modelo de monitorização colaborativa, no qual múltiplos veículos contribuem com observações sensoriais que podem ser agregadas para identificar problemas na superfície rodoviária.

2.1.2. Monitorização rodoviária baseada em acelerómetros

Os acelerómetros medem a aceleração ao longo de três eixos. Na monitorização rodoviária baseada em veículos, o eixo vertical é particularmente importante, uma vez que buracos, lombas e superfícies irregulares geram choques ascendentes e descendentes à medida que a suspensão do veículo responde às irregularidades da estrada. Os eixos longitudinal e lateral podem também conter informação útil, especialmente quando a velocidade do veículo, a travagem, as curvas ou a instabilidade lateral influenciam o sinal registado.

O pressuposto central na deteção de perigos rodoviários baseada em acelerómetros é que diferentes eventos rodoviários geram diferentes assinaturas vibratórias de curta duração. Os buracos estão frequentemente associados a padrões abruptos de descida e ressalto, enquanto as lombas tendem a produzir respostas verticais mais regulares e sustentadas. Contudo, estas assinaturas são afetadas pela velocidade do veículo, pelo tipo de suspensão, pela orientação do smartphone, pela posição de montagem, pelo material rodoviário e pelo ruído do sensor. Estes fatores de confusão explicam por que razão a deteção simples baseada em limiares é frequentemente insuficiente para uma classificação robusta.

2.1.3. De sinais contínuos a instâncias de classificação

Os dados de acelerómetros são recolhidos como séries temporais contínuas. Para utilizar estes dados em aprendizagem supervisionada, os investigadores segmentam frequentemente o sinal em janelas de comprimento fixo. A cada janela é então atribuída uma etiqueta de classe, sendo esta representada por características manualmente definidas ou por valores sequenciais brutos. Na aprendizagem automática clássica, a extração de características converte cada janela num vetor de dimensão fixa contendo

estatísticas como média, desvio-padrão, raiz do valor quadrático médio, energia, amplitude pico-a-pico e intervalo interquartil. Na aprendizagem profunda, são utilizadas diretamente janelas de sinal brutas ou minimamente processadas, permitindo que o modelo aprenda representações específicas da tarefa.

A segmentação por janelas deslizantes constitui, portanto, uma ponte metodológica fundamental entre a deteção sensorial bruta e a classificação supervisionada. O tamanho da janela e a sobreposição influenciam a captura integral do evento, a possível diluição do sinal por amostras de condução normal e o custo computacional introduzido. Este aspeto é importante para a presente dissertação, uma vez que a tarefa de classificação envolve eventos de curta duração que podem ocupar apenas uma pequena parte de cada janela.

2.2. Enquadramento Teórico

O enquadramento teórico deste estudo combina três perspetivas: deteção móvel participativa, classificação supervisionada de séries temporais e avaliação de classificação desequilibrada.

Em primeiro lugar, a deteção móvel participativa fornece a lógica infraestrutural do domínio. Trabalhos iniciais como o *Pothole Patrol*, de Eriksson et al., demonstraram que redes de sensores móveis podiam apoiar a monitorização da superfície rodoviária através da recolha de dados de acelerómetros e localização a partir de veículos [5]. Sistemas semelhantes baseados em *smartphones*, incluindo Nericell e Wolverine, demonstraram que dispositivos móveis podiam ser utilizados para inferir condições rodoviárias e de tráfego em larga escala [7], [8]. Esta perspetiva enquadra a deteção de perigos rodoviários não apenas como uma tarefa de processamento de sinal, mas como um problema de deteção distribuída, no qual muitos dispositivos de baixo custo podem contribuir para a monitorização da infraestrutura.

Em segundo lugar, a classificação supervisionada de séries temporais fornece o enquadramento computacional. Um sinal bruto de aceleração é transformado em exemplos etiquetados através de pré-processamento, segmentação, extração de características e treino de modelos. As abordagens clássicas de aprendizagem automática dependem de características construídas manualmente que sintetizam a amplitude, a variabilidade, a energia e as relações entre eixos do sinal. As abordagens de aprendizagem profunda, por contraste, procuram aprender padrões temporais

discriminativos diretamente a partir de sequências brutas. Esta distinção é central para a dissertação, uma vez que o estudo compara modelos baseados em características manualmente definidas com modelos de aprendizagem automática de características.
Em terceiro lugar, a teoria da classificação desequilibrada é essencial para avaliar a deteção de perigos rodoviários. Em dados reais de condução, os eventos perigosos são raros em comparação com os segmentos rodoviários normais. Consequentemente, a exatidão pode ser enganadora: um classificador pode alcançar elevada exatidão simplesmente ao prever a classe maioritária, neste caso a classe correspondente à estrada normal, falhando na deteção dos eventos minoritários, como buracos e lombas. A avaliação exige, portanto, métricas que deem visibilidade às classes minoritárias, tais como a pontuação F1 macro, a exatidão balanceada, a precisão por classe, a revocação por classe e as matrizes de confusão. Esta posição teórica apoia diretamente a presente dissertação, que avalia os modelos não apenas pela exatidão global, mas também pela sua capacidade de detetar buracos e lombas.

2.3. Revisão dos Principais Temas da Literatura

2.3.1. Deteção rodoviária participativa e baseada em *smartphones*

Os primeiros trabalhos influentes nesta área enquadraram a deteção de perigos rodoviários como um problema de deteção participativa. Eriksson et al. propuseram o *Pothole Patrol*, que utilizava deteção móvel e localização GPS para detetar buracos, reportando elevada exatidão de deteção nas suas condições experimentais [5]. Este trabalho foi importante porque demonstrou que a monitorização rodoviária podia ser realizada com sensores integrados em veículos, em vez de depender exclusivamente da inspeção manual. Contudo, a deteção baseada em limiares depende fortemente da calibração, da dinâmica do veículo e da configuração do sensor, o que limita a generalização.
Mednis et al. expandiram esta linha de investigação ao avaliar a deteção de buracos em tempo real utilizando *smartphones* Android e acelerómetros [4]. O seu trabalho introduziu classificadores de aprendizagem automática, como árvores de decisão, K-vizinhos mais próximos e Naive Bayes, na deteção de buracos. Yi et al. desenvolveram adicionalmente a monitorização colaborativa de pavimentos com recurso a tecnologias de deteção móvel, incluindo características nos domínios temporal e da frequência com máquinas de vetores de suporte [6]. Em conjunto, estes estudos estabeleceram que a

deteção móvel podia apoiar a monitorização das condições rodoviárias e que a aprendizagem supervisionada podia melhorar abordagens simples baseadas em regras. Outros sistemas baseados em *smartphones* alargaram o âmbito da deteção de buracos para a inferência de condições de tráfego e rodoviárias. O Wolverine utilizou sensores de smartphone para estimar condições de tráfego e da estrada [7], enquanto o Nericell demonstrou uma monitorização mais abrangente das condições rodoviárias e de tráfego utilizando *smartphones* móveis [8]. Estes estudos mostram que a classificação de perigos rodoviários integra um corpo mais amplo de investigação sobre deteção móvel aplicada ao transporte. A sua relevância para a presente dissertação reside na ideia de que dados de acelerómetros recolhidos em condições normais de condução podem fornecer informação útil sobre a infraestrutura rodoviária. Ao mesmo tempo, destacam também desafios práticos persistentes, incluindo a orientação do dispositivo, a variação ambiental e a necessidade de interpretar sinais sensoriais em condições não controladas.

2.3.2. Representação de sinais e engenharia de características

Um tema central na literatura diz respeito à forma como os sinais de acelerómetros devem ser representados para classificação. Os estudos de aprendizagem automática clássica dependem, geralmente, de características manualmente definidas extraídas de janelas fixas. Estas características são concebidas para captar a intensidade do sinal, a dispersão e alterações súbitas causadas por impactos rodoviários. Os descritores comuns incluem média, desvio-padrão, máximo, mínimo, amplitude pico-a-pico, raiz do valor quadrático médio, energia e características no domínio da frequência.

Yi et al. demonstraram que a combinação de descritores nos domínios temporal e da frequência com máquinas de vetores de suporte podia produzir um forte desempenho de classificação em condições controladas [6]. Silva et al. utilizaram um conjunto mais amplo de características estatísticas extraídas de dados de acelerómetros tri-axial e reportaram desempenho robusto com *Random Forest* na classificação binária de buracos [1]. Estes estudos indicam que características manualmente definidas podem captar padrões úteis em dados de aceleração, especialmente quando os eventos rodoviários geram alterações claras e consistentes no sinal.

Contudo, a engenharia de características apresenta limitações. Em primeiro lugar, as estatísticas manualmente definidas podem comprimir padrões de sinal breves e

localizados em valores de síntese, podendo, por isso, perder informação temporal. Em segundo lugar, a mesma característica pode comportar-se de forma diferente consoante a velocidade do veículo, a montagem do sensor e o material rodoviário. Em terceiro lugar, conjuntos de características concebidos para classificação binária podem não ser suficientes para distinguir múltiplos tipos de eventos rodoviários. Esta limitação é diretamente relevante para a presente dissertação, que compara modelos clássicos baseados em características estatísticas com modelos de aprendizagem profunda baseados em sequências brutas segmentadas em janelas.

Os resultados de importância das características na literatura apoiam, de forma geral, a relevância da aceleração vertical e das características baseadas na magnitude. A aceleração vertical é logicamente importante porque as perturbações da superfície rodoviária produzem movimento vertical no veículo. As características baseadas na magnitude combinam informação de todos os eixos e podem reduzir a sensibilidade à orientação do dispositivo. Isto fornece uma base conceptual para incluir, no conjunto de características do presente estudo, a magnitude da aceleração e a magnitude do *jerk*. Neste contexto, o termo *jerk* refere-se à taxa de variação da aceleração ao longo do tempo, permitindo captar alterações súbitas no sinal que podem estar associadas à passagem do veículo por irregularidades rodoviárias, como buracos ou lombas.

2.3.3. Aprendizagem automática clássica na deteção de perigos rodoviários

A aprendizagem automática clássica tem sido amplamente utilizada na deteção de perigos rodoviários, uma vez que pode apresentar bom desempenho com dados limitados e vetores de características interpretáveis. Mednis et al. compararam árvores de decisão, K-vizinhos mais próximos e *Naive Bayes* na deteção de buracos, tendo verificado que as árvores de decisão produziram a melhor exatidão entre os modelos testados [4]. Yi et al. utilizaram máquinas de vetores de suporte com características extraídas e reportaram elevada exatidão num conjunto de dados controlado [6]. Estes estudos demonstram que métodos não baseados em aprendizagem profunda podem classificar eficazmente anomalias rodoviárias quando estão disponíveis características informativas.

Os métodos de *ensemble* têm recebido particular atenção. O *Random Forest* combina múltiplas árvores de decisão treinadas sobre amostras *bootstrap* e subconjuntos aleatórios de características, melhorando a robustez em comparação com uma única

árvore de decisão. Silva et al. reportaram forte desempenho utilizando *Random Forest* com características estatísticas extraídas de dados de acelerómetros tri-axial [1]. Métodos de *gradient boosting* também foram reportados como eficazes; Anaissi et al. utilizaram *XGBoost* com sensores montados em veículos e aprendizagem automática para apoiar a deteção de buracos [13]. O sucesso destes modelos é consistente com a natureza do espaço de características: as características estatísticas de aceleração contêm frequentemente interações não lineares que os *ensembles* de árvores conseguem modelar eficazmente.

Ainda assim, os estudos de aprendizagem automática clássica apresentam frequentemente três limitações. Muitos são concebidos para classificação binária, tornando incerto o seu desempenho na distinção entre diferentes categorias de perigo. Muitos utilizam conjuntos de dados privados ou controlados, reduzindo a reprodutibilidade. Por fim, a elevada exatidão reportada pode não revelar o desempenho das classes minoritárias, especialmente quando o conjunto de dados é desequilibrado. Estas limitações motivam a comparação sistemática de múltiplos modelos clássicos no presente estudo, sob um enquadramento unificado de pré-processamento e avaliação.

2.3.4. Aprendizagem profunda e aprendizagem automática de características

A aprendizagem profunda tornou-se cada vez mais relevante por reduzir a dependência da engenharia manual de características, isto é, da definição prévia de estatísticas e descritores extraídos dos sinais de acelerómetro. Varona et al. propuseram em 2020 uma abordagem de aprendizagem profunda para a monitorização automática da superfície rodoviária e a deteção de buracos utilizando sequências brutas de acelerómetros [18]. A utilização de redes neuronais convolucionais 1D é significativa porque os filtros convolucionais podem aprender padrões temporais locais associados a eventos vibratórios de curta duração. Basavaraju et al. utilizaram abordagens de aprendizagem automática e aprendizagem profunda para a avaliação de anomalias da superfície rodoviária com sensores de smartphone, incluindo arquiteturas capazes de captar padrões espaciais e temporais [12].

As redes neuronais convolucionais são teoricamente adequadas à deteção de perigos rodoviários baseada em acelerómetros, uma vez que conseguem aprender motivos de sinal de curta duração que podem não ser plenamente captados por características estatísticas. Um sinal de buraco pode surgir como um evento breve e localizado dentro

de uma janela. Uma CNN consegue aprender filtros que respondem a esses padrões mesmo quando ocorrem em posições ligeiramente diferentes dentro da janela. As redes LSTM, por contraste, são concebidas para modelar dependências sequenciais e podem ser úteis quando a ordem das alterações de aceleração antes, durante e depois de um evento contém informação discriminativa. Os modelos híbridos CNN-LSTM procuram combinar a extração de padrões locais com a modelação temporal.

A literatura sugere que a aprendizagem profunda pode melhorar a aprendizagem de representações, mas também introduz novos desafios. Os modelos profundos exigem frequentemente mais dados, mais computação e afinação cuidadosa. O seu desempenho pode ser sensível ao desequilíbrio entre classes, ao desenho arquitetural e à estratégia de validação. Em aplicações práticas de monitorização rodoviária, a aprendizagem profunda deve, portanto, ser avaliada não apenas pelo desempenho preditivo, mas também pelo custo computacional e pela viabilidade de implementação. Esta questão é central para a presente dissertação, que compara modelos de aprendizagem profunda com métodos clássicos em termos de métricas sensíveis à classe e de eficiência computacional.

2.3.5. Desequilíbrio entre classes e métricas de avaliação

O desequilíbrio entre classes é um dos desafios metodológicos mais importantes na classificação de perigos rodoviários. Em condições reais de condução, os segmentos de estrada normais são muito mais frequentes do que buracos ou lombas. Consequentemente, um modelo pode alcançar elevada exatidão ao aprender a classe maioritária, enquanto falha na deteção dos eventos raros que são mais relevantes para a segurança e a manutenção.

Allouch et al. abordaram este problema utilizando aprendizagem sensível ao custo com funções de perda ponderadas no *RoadSense* [9]. Esta abordagem reconhece que os erros nas classes minoritárias podem ser mais importantes do que os erros em segmentos de estrada normal. De forma mais ampla, a avaliação consciente do desequilíbrio exige métricas que tratem as classes de forma mais equitativa. A pontuação F1 macro calcula a média dos valores F1 por classe e, portanto, confere maior influência às classes minoritárias do que a exatidão. A exatidão balanceada calcula a média da revocação entre classes, sendo mais informativa quando os suportes

das classes são desiguais. As matrizes de confusão são igualmente essenciais, uma vez que revelam se os erros se concentram em classes específicas.

A implicação para a presente dissertação é que a seleção dos modelos não deve basear-se apenas na exatidão global. Um modelo com exatidão ligeiramente inferior pode ser preferível se detetar mais buracos ou lombas. Inversamente, um modelo com elevada exatidão, mas fraca revocação das classes minoritárias, pode ser inadequado para aplicações críticas em termos de segurança. A literatura apoia, portanto, a utilização da pontuação F1 macro, da exatidão balanceada, da precisão por classe, da revocação por classe e da análise de matrizes de confusão no presente estudo.

2.3.6. Conjuntos de dados, reprodutibilidade e generalização

A disponibilidade de conjuntos de dados continua a constituir uma limitação importante na literatura. Muitos estudos publicados dependem de conjuntos de dados privados recolhidos em condições específicas. Isto dificulta a reprodução dos resultados, a comparação justa entre modelos e a avaliação da generalização entre veículos, sensores e ambientes rodoviários. Conjuntos de dados públicos, como o conjunto de dados da IEEE DataPort sobre estradas irregulares e planas, os Dados de Sensores de Buracos da Kaggle e os conjuntos de dados PVS de Sensores Veiculares Passivos, oferecem oportunidades para uma avaliação comparativa mais reprodutível [27], [28].

Contudo, a utilização de múltiplos conjuntos de dados públicos introduz os seus próprios desafios metodológicos. Os conjuntos de dados podem diferir em taxas de amostragem, posição dos sensores, definições de etiquetas, procedimentos de anotação e tipos de eventos. Estas diferenças criam heterogeneidade que deve ser tratada através de pré-processamento e normalização de etiquetas. Introduzem também desvio de domínio: um modelo treinado numa fonte pode não apresentar desempenho igualmente bom noutra. Esta questão é importante para a classificação de perigos rodoviários, uma vez que a implementação prática exige modelos que generalizem entre veículos, modelos de smartphone, posições de montagem, velocidades e materiais rodoviários.

A presente dissertação responde a esta lacuna ao combinar múltiplos conjuntos de dados públicos num enquadramento de classificação unificado. Esta estratégia melhora a representatividade do estudo em condições próximas do contexto real, ou seja, a sua validade ecológica, em comparação com um único conjunto de dados controlado. No

entanto, também torna a tarefa de classificação mais difícil, uma vez que introduz maior heterogeneidade entre fontes de dados.

2.4. Estudos Empíricos e Investigação Anterior

A literatura empírica pode ser agrupada em quatro fases principais. A primeira fase consiste em sistemas de deteção participativa e baseados em limiares. Eriksson et al. demonstraram que redes de sensores móveis podiam detetar buracos utilizando dados de acelerómetros e GPS [5]. Este trabalho estabeleceu a viabilidade da monitorização rodoviária através de veículos em movimento, mas também evidenciou a dependência dos métodos baseados em limiares relativamente à calibração e a características específicas do veículo. O Nericell e o Wolverine demonstraram igualmente o potencial mais amplo dos sensores de *smartphones* para a monitorização das condições rodoviárias e de tráfego [7], [8].

A segunda fase introduziu a aprendizagem automática clássica. Mednis et al. compararam vários classificadores para a deteção de buracos em tempo real utilizando acelerómetros de *smartphones* Android e reportaram forte desempenho para árvores de decisão [4]. Yi et al. utilizaram máquinas de vetores de suporte com características nos domínios temporal e da frequência, demonstrando que a aprendizagem automática podia melhorar a monitorização de pavimentos com recurso a deteção móvel [6]. Estes estudos forneceram evidência importante de que os sinais de acelerómetros contêm informação discriminativa para a deteção de anomalias rodoviárias.

A terceira fase centrou-se na aprendizagem por ensemble e em conjuntos de características mais robustos. Silva et al. reportaram elevada exatidão na classificação binária de buracos utilizando *Random Forest* com características estatísticas de acelerómetros [1]. Anaissi et al. demonstraram de forma semelhante que o *gradient boosting* podia ser eficaz quando treinado com dados de sensores montados em veículos [13]. Estes estudos indicam que os métodos de ensemble baseados em árvores são candidatos fortes para a deteção de perigos rodoviários baseada em características. Contudo, o seu foco na deteção binária deixa em aberto a questão do seu desempenho quando múltiplas classes de perigos devem ser distinguidas.

A quarta fase introduziu a aprendizagem profunda. Varona et al. utilizaram 1D CNNs para aprender diretamente a partir de sequências de acelerómetro, reduzindo a dependência da engenharia manual de características [18]. Basavaraju et al. utilizaram

abordagens CNN e LSTM para captar padrões espaciais e temporais em dados sensoriais de *smartphones* [12]. Estes estudos sugerem que a aprendizagem profunda pode oferecer vantagens para padrões de eventos rodoviários complexos ou subtis. Contudo, a literatura ainda não estabeleceu de forma conclusiva se a aprendizagem profunda supera consistentemente a aprendizagem automática clássica quando são considerados o desequilíbrio entre classes, o custo computacional e as restrições de implementação.

Ao longo destas fases, existe amplo consenso de que os dados de acelerómetros são úteis para a monitorização das condições rodoviárias. Existe também consenso de que a aceleração vertical e as características relacionadas com a magnitude são informativas, que a aprendizagem supervisionada melhora os métodos simples baseados em limiares e que o desequilíbrio entre classes complica a avaliação. A principal divergência ou incerteza diz respeito à escolha do modelo: os métodos clássicos baseados em características são frequentemente eficientes e interpretáveis, enquanto os métodos de aprendizagem profunda podem ser mais adequados para aprender padrões temporais subtis, embora exijam mais dados e capacidade computacional. Este compromisso por resolver é central para a presente dissertação.

2.5. Lacunas na Literatura

Da literatura revista emergem várias lacunas.

Em primeiro lugar, continuam a ser limitadas as comparações diretas e justas entre aprendizagem automática clássica e aprendizagem profunda. Muitos estudos centram-se numa única família algorítmica, utilizam diferentes conjuntos de dados ou adotam diferentes fluxos de pré-processamento. Isto dificulta determinar se as diferenças observadas se devem à arquitetura do modelo ou ao desenho experimental.

Em segundo lugar, grande parte do trabalho anterior centra-se na classificação binária. Embora a deteção binária de buracos seja útil, a monitorização rodoviária prática exige frequentemente distinguir diferentes tipos de eventos. Lombas, buracos e superfícies irregulares podem produzir padrões de vibração relacionados, mas têm implicações distintas para a manutenção. A classificação multiclasse é, portanto, mais realista, mas encontra-se menos explorada.

Em terceiro lugar, o desequilíbrio entre classes nem sempre é tratado adequadamente. A exatidão continua a ser amplamente reportada, mas pode ocultar a fraca deteção de

perigos raros. Estudos que utilizam aprendizagem sensível ao custo ou avaliação sensível às classes demonstram a importância de métodos conscientes do desequilíbrio, mas é necessária uma comparação sistemática adicional.

Em quarto lugar, a reprodutibilidade e a generalização continuam condicionadas por limitações dos conjuntos de dados. Conjuntos de dados privados e controlados limitam a validação externa. Existem conjuntos de dados públicos, mas estes diferem em posição dos sensores, taxa de amostragem, etiquetas e condições de recolha. É necessária mais investigação sobre pré-processamento unificado, teste entre conjuntos de dados e adaptação de domínio. Esta limitação é igualmente documentada na revisão sistemática conduzida pela autora no contexto desta dissertação, que constata que apenas duas das 21 publicações analisadas apresentam validação real completa e que as práticas de privacidade e consentimento raramente são formalizadas, evidenciando uma lacuna estrutural de validade externa em todo o domínio [33].

Em quinto lugar, as considerações de implementação permanecem pouco desenvolvidas. Os sistemas de deteção de perigos rodoviários devem operar sob restrições reais, tais como alterações na orientação do smartphone, variabilidade dos veículos, ruído sensorial, latência, consumo de bateria e gestão de falsos alarmes. Estudos que reportam elevada exatidão offline podem não demonstrar plenamente a adequação prática dos modelos.

Em sexto lugar, a interpretabilidade das características e a transparência dos modelos requerem maior atenção. Os modelos clássicos permitem análise da importância das características, mas os modelos de aprendizagem profunda são frequentemente menos interpretáveis. Em aplicações relacionadas com segurança, compreender por que razão um modelo prevê um buraco ou uma lomba é importante para a confiança, a validação e a integração em fluxos de trabalho de manutenção municipal.

Estas lacunas justificam o desenho comparativo do presente estudo. Ao utilizar múltiplos conjuntos de dados públicos, um fluxo de pré-processamento unificado, modelos clássicos e de aprendizagem profunda, métricas sensíveis às classes, análise computacional e interpretação da importância das características, a dissertação aborda várias limitações identificadas na literatura.

2.6. Síntese do Capítulo e Ligação ao Presente Estudo

Este capítulo reviu a literatura sobre classificação de perigos rodoviários baseada em acelerómetros. A literatura demonstra que sensores de *smartphones* e sensores montados em veículos podem apoiar a monitorização rodoviária escalável, especialmente no âmbito da deteção participativa e dos sistemas de transporte inteligentes. As abordagens iniciais baseadas em limiares demonstraram viabilidade, mas eram sensíveis à calibração e às condições do veículo. A aprendizagem automática clássica melhorou a deteção através da utilização de características estatísticas e no domínio da frequência, com métodos de ensemble, como o *Random Forest e o gradient boosting*, a demonstrarem forte desempenho. As abordagens de aprendizagem profunda, particularmente modelos baseados em CNN e LSTM, introduziram a aprendizagem automática de características a partir de sequências brutas de aceleração e ofereceram uma potencial solução para as limitações das características manualmente definidas.

A revisão também demonstrou que a deteção de perigos rodoviários continua a ser metodologicamente desafiante. A exatidão global pode induzir interpretações incorretas em cenários de forte desequilíbrio entre classes. Além disso, a classificação multiclasse é menos estudada do que a deteção binária, os conjuntos de dados públicos continuam limitados e a generalização entre conjuntos de dados ainda não está plenamente estabelecida. Estas questões são diretamente relevantes para a presente dissertação.

O presente estudo posiciona-se, portanto, como uma investigação empírica comparativa entre abordagens de aprendizagem automática clássica e de aprendizagem profunda para a classificação de perigos rodoviários com recurso a dados de acelerómetro. O estudo baseia-se em trabalhos anteriores ao combinar conjuntos de dados públicos, harmonizar etiquetas, aplicar segmentação por janelas deslizantes, comparar múltiplas famílias de modelos, avaliar o desempenho com métricas sensíveis ao desequilíbrio e analisar compromissos computacionais. Ao fazê-lo, a dissertação procura fornecer evidência mais clara para a seleção de modelos em sistemas práticos de monitorização rodoviária e identificar as condições em que abordagens clássicas ou de aprendizagem profunda são mais adequadas.

Capítulo 3 – Metodologia

Este capítulo descreve a metodologia experimental adotada para investigar a classificação de perigos rodoviários com recurso a dados adquiridos através do sensor acelerómetro. A metodologia abrange a aquisição de dados, o pré-processamento, a engenharia de características, a seleção de modelos, os procedimentos de treino e os protocolos de avaliação. A abordagem foi concebida de modo a permitir uma comparação rigorosa entre paradigmas de aprendizagem automática clássica e de aprendizagem profunda, respondendo simultaneamente aos desafios inerentes a dados sensoriais de diferentes fontes e desequilibrados.

3.1. Fontes de dados e aquisição

Foram selecionados três conjuntos de dados que estão disponíveis publicamente, com o objetivo de abranger condições rodoviárias, configurações de sensores e tipologias de perigos distintas. Os critérios de seleção privilegiaram conjuntos de dados que continham medições de acelerómetros tri-axial, acompanhadas de anotações de referência relativas aos perigos identificados, bem como uma resolução temporal suficiente para a deteção de eventos. O Anexo A – Comparação dos Conjuntos de Dados de Origem apresenta uma comparação detalhada dos três conjuntos de dados, incluindo o número de amostras, a taxa de amostragem e a distribuição de classes antes da integração.

3.1.1. Conjunto de dados IEEE DataPort

O primeiro conjunto de dados, intitulado "*Dataset Collected on Rough and Plane Roads*", foi obtido a partir da plataforma IEEE DataPort [27]. Este conjunto de dados é composto por registos de acelerómetros amostrados a 100 Hz, captando o movimento do veículo em segmentos rodoviários que incluem buracos e superfícies regulares. As colunas originais dos dados (X, Y, Z) foram renomeadas para accX, accY e accZ, de modo a assegurar a consistência entre os diferentes conjuntos de dados. Foram disponibilizadas etiquetas binárias, nas quais 0 representa uma superfície rodoviária normal e 1 indica a presença de um buraco. As marcas temporais foram sintetizadas em intervalos de 10 ms (1/100 Hz), com o objetivo de facilitar o alinhamento temporal durante a etapa de pré-processamento.

3.1.2. Dados de Sensores de Buracos da plataforma Kaggle

O segundo conjunto de dados, proveniente da plataforma Kaggle [29], contém cinco percursos de condução, com medições de acelerómetros registadas através de sensores de smartphone. As colunas dos dados accelerometerX, accelerometerY e accelerometerZ representam a aceleração tri-axial. Os eventos correspondentes a buracos foram anotados com marcas temporais precisas, aos quais foram associados às marcas temporais dos sensores através de correspondência temporal exata (*tolerance* = 0 ms). As etiquetas distinguem entre condições normais de condução (*none*) e ocorrências de buracos (*pothole*).

3.1.3. Conjunto de Dados de Sensores Veiculares Passivos

O terceiro conjunto de dados utilizado corresponde ao PVS, obtido a partir da plataforma Kaggle e originalmente publicado por Menegazzo e von Wangenheim [28]. Este conjunto de dados reúne várias sessões de recolha em contexto veicular, contendo medições provenientes de sensores instalados em diferentes posições do veículo, bem como informação de localização e velocidade.

Uma particularidade deste conjunto de dados é a existência de medições registadas em diferentes posições físicas do veículo. Em cada sessão, foram considerados os dados relativos aos dois lados do veículo e, para cada lado, os sinais encontravam-se organizados segundo diferentes posições de sensor, nomeadamente acima da suspensão, abaixo da suspensão e no painel de instrumentos. Esta organização torna o conjunto de dados particularmente relevante para o presente estudo, uma vez que permite considerar sinais recolhidos em locais distintos do veículo, os quais podem refletir de forma diferente as vibrações provocadas por irregularidades rodoviárias.

Para permitir a integração deste conjunto de dados no fluxo experimental comum, os dados originais foram reorganizados e harmonizados. Em primeiro lugar, as etiquetas disponibilizadas no conjunto de dados foram associadas às respetivas observações. De seguida, as diferentes posições de sensor foram tratadas como fontes de medição distintas, originando subconjuntos independentes para cada combinação de sessão, lado do veículo e posição do sensor.

Este procedimento não consistiu na fusão ou média dos sinais recolhidos nas várias posições. Em vez disso, cada posição foi preservada como uma fonte autónoma de

dados, sendo posteriormente convertida para uma estrutura comum. Esta harmonização permitiu que os sinais provenientes de diferentes localizações do veículo fossem processados pelo mesmo fluxo metodológico, tornando-os comparáveis aos restantes conjuntos de dados utilizados no estudo.

No que respeita às etiquetas, as ocorrências associadas à presença de lombas foram agregadas numa única classe correspondente a *speed_bump*, independentemente do tipo de superfície rodoviária em que tinham sido registadas. As restantes observações foram associadas à classe *none*. Desta forma, o conjunto PVS contribuiu sobretudo com exemplos das classes *speed_bump* e *none*, complementando os restantes conjuntos de dados utilizados na classificação de perigos rodoviários [3], [10].

3.2. Fluxo de Pré-processamento dos Dados

Foi implementado um fluxo de pré-processamento unificado com o objetivo de uniformizar os conjuntos de dados heterogéneos e preparar os dados para o treino dos modelos. Este fluxo, encapsulado no módulo *road_hazard_common.py*, consiste na normalização das etiquetas, na segmentação por janelas deslizantes, na extração de características e na divisão entre conjuntos de treino e teste com estratificação baseada em grupos.

3.2.1. Normalização de etiquetas

De modo a assegurar a consistência entre os diferentes conjuntos de dados, todas as etiquetas de perigos rodoviários foram normalizadas de acordo com uma taxonomia comum. Foram aplicados os seguintes mapeamentos: as variantes das etiquetas relativas a buracos, como "*potholes*", foram mapeadas para "*pothole*"; as variantes das etiquetas relativas a lombas, como "*speed_bumps*", foram mapeadas para "*speed_bump*"; e as condições rodoviárias normais, como "*normal*", "*smooth*" e "*no_hazard*", foram mapeadas para "*none*".

Esta normalização resultou num problema de classificação composto por três classes: *none*, *pothole* e *speed_bump*.

3.2.2. Segmentação por Janelas deslizantes

Foi utilizada a segmentação por janelas deslizantes para transformar séries temporais contínuas de dados de acelerómetros em janelas de análise de comprimento fixo, uma

abordagem amplamente adotada no reconhecimento de atividades baseado em sensores [24], [25]. Foi selecionado um tamanho de janela de 50 amostras, com um passo de 25 amostras, correspondente a uma sobreposição de 50%. Esta escolha baseou-se em trabalhos anteriores que demonstram que uma sobreposição moderada melhora a resolução temporal, mantendo simultaneamente a eficiência computacional.
O modo de segmentação foi definido como "*auto*", permitindo selecionar a segmentação baseada em índices de amostras quando estes se encontram disponíveis e a segmentação baseada em marcas temporais nos restantes casos. Esta estratégia adaptativa permite acomodar conjuntos de dados com diferentes estruturas de metadados temporais.

3.2.3. Estratégia de Rotulagem das Janelas

A cada janela foi atribuída uma única etiqueta, de acordo com um esquema baseado em prioridades, concebido para preservar eventos de perigo raros na presença de desequilíbrio entre classes. A prioridade de rotulagem foi definida da seguinte forma: *pothole* > *speed_bump* > *none*. Consequentemente, se qualquer amostra dentro de uma janela estivesse etiquetada como *pothole*, toda a janela recebia a etiqueta *pothole*. Caso não existissem amostras de *pothole*, mas estivesse presente pelo menos uma amostra de *speed_bump*, a janela era etiquetada como *speed_bump*. As janelas que continham exclusivamente amostras da classe *none* eram etiquetadas como *none*.
Este esquema de prioridades mitiga o risco de os eventos de perigo serem atenuados por amostras pertencentes à classe maioritária dentro de janelas sobrepostas, um aspeto crítico tendo em conta o acentuado desequilíbrio entre classes observado em dados reais de monitorização rodoviária [21], [22].

3.2.4. Extração de características

Para os modelos de aprendizagem automática clássica, foi extraído de cada janela um conjunto abrangente de 58 características estatísticas e de processamento de sinal. As características foram calculadas de forma independente para cada um dos quatro canais de sinal: accX, accY, accZ e magnitude, sendo a magnitude definida como:

$$\text{magnitude} = \sqrt{\text{accX}^2 + \text{accY}^2 + \text{accZ}^2}$$

Adicionalmente, a magnitude do jerk foi calculada como a diferença de primeira ordem do sinal de magnitude:

$$\text{jerk\_magnitude} = \Delta(\text{magnitude})$$

Para cada um dos cinco canais (*accX*, *accY*, *accZ*, magnitude e *jerk_magnitude*), foram extraídas as seguintes 11 características: média, desvio-padrão (std), mínimo (min), máximo (max), amplitude pico-a-pico (ptp), raiz do valor quadrático médio (rms), intervalo interquartil (iqr), mediana, variância (var), média dos valores absolutos (abs_mean) e energia, definida como a soma dos valores ao quadrado.

Este processo resultou em 55 características, correspondentes a 11 características aplicadas a 5 canais. Adicionalmente, foram calculadas três características destinadas a captar correlações entre eixos: correlação entre *accX* e *accY*, correlação entre *accX* e *accZ*, e correlação entre *accY* e *accZ*.

O vetor final de características compreendeu 58 dimensões. Este conjunto de características está alinhado com práticas estabelecidas na classificação baseada em acelerómetros, incorporando estatísticas no domínio temporal e descritores baseados em energia, capazes de captar tanto a tendência central como a variabilidade dos sinais [3], [30], [14].

Para os modelos de aprendizagem profunda, as sequências brutas segmentadas em janelas foram utilizadas diretamente como entrada, sendo cada janela representada por uma matriz de dimensão 50 x 4, correspondente a 50 passos temporais e 4 canais: *accX*, *accY*, *accZ* e magnitude. Esta abordagem permite que arquiteturas convolucionais e recorrentes aprendam representações específicas da tarefa diretamente a partir dos dados segmentados, sem necessidade de definir manualmente características estatísticas como média, desvio-padrão, RMS, energia ou amplitude pico-a-pico [18], [19], [20].

3.2.5. Divisão entre Conjuntos de Treino e Teste

De modo a prevenir a fuga de informação (*data leakage*) e assegurar estimativas mais robustas de generalização, foi implementada uma divisão entre conjuntos de treino e teste baseada em grupos, recorrendo ao método *GroupShuffleSplit*, com uma proporção de 80/20. A divisão foi realizada ao nível do ficheiro de origem, e não ao nível das janelas individuais, uma vez que janelas consecutivas podem partilhar amostras e apresentar elevada correlação temporal. A variável de agrupamento utilizada foi o identificador do

ficheiro de origem, garantindo que todas as janelas derivadas de um determinado ficheiro fossem atribuídas exclusivamente ao conjunto de treino ou ao conjunto de teste. Esta estratégia é particularmente crítica no contexto de dados sensoriais temporalmente correlacionados, uma vez que uma divisão aleatória pode conduzir a estimativas de desempenho excessivamente otimistas, devido à proximidade temporal entre amostras de treino e de teste [25].

3.2.6. Reequilíbrio do Conjunto de Treino

Dado o acentuado desequilíbrio entre classes no conjunto de dados combinado, foi aplicada uma estratégia de reequilíbrio exclusivamente ao conjunto de treino, com o objetivo de reduzir a tendência dos modelos para privilegiar a classe maioritária (*none*) e melhorar a aprendizagem das classes minoritárias (*pothole* e *speed_bump*). As classes minoritárias (*pothole* e *speed_bump*) foram sobreamostradas até igualarem o número de amostras da classe minoritária mais frequente. A classe maioritária (*none*) foi limitada ao dobro do número máximo de amostras de perigo, de modo a evitar a sua dominância completa, preservando simultaneamente um número suficiente de exemplos negativos.

Esta abordagem de reequilíbrio visa melhorar a sensibilidade dos modelos à deteção de eventos de perigo raros, sem introduzir uma sobrecarga computacional excessiva [21], [22].

O conjunto de teste foi mantido desequilibrado, de forma a refletir as distribuições de classes observadas em cenários reais e a proporcionar estimativas de desempenho mais realistas.

3.3. Modelos Clássicos de Aprendizagem Automática

Foram avaliados nove algoritmos clássicos de aprendizagem automática, representando diversos paradigmas de aprendizagem, incluindo métodos de grupo baseados em árvores, máquinas de vetores de suporte, redes neuronais e métodos baseados em instâncias. Todos os modelos foram implementados com recurso à biblioteca *scikit-learn* [15] e encapsulados em objetos *Pipeline*, os quais incluíam o *StandardScaler* para a normalização das características, sempre que aplicável.

3.3.1. Seleção dos Modelos

Foram selecionados os seguintes modelos:

1. **Gradient Boosting baseado em Histogramas (*Hist Gradient Boosting*)**: uma implementação de *gradient boosting* otimizada para grandes conjuntos de dados, recorrendo a divisões baseadas em histogramas para maior eficiência computacional [16].
2. **Gradient Boosting**: uma abordagem tradicional de *gradient boosting* com árvores de decisão como modelos base, reconhecida pelo seu elevado desempenho em dados tabulares [16].
3. **Random Forest**: um grupo de árvores de decisão treinadas sobre amostras *bootstrap* e subconjuntos aleatórios de características, proporcionando robustez e interpretabilidade [11], [17], [34].
4. **Rede Neuronal Perceptrão Multicamada (MLP)**: uma rede neuronal *feedforward* com camadas ocultas, capaz de aprender fronteiras de decisão não lineares [30].
5. **Extra Trees**: um método de agrupamento semelhante ao *Random Forest*, mas com divisões totalmente aleatórias, oferecendo menor variância e tempos de treino mais reduzidos [16].
6. **Máquina de Vetores de Suporte com núcleo RBF (SVM RBF)**: um classificador baseado em núcleos que projeta os dados para um espaço de maior dimensionalidade, permitindo a separação não linear [35].
7. **Regressão Logística**: um classificador linear que fornece saídas probabilísticas e funciona como modelo de referência para comparação [2].
8. **K-Vizinhos Mais Próximos (KNN)**: um classificador baseado em instâncias que atribui etiquetas com base na votação maioritária entre os k exemplos de treino mais próximos [10].
9. **Árvore de Decisão**: uma única árvore de decisão que fornece regras interpretáveis, embora seja suscetível a sobreajustamento (*overfitting*) [16].

Estes modelos foram escolhidos de modo a abranger diferentes níveis de complexidade, interpretabilidade e requisitos computacionais, permitindo uma comparação abrangente entre diferentes famílias algorítmicas [3], [17].

3.3.2. Procedimento de Treino

Cada modelo foi treinado sobre o conjunto de treino reequilibrado, utilizando hiperparâmetros predefinidos, com exceção dos parâmetros *random_state*, definidos com o objetivo de assegurar que os resultados são reproduzíveis. O tempo de treino e o tempo de predição foram registados para efeitos de análise da eficiência computacional. Os modelos foram avaliados no conjunto de teste não reequilibrado, de modo a aferir o seu desempenho em condições representativas de cenários reais.

3.4. Otimização de Hiperparâmetros do *Random Forest*

Dado o elevado desempenho de referência apresentado pelo *Random Forest* e a sua ampla utilização em tarefas de classificação baseadas em dados de acelerómetros [11], [17], [34], foi conduzida uma otimização sistemática dos hiperparâmetros, com o objetivo de maximizar o desempenho de classificação.

3.4.1. Espaço de Pesquisa de Hiperparâmetros

Foi adotada uma estratégia de pesquisa aleatória sobre 40 configurações, recorrendo-se à validação cruzada com 5 partições (*5-fold cross-validation*) para estimar o desempenho de generalização. O espaço de pesquisa incluiu os seguintes hiperparâmetros:

- *n_estimators*: número de árvores na floresta
- *max_depth*: profundidade máxima de cada árvore
- *min_samples_split*: número mínimo de amostras necessário para dividir um nó interno
- *min_samples_leaf*: número mínimo de amostras exigido num nó folha
- *max_features*: número de características consideradas em cada divisão, por exemplo, *sqrt* ou *log2*
- *max_samples*: fração de amostras utilizada para o treino de cada árvore, no contexto da amostragem bootstrap
- *criterion*: medida de qualidade da divisão, nomeadamente *gini* ou *entropy*
- *bootstrap*: indicação de utilização, ou não, de amostragem *bootstrap*

O objetivo da otimização foi a pontuação F1 com média macro (*macro-averaged F1-score*), selecionada por permitir equilibrar o desempenho entre as três classes, apesar do acentuado desequilíbrio existente entre elas [21].

3.4.2. Melhor Configuração

A configuração ótima identificada através da pesquisa aleatória foi a seguinte: *n_estimators* = 500; *max_depth* = 30; *min_samples_split* = 5; *min_samples_leaf* = 1; *max_features* = log2; *max_samples* = 0.8; *criterion* = *entropy*; e *bootstrap* = *True*.
Esta configuração obteve uma pontuação F1 com média macro, em validação cruzada, de 0.9456 no conjunto de treino, tendo sido posteriormente avaliada no conjunto de teste reservado.

3.5. Modelos de Aprendizagem Profunda

Foram implementadas três arquiteturas de aprendizagem profunda com o objetivo de avaliar a eficácia da aprendizagem automática de características a partir de sequências brutas de dados de acelerómetro. Todos os modelos foram desenvolvidos com recurso ao *TensorFlow/Keras* [26] e treinados sobre sequências brutas segmentadas em janelas, com uma dimensão de entrada de 50 x 4.

3.5.1. Rede Neuronal Convolucional 1D (*1D CNN*)

A arquitetura *1D CNN* foi concebida para extrair padrões temporais locais dos sinais de acelerómetros através de camadas convolucionais hierárquicas [19], [20]. A arquitetura compreendeu: múltiplas camadas convolucionais 1D com função de ativação ReLU; camadas de *max pooling* para subamostragem temporal; camadas de *dropout* para regularização; camadas totalmente ligadas para classificação; e uma camada de saída *softmax* para predição em três classes.
As *1D CNNs* têm demonstrado um desempenho robusto em dados sensoriais de séries temporais, ao aprenderem detetores de características invariantes a deslocamentos, capazes de captar padrões de sinal característicos associados a perigos rodoviários [18], [19].

3.5.2. Rede de Memória de Curto e Longo Prazo (LSTM)

A arquitetura LSTM foi selecionada com o objetivo de modelar dependências temporais de longo alcance em sequências de dados de acelerómetros [36], [37]. A arquitetura consistiu em: camadas LSTM empilhadas com *dropout* recorrente; camadas densas para integração de características; e uma camada de saída *softmax*.

As LSTM são particularmente adequadas a dados sequenciais nos quais o contexto temporal e a ordenação dos eventos são fatores críticos, uma vez que mantêm estados internos de memória capazes de captar informação histórica [36], [37].

3.5.3. Arquitetura Híbrida CNN + LSTM

O modelo híbrido CNN+LSTM combina as vantagens da extração de características convolucional com a modelação temporal recorrente [21], [38], [37]. A arquitetura compreendeu: camadas convolucionais 1D iniciais para extração de características locais, camadas LSTM para modelar dependências temporais nas características aprendidas, camadas densas para classificação e uma camada de saída *softmax*.

Esta abordagem híbrida tem demonstrado superar modelos CNN ou LSTM isolados em tarefas de reconhecimento de atividades, ao tirar partido simultaneamente da abstração espacial e temporal [21], [37].

3.5.4. Configuração de Treino

Todos os modelos de aprendizagem profunda foram treinados com a seguinte configuração: função de perda *categorical cross-entropy*, otimizador Adam com taxa de aprendizagem inicial de 0.001, tamanho de lote (*batch size*) de 64, paragem antecipada (*early stopping*) com paciência de 5 *epochs* relativamente à perda de validação, redução da taxa de aprendizagem através do método *ReduceLROnPlateau*, com fator de 0.5 e paciência de 3 *epochs*, e divisão de validação correspondente a 20% dos dados de treino.

Nos modelos de aprendizagem profunda, o desequilíbrio entre classes foi tratado através de pesos de classe calculados a partir da distribuição do conjunto de treino. Ao contrário dos modelos clássicos, as sequências brutas não foram sobreamostradas diretamente. O desempenho de validação foi monitorizado com o objetivo de reduzir o risco de sobreajustamento, tendo sido registado o número de *epochs* treinados, a perda final de validação e a exatidão de validação para cada arquitetura.

3.6. Métricas de avaliação

Dado o acentuado desequilíbrio entre classes no conjunto de teste, composto por 96,2% de amostras da classe *none*, 2,1% da classe *pothole* e 1,7% da classe *speed_bump*, a exatidão (*accuracy*) isoladamente revela-se insuficiente para a avaliação do desempenho dos modelos. Assim, foi utilizado um conjunto abrangente de métricas de avaliação:

3.6.1. Métricas Globais

- **Exatidão (*Accuracy*)**: proporção de amostras corretamente classificadas.
- **Exatidão balanceada (*Balanced Accuracy*)**: média da revocação por classe, atribuindo igual peso a cada classe, independentemente do respetivo suporte.
- **Precisão, revocação e pontuação F1 com média macro (*macro-averaged Precision*, *Recall* and *F1-score*)**: média aritmética das métricas calculadas por classe, tratando todas as classes de forma equitativa.

3.6.2. Métricas por Classe

Para cada classe (*none*, *pothole* e *speed_bump*), foram calculadas as seguintes métricas:

- **Precisão (*Precision*)**: proporção de verdadeiros positivos entre as amostras classificadas como positivas.
- **Revocação (*Recall*)**: proporção de verdadeiros positivos entre as amostras efetivamente positivas.
- **Pontuação F1 (*F1-score*)**: média harmónica entre a precisão e a revocação.

3.6.3. Matriz de Confusão

Foram geradas matrizes de confusão com o objetivo de visualizar os padrões de classificação e identificar tendências sistemáticas de classificação incorreta.

3.6.4. Métricas Computacionais

- **Tempo de treino**: tempo de relógio (*wall-clock time*) necessário para o treino do modelo.

- **Tempo de predição**: tempo de relógio necessário para a inferência sobre o conjunto de teste.

Estas métricas proporcionam uma visão global do desempenho dos modelos, equilibrando a precisão da classificação, a sensibilidade específica por classe e a eficiência computacional [3], [21], [22].

3.6.5. Estimativa de Incerteza por *Bootstrap*

Para complementar a avaliação pontual dos modelos, foi realizada uma análise de incerteza baseada em *bootstrap* sobre as métricas calculadas a partir das predições obtidas no conjunto de teste. O *bootstrap* é uma técnica estatística de reamostragem com reposição que permite estimar a variabilidade de uma métrica a partir da distribuição empírica dos dados observados, sendo amplamente utilizado para obter estimativas de incerteza e intervalos de confiança sem assumir uma distribuição paramétrica específica [39], [40].

O procedimento foi aplicado aos ficheiros de predições gerados para cada modelo, contendo a etiqueta real e a etiqueta prevista para cada janela do conjunto de teste. Para cada modelo, foram efetuadas 5000 reamostragens com reposição das observações do conjunto de teste. Em cada reamostragem, foram recalculadas as principais métricas de avaliação, nomeadamente a exatidão, a exatidão balanceada, a precisão macro, a revocação macro, a pontuação F1 macro, a pontuação F1 ponderada e as métricas específicas por classe.

A escolha de 5000 reamostragens resultou de um compromisso entre estabilidade estatística e custo computacional. Este número permite obter estimativas mais estáveis dos percentis utilizados na construção dos intervalos de confiança, sem implicar um custo computacional excessivo, uma vez que o procedimento foi aplicado às predições previamente obtidas e não exigiu novo treino dos modelos [39], [40].

Os intervalos de confiança a 95% foram estimados através dos percentis 2,5% e 97,5% da distribuição empírica das métricas obtida por bootstrap. Esta abordagem permitiu complementar os valores pontuais reportados com uma estimativa da sua incerteza, sendo particularmente relevante num contexto de forte desequilíbrio entre classes, no qual as métricas associadas às classes minoritárias podem apresentar maior variabilidade [21], [22].

Importa salientar que esta análise quantifica a incerteza associada às métricas calculadas a partir das predições no conjunto de teste, mas não substitui uma validação externa independente. Além disso, uma vez que as janelas foram obtidas a partir de séries temporais segmentadas com sobreposição, os intervalos de confiança devem ser interpretados com prudência, pois as observações não são totalmente independentes [24], [25]. Ainda assim, o *bootstrap* fornece uma análise adicional da robustez dos resultados e contribui para uma interpretação mais rigorosa do desempenho dos modelos avaliados [39], [40].

3.7. Ambiente Experimental

Todos os ensaios foram conduzidos num ambiente computacional uniformizado, de modo a assegurar a reprodutibilidade dos resultados. A pilha de software incluiu Python 3.x, scikit-learn para os modelos clássicos, *TensorFlow/Keras* para os modelos de aprendizagem profunda e bibliotecas científicas padrão, nomeadamente *NumPy*, *Pandas* e *Matplotlib*. As sementes aleatórias foram fixadas em todos os ensaios, com o objetivo de controlar os componentes pseudoaleatórios do processo experimental e aumentar a reprodutibilidade dos resultados, sem eliminar a aleatoriedade inerente a alguns métodos.

Capítulo 4 – Resultados

Este capítulo apresenta os resultados experimentais obtidos a partir da avaliação comparativa entre abordagens de aprendizagem automática clássica e de aprendizagem profunda aplicadas à classificação de perigos rodoviários. Os resultados encontram-se organizados por categoria de modelo, sendo reportadas métricas de desempenho detalhadas, matrizes de confusão, análise da importância das características e dinâmicas de treino.

4.1. Características do Conjunto de Teste

O conjunto de teste compreendeu 47 706 janelas derivadas dos ficheiros de origem reservados para avaliação. A distribuição das classes refletiu cenários reais de monitorização rodoviária, caracterizados por um acentuado desequilíbrio:

- *none*: 45 911 janelas (96,2%)

- *pothole*: 980 janelas (2,1%)
- *speed_bump*: 815 janelas (1,7%)

Esta distribuição evidencia o desafio associado à deteção de eventos de perigo raros em condições de condução predominantemente normais, uma característica típica das aplicações de monitorização rodoviária [3], [21].

4.2. Resultados dos Modelos Clássicos de Aprendizagem Automática

Foram avaliados nove modelos clássicos de aprendizagem automática no conjunto de teste. A Tabela 1 apresenta as métricas de desempenho abrangentes para todos os modelos, ordenadas de acordo com a pontuação F1 com média macro.

Tabela 1 - Desempenho dos Modelos Clássicos de Aprendizagem Automática

Modelo	Exatidão	Exatidão balanceada	F1 Macro	F1 Buraco (Pothole)	F1 Lomba (speed_bump)	Tempo de Treino (s)
Hist Gradient Boosting	0.9397	0.5365	0.4475	0.0197	0.3542	2.0
Gradient Boosting	0.9344	0.5267	0.4342	0.0198	0.3173	135.3
Random Forest	0.9304	0.5260	0.4281	0.0195	0.3011	2.4
Rede Neuronal MLP	0.9213	0.5362	0.4192	0.0193	0.2799	152.6
Extra Trees	0.9131	0.5365	0.4099	0.0195	0.2560	1.0
SVM com núcleo RBF	0.8980	0.5542	0.4006	0.0190	0.2373	11.7
Regressão Logística	0.8704	0.5395	0.3831	0.0376	0.1817	0.8
KNN	0.8929	0.5006	0.3828	0.0192	0.1865	0.03
Árvore de Decisão	0.8566	0.5122	0.3641	0.0172	0.1528	0.5

4.2.1. Padrões Globais de Desempenho

O *Histogram-based Gradient Boosting* obteve a pontuação F1 macro mais elevada (0.4475), seguido pelo *Gradient Boosting* (0.4342) e pelo *Random Forest* (0.4281). Estes métodos de *ensemble* baseados em árvores superaram de forma consistente as restantes famílias algorítmicas, confirmando resultados anteriores que indicam que estes métodos apresentam desempenho superior em dados sensoriais tabulares [3], [11], [17].

Todos os modelos alcançaram uma exatidão global elevada (>85%), refletindo a predominância da classe *none*. No entanto, os valores de exatidão balanceada (0.50 - 0.55) evidenciaram alguma dificuldade na deteção das classes minoritárias, com o desempenho a aproximar-se de uma classificação aleatória quando as classes são ponderadas de forma equitativa.

4.2.2. Desempenho por Classe

A Figura 1 ilustra a exatidão global e a pontuação F1 macro dos nove modelos clássicos de aprendizagem automática, ordenados de acordo com a pontuação F1 macro, observando que o *Histogram Gradient Boosting* alcançou a pontuação F1 macro mais elevada, com um valor de 0.4475, bem como a maior exatidão, atingindo 0.9397.

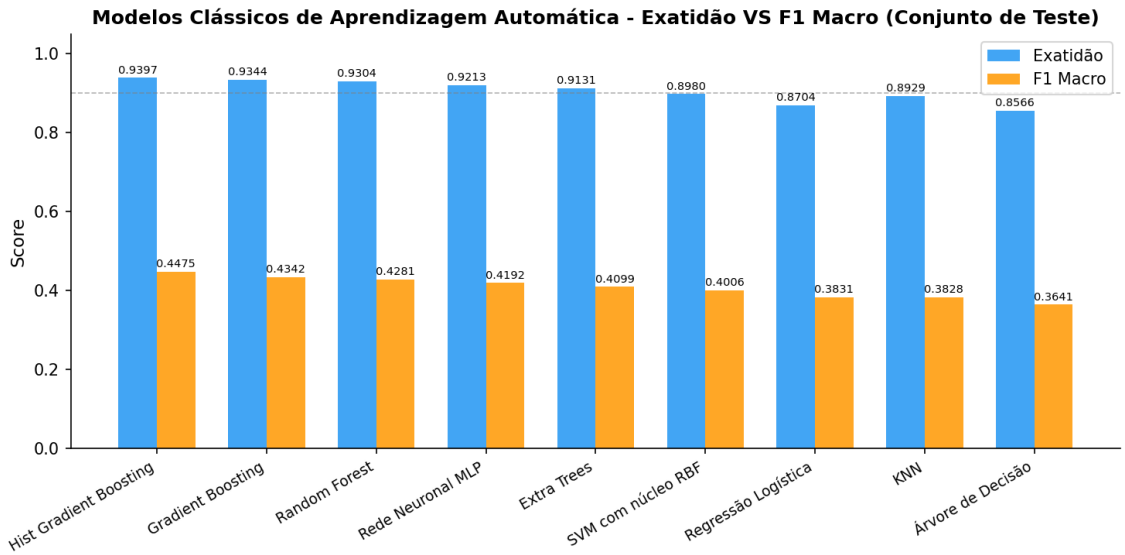

Figura 1 - Modelos Clássicos de Aprendizagem Automática: Exatidão VS F1 Macro

A Figura 2 ilustra as pontuações F1 por classe para a deteção de buracos e lombas em todos os modelos clássicos, ordenados de acordo com a pontuação F1 específica por classe. Os resultados demonstram que as pontuações F1 relativas à deteção de lombas

foram substancialmente superiores às pontuações F1 relativas à deteção de buracos em todos os modelos.

No que respeita à deteção de buracos, todos os modelos clássicos apresentaram um desempenho extremamente reduzido, com pontuações F1 variando entre 0.0172, no caso da Árvore de Decisão, e 0.0376, no caso da Regressão Logística. A maior revocação para a classe buraco foi de 0.0204, obtida pela Regressão Logística, indicando que menos de 3% dos buracos reais foram corretamente identificados. Este desempenho severamente limitado sugere que as características extraídas e/ou a estratégia de treino adotada foram insuficientes para captar características de sinal específicas dos buracos na presença de um desequilíbrio extremo entre classes [21], [22].

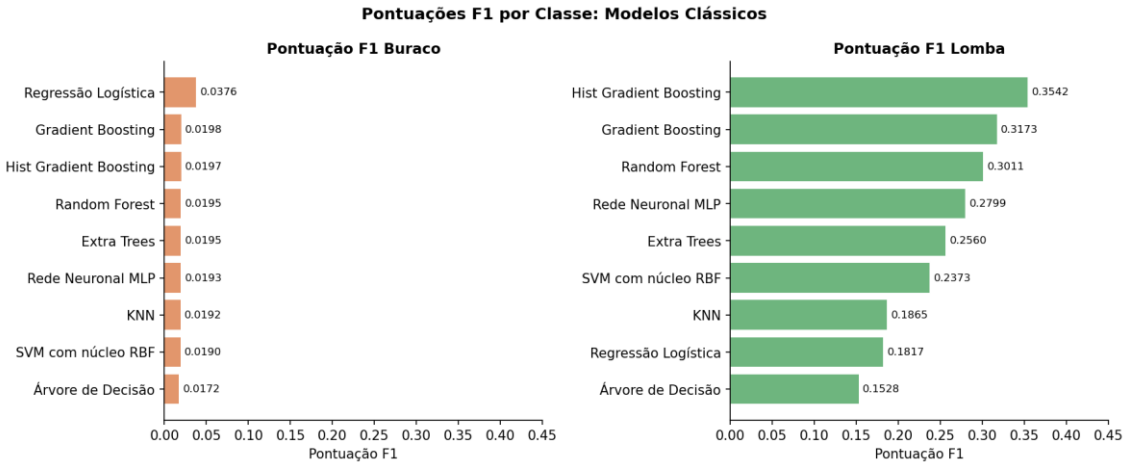

Figura 2 - Pontuações F1 por Classe: Modelos Clássicos

4.2.3. Eficiência Computacional

Os tempos de treino variaram substancialmente entre os modelos. O KNN foi o mais rápido (0.03 s), seguido pela Regressão Logística (0.8 s) E pelo Extra Trees (1,0 s). O *Gradient Boosting* (135,3 s) e a Rede Neuronal MLP (152,6 s) exigiram tempos de treino significativamente superiores. O *Histogram-based Gradient Boosting* alcançou um desempenho competitivo com apenas 2,0 s de tempo de treino, demonstrando uma excelente eficiência computacional [16].

4.3. Resultados da Otimização de Hiperparâmetros do *Random Forest*

O *Random Forest* foi selecionado para otimização de hiperparâmetros devido ao seu forte desempenho de referência, à sua interpretabilidade e à sua ampla adoção em tarefas de classificação baseadas em dados de acelerómetros [11], [17], [34].

4.3.1. Resultado da Otimização

A Figura 3 apresenta a distribuição das pontuações F1 macro obtidas em validação cruzada ao longo de 40 configurações aleatórias de hiperparâmetros. Recorrendo à pesquisa aleatória com validação cruzada de 5 partições, a configuração ótima, detalhada na Secção 3.4.2, alcançou uma pontuação F1 macro de 0.9456 em validação cruzada no conjunto de treino e uma pontuação F1 macro de 0.4498 no conjunto de teste. A diferença substancial entre o desempenho em validação cruzada e o desempenho no teste, 0.9456 versus 0.4498, indica a ocorrência de sobreajustamento à distribuição dos dados de treino, provavelmente devido a diferenças nas características dos ficheiros de origem entre os conjuntos de treino e de teste [25]. As 40 configurações de hiperparâmetros avaliadas, ordenadas por desempenho, encontram-se detalhadas no Anexo B – Resultados Completos da Otimização de Hiperparâmetros do *Random Forest*.

Importa salientar que a validação cruzada da pesquisa aleatória foi realizada sobre o conjunto de treino previamente reequilibrado, pelo que a pontuação F1 macro obtida em validação cruzada pode estar otimista. Assim, este valor deve ser interpretado sobretudo como critério interno de seleção de hiperparâmetros, não como estimativa final de generalização. A avaliação no conjunto de teste independente, mantido desequilibrado e separado por grupos, constitui a referência principal para interpretar o desempenho real do modelo.

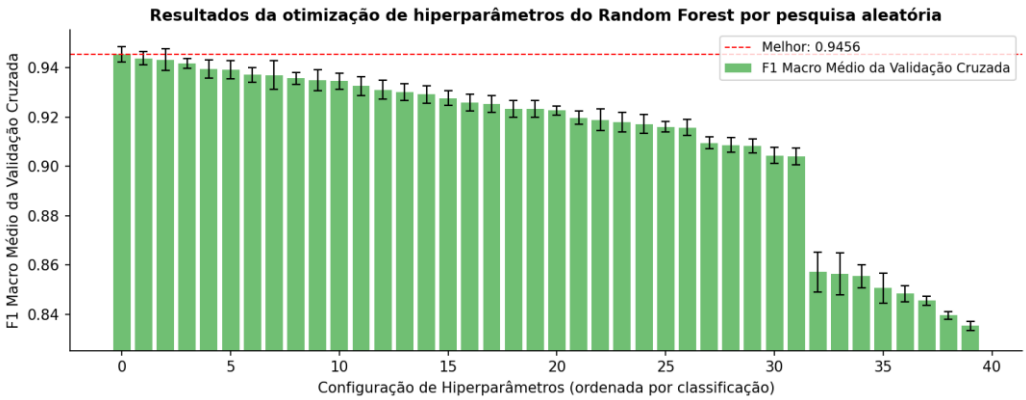

Figura 3 - Resultados da otimização de hiperparâmetros do Random Forest por pesquisa aleatória

4.3.2. Desempenho detalhado da classificação

A Tabela 2 apresenta o relatório de classificação detalhado para o modelo *Random Forest* otimizado. O modelo otimizado alcançou um desempenho excelente na classe none, com uma pontuação F1 de 0.9708, e um desempenho moderado na deteção de *speed_bump*, com uma pontuação F1 de 0.3590 e um *recall* de 0.5742. Contudo, a deteção de *pothole* permaneceu comprometida, com uma pontuação F1 de 0.0195 e um *recall* de 0.0102, uma vez que apenas 10 dos 980 buracos foram corretamente identificados.

Tabela 2 - Relatório de Classificação do Random Forest Otimizado

Classe	**Precisão**	***Recall***	**Pontuação F1**	**Suporte**
None	0.9713	0.9704	0.9708	45,911
Pothole	0.2222	0.0102	0.0195	980
speed_bump	0.2612	0.5742	0.3590	815
Média macro	**0.4849**	**0.5183**	**0.4498**	**47,706**
Média ponderada	**0.9438**	**0.9439**	**0.9409**	**47,706**

4.3.3. Análise da Matriz de Confusão

A Figura 4 apresenta a matriz de confusão do modelo *Random Forest* otimizado, apresentando tanto as contagens absolutas como os valores de *recall* normalizados por linha. Os resultados indicam que o modelo classificou corretamente 97,04% das janelas normais, mas não identificou 99% dos buracos.

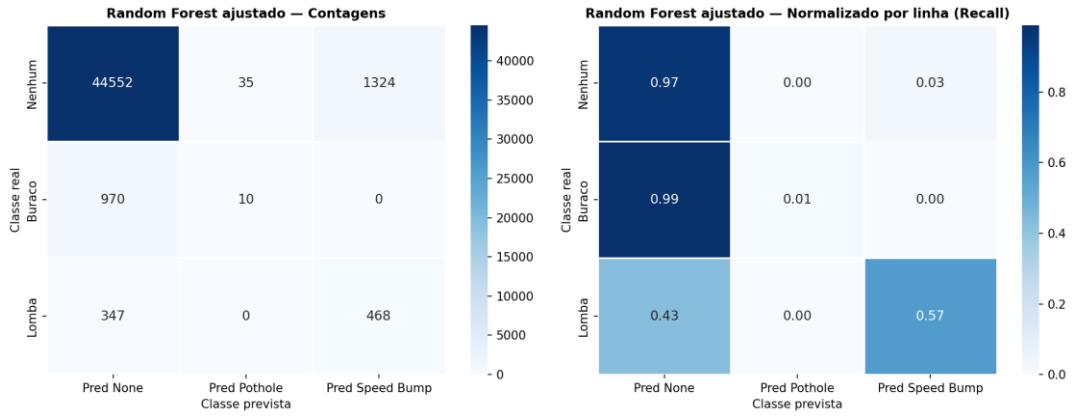

Figura 4 - Matriz de Confusão do Random Forest Otimizado

A Tabela 3 apresenta a matriz de confusão do modelo *Random Forest* otimizado, incluindo contagens absolutas e valores de *recall* normalizados por linha. A matriz revela vários padrões críticos no desempenho de classificação do modelo. Em primeiro lugar, a deteção de *pothole* revelou uma falha quase completa, uma vez que 970 dos 980 buracos, correspondentes a 99,0%, foram incorretamente classificados como none. Em segundo lugar, 347 das 815 lombas, ou 42,6%, foram igualmente incorretamente classificadas como *none*, enquanto 468 lombas, correspondentes a 57,4%, foram corretamente identificadas. Importa salientar que nenhuma lomba foi incorretamente classificada como *pothole*, sugerindo características de sinal distintas. Além disso, o modelo gerou 1 324 predições falsas de *speed_bump*, representando 2,9% das amostras da classe *none*, o que indica uma tendência para sobre prever lombas em situações de incerteza.

Globalmente, estes padrões sugerem que o modelo aprendeu a distinguir lombas de condições rodoviárias normais com sucesso moderado, mas não conseguiu aprender características discriminativas para buracos, provavelmente devido à representação insuficiente de buracos nos dados de treino e/ou à expressividade inadequada das características [21], [22].

Tabela 3 - Matriz de Confusão do Random Forest Otimizado

	Previsto como none	**Previsto como pothole**	**Previsto como speed_bump**
Verdadeiro none	44 552	35	1 324
Verdadeiro pothole	970	10	0
Verdadeiro speed_bump	347	0	468

4.3.4. Análise da Importância das Caraterísticas

A Figura 5 apresenta as 20 características mais importantes identificadas pelo modelo *Random Forest* otimizado, medidas pela diminuição média da impureza. A análise da importância das características revela que as características derivadas da magnitude, incluindo *mag_rms*, *mag_max*, *mag_abs_mean*, *mag_energy*, *mag_median* e *mag_mean*, ocupam 6 das 10 primeiras posições, indicando que a magnitude global da aceleração é altamente informativa para a deteção de perigos [3], [30]. As características de aceleração vertical, como *accZ_energy*, *accZ_abs_mean*, *accZ_rms*, *accZ_median*, *accZ_mean* e *accZ_max*, também estão fortemente representadas, o que é consistente com a expectativa de que os perigos rodoviários induzam principalmente movimento vertical no veículo [10], [2]. As características de energia e RMS surgem frequentemente entre as variáveis mais importantes, sugerindo que a potência e a amplitude do sinal são discriminadores críticos [14]. Em contraste, as características da magnitude do *jerk*, incluindo *jerk_mag_abs_mean* e *jerk_mag_iqr*, surgem em posições inferiores no ranking, indicando que a informação relativa à taxa de variação é menos discriminativa do que as características de magnitude absoluta para esta tarefa. As características de correlação entre eixos também apresentaram importância limitada, com *corr_accY_accZ*, *corr_accX_accZ* e *corr_accX_accY* classificadas nas 37.ª, 53.ª e 54.ª posições, respetivamente, não sendo, por isso, apresentadas na tabela. Globalmente, estes resultados estão alinhados com trabalhos anteriores que enfatizam a importância das características de magnitude e de aceleração vertical na deteção de anomalias rodoviárias.

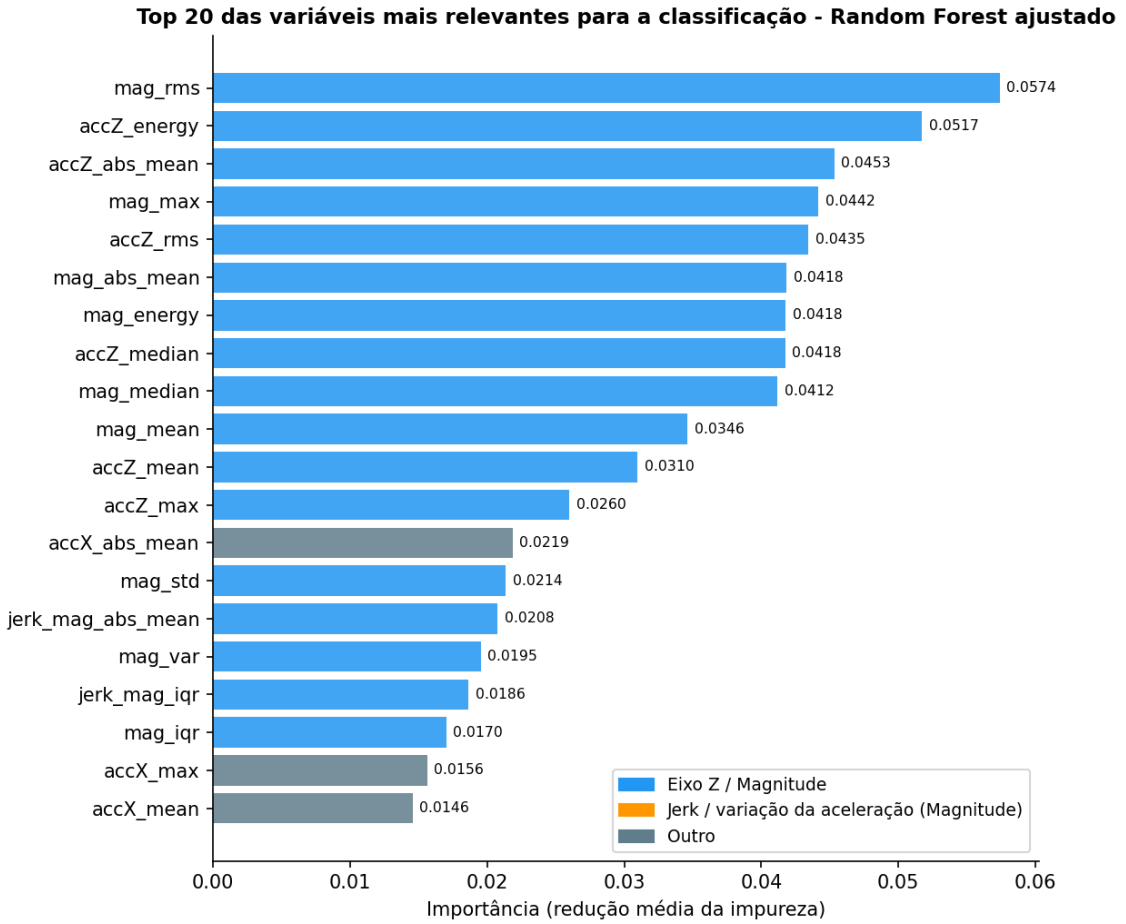

Figura 5 - Top 20 das variáveis mais relevantes para a classificação - Random Forest ajustado

4.4. Resultados dos Modelos de Aprendizagem Profunda

Foram avaliadas três arquiteturas de aprendizagem profunda: 1D CNN, LSTM e CNN+LSTM. A Tabela 4 apresenta as métricas de desempenho abrangentes.

Tabela 4 - Desempenho dos Modelos de Aprendizagem Profunda

Modelo	**Exatidão**	**Exatidão balanceada**	**F1 Macro**	**F1 Buraco**	**F1 Lomba**	**Tempo de treino(s)**	***Epoch***
1D CNN	0.9030	0.6143	0.5108	0.3566	0.2277	443	11
LSTM	0.9235	0.5429	0.4245	0.0190	0.2947	6,483	24
CNN + LSTM	0.9106	0.5443	0.4145	0.0352	0.2558	1,077	13

4.4.1. Comparação Global de Desempenho

A Figura 6 apresenta a exatidão e a pontuação F1 macro das três arquiteturas de aprendizagem profunda. Entre todos os modelos avaliados, incluindo abordagens clássicas e de aprendizagem profunda, a 1D CNN alcançou a pontuação F1 macro mais elevada, atingindo 0.5108, bem como a melhor exatidão balanceada, com um valor de 0.6143. Isto representa uma melhoria relativa de 13,6% face ao melhor modelo clássico, o *Random Forest* otimizado, que obteve uma pontuação F1 macro de 0.4498. A exatidão balanceada mais elevada indica um desempenho superior quando todas as classes são ponderadas de forma equitativa [19], [20]. Em comparação, os modelos LSTM e CNN+LSTM alcançaram pontuações F1 macro de 0.4245 e 0.4145, respetivamente, valores comparáveis aos dos melhores modelos clássicos, mas substancialmente inferiores ao da 1D CNN. Isto sugere que as arquiteturas puramente recorrentes e os modelos híbridos não exploraram eficazmente as dependências temporais nesta tarefa, possivelmente devido ao comprimento relativamente curto da janela, de 50 amostras, o que poderá ter limitado o contexto temporal de longo alcance [36], [37].

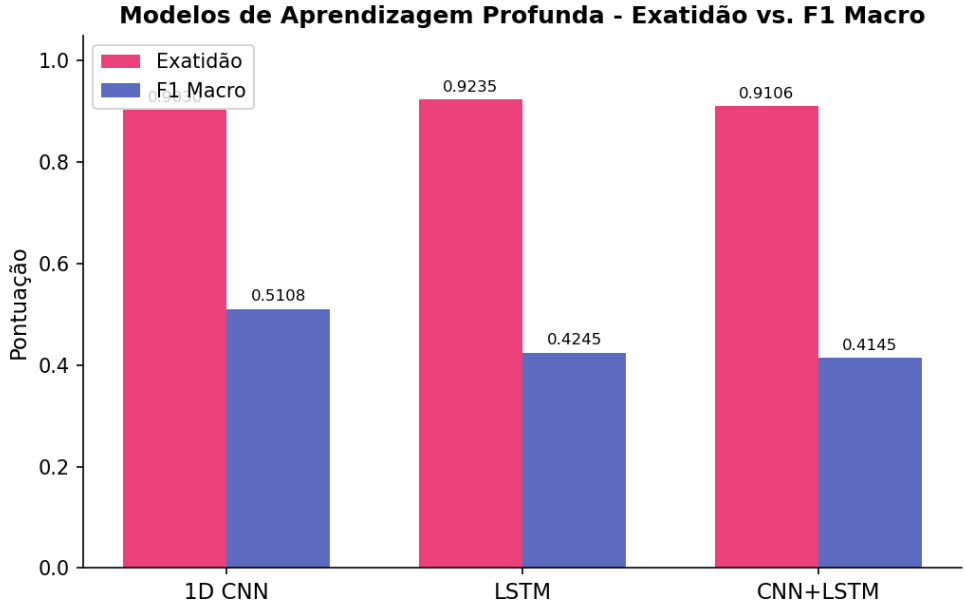

Figura 6 - Modelos de Aprendizagem Profunda - Exatidão VS F1 Macro

4.4.2. Desempenho por Classe

Para a deteção de buracos, a *1D CNN* alcançou uma melhoria substancial em comparação com todos os modelos clássicos, atingindo uma pontuação F1 de 0.3566, enquanto o melhor valor de F1 obtido entre os modelos clássicos para esta classe foi

de apenas 0.0376 (Regressão Logística). A 1D CNN alcançou uma precisão de 0.7419 e *recall* de 0.2347 para a classe *pothole*, identificando corretamente aproximadamente 23% dos buracos com elevada confiança. Isto representa uma melhoria de aproximadamente 9,5 vezes face ao melhor valor obtido entre os modelos clássicos nesta classe, valor próximo de uma ordem de grandeza, e demonstra a eficácia das características convolucionais aprendidas na captação de padrões de sinal específicos associados a buracos [18], [19]. Em contraste, os modelos LSTM e CNN+LSTM apresentaram um desempenho reduzido na deteção de buracos, com pontuações F1 de 0.0190 e 0.0352, respetivamente, valores comparáveis aos dos modelos clássicos. Isto sugere que as arquiteturas recorrentes, por si só, são insuficientes para a deteção de buracos sem uma extração eficaz de características [36].

Para a deteção de lombas, o desempenho variou mais entre os modelos de aprendizagem profunda. A LSTM alcançou a pontuação F1 mais elevada para a classe *speed_bump* entre os modelos profundos, com um valor de 0.2947 e um *recall* de 0.6712. A 1D CNN obteve uma pontuação F1 de 0.2277, com *recall* de 0.6871, enquanto a CNN+LSTM alcançou uma pontuação F1 de 0.2558, com um *recall* de 0.6798. Todos os modelos de aprendizagem profunda obtiveram valores de *recall* para a classe *speed_bump* superiores a 67%, indicando uma elevada sensibilidade a eventos de lomba [36], [37]. Contudo, a precisão para a classe *speed_bump* manteve-se baixa em todos os modelos profundos, variando entre 0.1365 e 0.1888, o que indica taxas elevadas de falsos positivos. Este resultado contrasta com os modelos clássicos, que alcançaram maior precisão na deteção de lombas, atingindo até 0.2612 no caso do *Random Forest* otimizado, embora à custa de um *recall* inferior.

4.4.3. Dinâmicas de Treino

As Figura 7 a Figura 9 ilustram as curvas de perda e exatidão de treino e validação para os três modelos de aprendizagem profunda. A Figura 7 apresenta o histórico de treino da 1D CNN, que convergiu rapidamente após 11 *epochs*, com uma redução da taxa de aprendizagem de 0.001 para 0.0005. A sua perda final de validação foi de aproximadamente 0.306, enquanto a exatidão de validação atingiu 0.856. As curvas de treino evidenciam uma convergência suave, com sobreajustamento mínimo, indicando uma regularização eficaz [19]. A Figura 8 apresenta o histórico de treino da LSTM, mostrando que este modelo necessitou de 24 *epochs* para convergir, correspondendo

à duração de treino mais longa entre os modelos de aprendizagem profunda. A sua taxa de aprendizagem foi reduzida duas vezes, de 0.001 para 0.0005 e, posteriormente, para 0.00025. A perda final de validação foi de aproximadamente 0.162, com a exatidão de validação a atingir 0.925. As curvas de treino demonstram uma melhoria gradual, com alguma oscilação nas métricas de validação, sugerindo sensibilidade à taxa de aprendizagem e à amostragem por lotes [36]. A Figura 9 apresenta o histórico de treino do modelo híbrido CNN+LSTM, que convergiu após 13 *epochs*, com uma redução da taxa de aprendizagem de 0.001 para 0.0005. O modelo alcançou uma perda final de validação de aproximadamente 0.184 e uma exatidão de validação de 0.911, evidenciando um comportamento de convergência intermédio entre a 1D CNN e a LSTM, com desempenho de validação estável [21], [37]. Apesar de terem alcançado uma exatidão de validação superior, os modelos LSTM e CNN+LSTM não traduziram esse desempenho num desempenho superior no conjunto de teste, particularmente na deteção de buracos. Esta discrepância sugere que o desempenho de validação na distribuição de treino reequilibrada não prevê de forma fiável o desempenho na distribuição de teste desequilibrada [22].

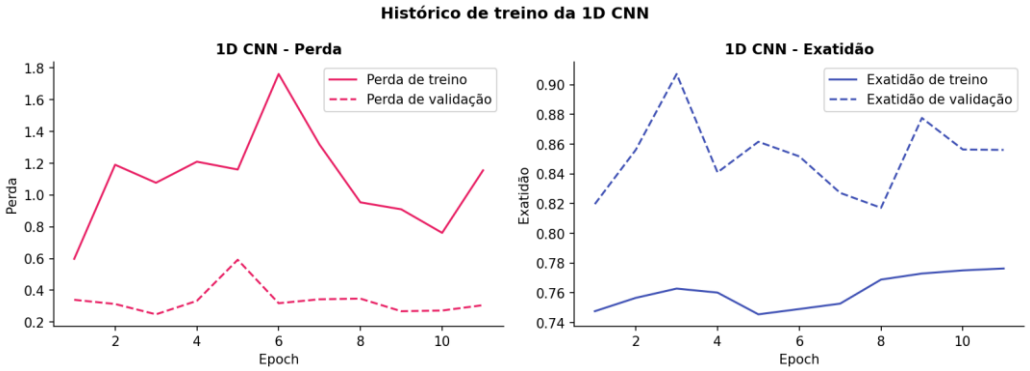

Figura 7 - Histórico de treino da 1D CNN

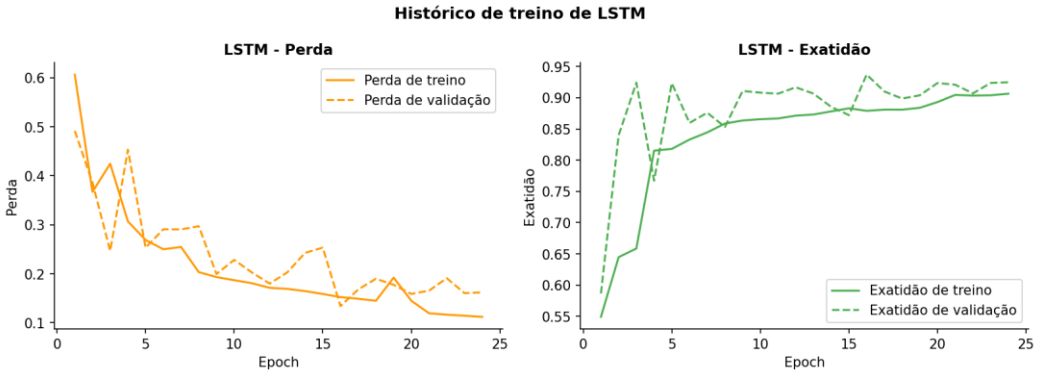

Figura 8 - Histórico de treino de LSTM

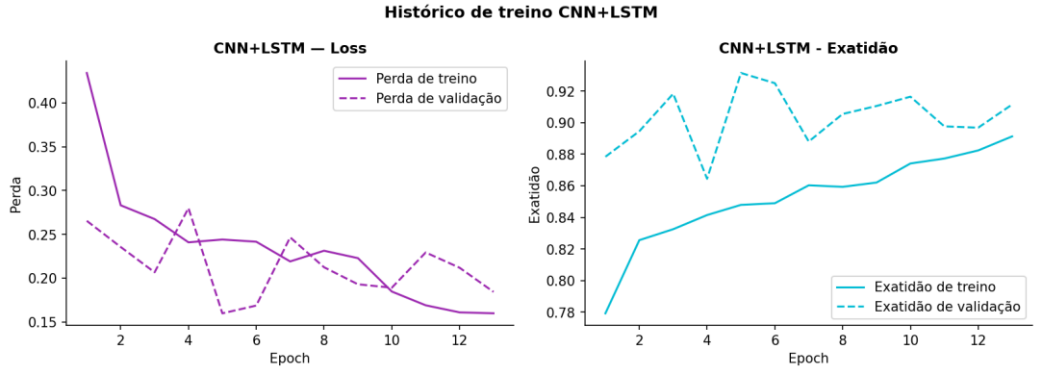

Figura 9 - Histórico de treino CNN+LSTM

4.4.4. Custo Computacional

Os tempos de treino variaram substancialmente entre os modelos de aprendizagem profunda. A 1D CNN foi a mais eficiente (443 s), seguida pela CNN+LSTM (1 077 s) e pela LSTM (6 483 s). O longo tempo de treino da LSTM – aproximadamente 2 horas – reflete o custo computacional das operações recorrentes e o maior número de épocas necessário para a convergência [36].

Em comparação com os modelos clássicos, as abordagens de aprendizagem profunda exigiram tempos de treino substancialmente superiores, variando a diferença consoante o modelo considerado. Em termos gerais, uma diferença de uma a duas ordens de grandeza corresponde aproximadamente a um aumento entre 10 e 100 vezes. Contudo, o desempenho superior da 1D CNN, particularmente na deteção de buracos, poderá justificar o aumento do custo computacional em aplicações nas quais a deteção de classes minoritárias é crítica [19], [21].

4.5. Análise Comparativa entre Todos os Modelos

A Figura 10 apresenta uma comparação abrangente das pontuações F1 macro entre os 12 modelos avaliados, incluindo nove modelos clássicos de aprendizagem automática e três modelos de aprendizagem profunda. A 1D CNN alcançou a pontuação F1 macro mais elevada, com um valor de 0.5108, seguida *pelo Histogram Gradient Boosting*, que obteve uma pontuação F1 macro de 0.4475. Na figura, os modelos de aprendizagem profunda são apresentados a vermelho, enquanto os modelos clássicos são apresentados a azul.

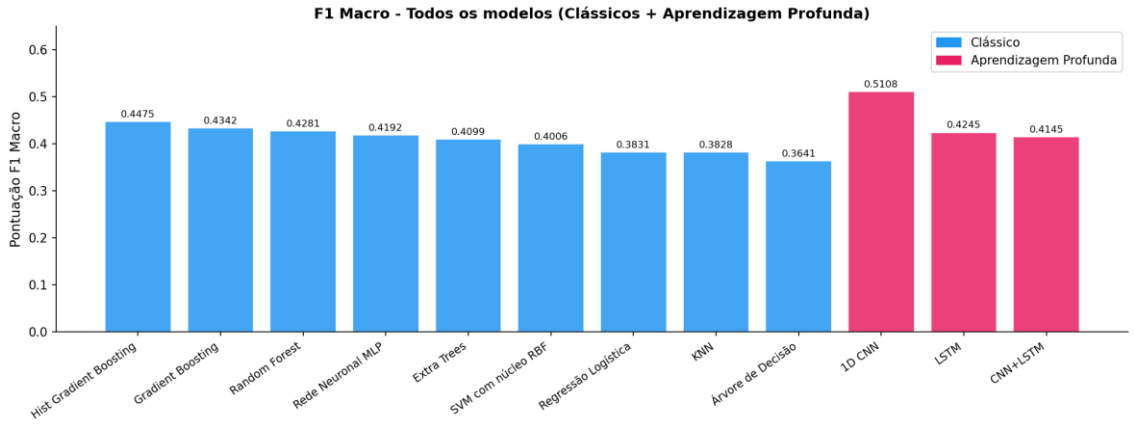

Figura 10 - F1 Macro - Todos os modelos (Clássicos + Aprendizagem Profunda)

4.5.1. Classificação Global

Os cinco melhores modelos segundo a pontuação F1 macro foram: 1. 1D CNN: 0.5108 2. *Random Forest* otimizado: 0.4498 3. *Hist Gradient Boosting*: 0.4475 4. *Gradient Boosting*: 0.4342 5. *Random Forest* predefinido: 0.4281.

A superioridade da 1D CNN é evidente, alcançando uma pontuação F1 macro 13,6% superior à do melhor modelo clássico. Entre os modelos clássicos, os *ensembles* baseados em árvores (*Hist Gradient Boosting*, *Gradient Boosting* e *Random Forest*) superaram consistentemente as restantes famílias algorítmicas [3], [11], [17].

4.5.2. Equilíbrios entre a Deteção de Buracos e de Lombas

A Figura 11 ilustra o *recall* por classe para a deteção de buracos e lombas nos nove modelos clássicos, revelando um compromisso crítico. Os modelos clássicos alcançaram um *recall* elevado para lombas, atingindo até 0.7325, mas extremamente baixa para buracos, permanecendo abaixo de 0.03 em todos os modelos. Em

comparação, a 1D CNN alcançou um *recall* moderado para buracos, com um valor de 0.2347, e um valor elevado para lombas, atingindo 0.6871. Os modelos LSTM e CNN+LSTM apresentaram um *recall* reduzido para buracos, inferior a 0.02, mas um elevado para lombas, superior a 0.67. Este padrão sugere que a deteção de buracos requer representações de características mais sofisticadas do que a deteção de lombas, as quais as características estatísticas clássicas e as arquiteturas recorrentes têm dificuldade em captar [18], [19].

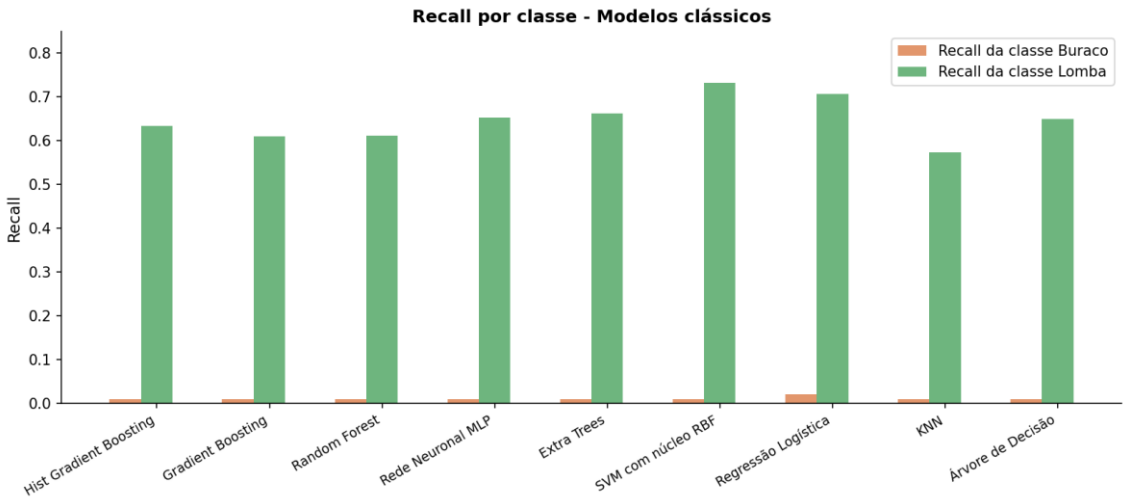

Figura 11 - Recall por classe - Modelos Clássicos

4.5.3. Equilíbrios de Eficiência Computacional

A Figura 12 apresenta a comparação dos tempos de treino entre os 12 modelos, numa escala logarítmica, revelando uma variação substancial entre as abordagens clássicas e as abordagens de aprendizagem profunda. Os modelos de aprendizagem profunda exigiram tempos de treino substancialmente superiores, variando entre 443 e 6 483 segundos, em comparação com os modelos clássicos, cujos tempos variaram entre 0.03 e 153 segundos. Os modelos mais rápidos foram o KNN, com 0.03 segundos, a Regressão Logística, com 0.8 segundos, e o *Extra Trees*, com 1 segundo. Os modelos com tempos de treino moderados incluíram o *Histogram Gradient Boosting*, com 2 segundos, o *Random Forest*, com 2.4 segundos, e o SVM RBF, com 11.7 segundos. Os modelos mais lentos incluíram o *Gradient Boosting*, com 135.3 segundos, o MLP, com 152.6 segundos, a 1D CNN, com 443 segundos, e o CNN+LSTM, com 1 077 segundos, enquanto o LSTM foi o modelo mais lento, exigindo 6 483 segundos. Globalmente, a 1D

CNN alcançou o melhor desempenho com um tempo de treino moderado de 443 segundos, representando um compromisso favorável entre desempenho e eficiência em comparação com o LSTM, que exigiu 6 483 segundos, e demonstrando um desempenho F1 macro comparável ao do *Gradient Boosting*, que exigiu 135.3 segundos [19], [16].

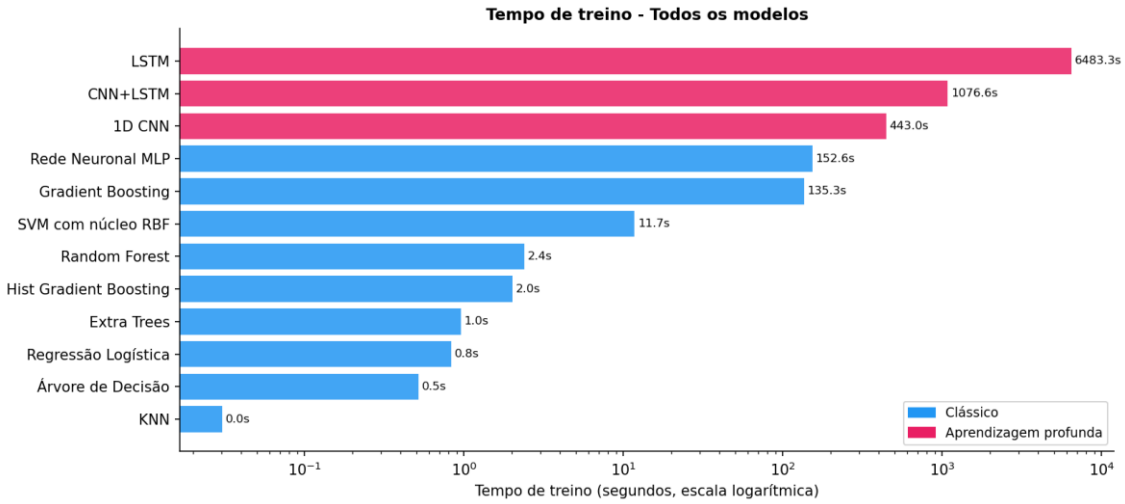

Figura 12 - Tempo de treino - Todos os modelos

4.5.4. Robustez dos Resultados por *Bootstrap*

Com o objetivo de avaliar a estabilidade dos resultados obtidos, foi realizada uma análise *bootstrap* com 5000 reamostragens das predições do conjunto de teste. Esta análise permitiu estimar intervalos de confiança a 95% para as principais métricas de desempenho, em particular a pontuação F1 macro, a exatidão balanceada e as métricas específicas das classes minoritárias [39], [40].

A Tabela 5 apresenta os intervalos de confiança a 95% para os modelos com melhor desempenho global e para os modelos mais relevantes na comparação entre abordagens clássicas e de aprendizagem profunda.

Tabela 5 - Intervalos de confiança a 95% por bootstrap para os principais modelos

Modelo	F1 macro	IC 95% F1 macro	Exatidão balanceada	IC 95% Exatidão balanceada	F1 pothole	IC 95% F1 pothole	F1 speed_bump	IC 95% F1 speed_bump
1D CNN	0.5108	[0.4989; 0.5222]	0.6143	[0.6003; 0.6278]	0.3566	[0.3234; 0.3891]	0.2277	[0.2169; 0.2380]
Random Forest otimizado	0.4498	[0.4421; 0.4577]	0.5183	[0.5070; 0.5298]	0.0195	[0.0079; 0.0327]	0.3590	[0.3399; 0.3793]
Hist Gradient Boosting	0.4475	[0.4404; 0.4551]	0.5365	[0.5254; 0.5480]	0.0197	[0.0080; 0.0332]	0.3542	[0.3364; 0.3726]
Gradient Boosting	0.4342	[0.4273; 0.4415]	0.5266	[0.5151; 0.5380]	0.0198	[0.0080; 0.0332]	0.3173	[0.3005; 0.3338]
Random Forest	0.4281	[0.4214; 0.4350]	0.5260	[0.5147; 0.5376]	0.0195	[0.0079; 0.0328]	0.3011	[0.2860; 0.3166]

Os resultados da análise *bootstrap* confirmam a superioridade da arquitetura 1D CNN em termos de desempenho global. A pontuação F1 macro da 1D CNN foi de 0.5108, com intervalo de confiança a 95% entre 0.4989 e 0.5222, não se sobrepondo aos intervalos dos restantes modelos principais. Este resultado é consistente com uma vantagem real da 1D CNN, embora a magnitude exata da incerteza estimada possa estar subestimada, uma vez que as janelas do conjunto de teste não são totalmente independentes entre si (ver discussão na secção 3.6.5) [39], [40]. Esta conclusão está também alinhada com trabalhos anteriores que demonstram a adequação das redes neuronais convolucionais à aprendizagem automática de padrões relevantes em sinais de acelerómetros e vibração rodoviária [18], [19].

A análise por classe revela, contudo, diferenças importantes entre os modelos. A 1D CNN apresentou o melhor desempenho na classe *pothole*, com uma pontuação F1 de 0.3566 e intervalo de confiança entre 0.3234 e 0.3891, muito acima dos modelos clássicos avaliados. Este resultado confirma que a aprendizagem profunda foi mais eficaz na deteção de buracos, a classe minoritária mais difícil do problema, reforçando a importância de métricas sensíveis ao desempenho por classe em cenários desequilibrados [21], [22].

Por outro lado, o *Random Forest* otimizado e o *Hist Gradient Boosting* apresentaram os melhores resultados na classe *speed_bump*, com pontuações F1 de 0.3590 e 0.3542, respetivamente. Este comportamento é consistente com a capacidade dos ensembles baseados em árvores para modelar relações não lineares em características estatísticas extraídas de sinais de acelerómetros [3], [11], [17]. No entanto, os respetivos intervalos de confiança apresentam sobreposição, pelo que não é prudente afirmar uma superioridade inequívoca de um destes modelos sobre o outro nesta classe.

A análise *bootstrap* reforça ainda a limitação da exatidão global como métrica principal. Embora o *Random Forest* otimizado tenha apresentado uma exatidão global elevada, o seu desempenho na classe *pothole* permaneceu muito reduzido, com uma pontuação F1 de apenas 0.0195. Assim, os resultados confirmam que, em cenários fortemente desequilibrados, a seleção do melhor modelo deve privilegiar métricas como a pontuação F1 macro, a exatidão balanceada e o desempenho por classe, em vez da exatidão isolada [21], [22].

4.6. Síntese dos Principais Resultados

Os resultados experimentais permitiram identificar vários aspetos críticos. Em primeiro lugar, a 1D CNN demonstrou um desempenho superior na deteção de buracos, alcançando uma melhoria de uma ordem de grandeza na pontuação F1 da classe *pothole* (0.3566) em comparação com o melhor valor de F1-pothole obtido entre os modelos clássicos (0.0376, Regressão Logística), evidenciando o valor das representações aprendidas para eventos raros e subtis [18], [19]. Em segundo lugar, os *ensembles* clássicos baseados em árvores apresentaram um desempenho robusto na deteção de lombas, alcançando pontuações F1 até 0.3590, com um custo computacional substancialmente inferior ao dos modelos de aprendizagem profunda [3], [11], [17]. Em terceiro lugar, os modelos LSTM e CNN+LSTM não superaram a 1D CNN, sugerindo que as dependências temporais para além da janela de 50 amostras não foram críticas para esta tarefa, ou que estes modelos não conseguiram aprender eficazmente tais dependências [36], [37]. Em quarto lugar, o acentuado desequilíbrio entre classes permaneceu um desafio importante em todas as abordagens, sendo o conjunto de dados composto por 96,2% de amostras da classe *none*, 2,1% da classe *pothole* e 1,7% da classe *speed_bump*, e continuando a deteção de buracos a constituir a tarefa mais difícil para todos os modelos [21], [22]. Em quinto lugar, a análise da

importância das características demonstrou que as características de magnitude e de aceleração vertical dominaram as classificações do *Random Forest*, em consonância com o conhecimento do domínio relativo ao movimento veicular induzido por perigos rodoviários [10], [30]. Em sexto lugar, a análise *bootstrap* das predições reforçou a estabilidade das conclusões obtidas, demonstrando que a superioridade da 1D CNN em termos de F1 macro se manteve quando considerada a incerteza associada ao conjunto de teste [39], [40]. Em particular, a 1D CNN apresentou uma pontuação F1 macro de 0.5108, com intervalo de confiança a 95% entre 0.4989 e 0.5222, não se sobrepondo aos intervalos dos restantes modelos principais. Esta análise confirmou também que os modelos clássicos baseados em árvores, embora competitivos na deteção de lombas, continuaram a apresentar desempenho reduzido na classe *pothole*, reforçando a necessidade de métricas adequadas a cenários de desequilíbrio entre classes. Por fim, a diferença substancial entre o desempenho em validação cruzada e o desempenho no conjunto de teste, como no caso da diferença da pontuação F1 macro do *Random Forest* entre 0.9456 e 0.4498, indica uma generalização limitada entre ficheiros de origem e evidencia a dificuldade de construir modelos robustos para dados sensoriais heterogéneos [25].

Capítulo 5 – Discussão

Este capítulo interpreta os resultados experimentais, contextualiza-os no âmbito da literatura existente, examina as limitações metodológicas e explora as implicações para sistemas práticos de monitorização de perigos rodoviários.

5.1. Interpretação do Desempenho dos Modelos

5.1.1. Vantagens da Aprendizagem Profunda na Deteção de Buracos

A superioridade expressiva da 1D CNN na deteção de buracos, com uma pontuação F1 de 0.3566 em comparação com 0.0376, o melhor valor obtido entre os modelos clássicos para esta classe (Regressão Logística), representa o resultado mais significativo deste estudo. Esta diferença de desempenho pode ser atribuída a vários fatores. Em primeiro lugar, as camadas convolucionais permitem a aprendizagem automática de características ao extrair representações hierárquicas diretamente dos sinais brutos, captando potencialmente padrões temporais subtis e características de frequência que não são adequadamente representados por características estatísticas

manualmente definidas [18], [19]. Os buracos podem induzir picos de aceleração breves e de alta frequência, que se diluem quando agregados em estatísticas ao nível da janela, como a média, o desvio-padrão ou o RMS [30], [14]. Em segundo lugar, as convoluções 1D com tamanhos de núcleo reduzidos, tipicamente entre 3 e 7 amostras, conseguem detetar eventos de sinal localizados dentro da janela de 50 amostras, ao passo que as características estatísticas agregam informação ao longo de toda a janela [20]. Assim, se as assinaturas associadas aos buracos ocuparem apenas uma pequena fração da janela, por exemplo, entre 5 e 10 amostras, os filtros convolucionais podem isolar esses eventos, enquanto as características estatísticas podem ser dominadas por amostras não associadas a perigos [24]. Em terceiro lugar, as CNN profundas com múltiplas camadas convolucionais podem aprender representações multiescala, desde alterações instantâneas de granularidade fina até padrões mais amplos ao nível da janela [38]. Esta representação multiescala poderá ser crítica para distinguir buracos, que correspondem a impactos breves e abruptos, de lombas, que envolvem transições mais suaves e de maior duração, bem como de vibrações rodoviárias normais. Globalmente, estes resultados estão alinhados com trabalhos anteriores que demonstram a superioridade das CNN face à engenharia tradicional de características na monitorização da superfície rodoviária e no reconhecimento de atividades a partir de dados de acelerómetros [41].

5.1.2. Pontos Fortes dos Modelos Clássicos na Deteção de Lombas

Apesar das vantagens da aprendizagem profunda na deteção de buracos, os *ensembles* clássicos baseados em árvores alcançaram um desempenho competitivo na deteção de lombas, com o *Random Forest* otimizado a atingir uma pontuação F1 de 0.3590, em comparação com valores entre 0.2277 e 0.2947 obtidos pelos modelos de aprendizagem profunda. Isto sugere que as lombas produzem padrões de aceleração mais estereotipados e consistentes, adequadamente captados por características estatísticas [10], [2]. As lombas induzem, tipicamente, uma aceleração vertical sustentada, causada pelo movimento ascendente e descendente de maior duração à medida que o veículo atravessa a lomba, bem como alterações previsíveis na magnitude, refletidas em aumentos consistentes da magnitude da aceleração e da energia no eixo vertical. Podem também produzir padrões simétricos, com perfis de aceleração semelhantes para os eixos dianteiro e traseiro, os quais podem ser captados através de estatísticas ao nível da janela [30]. Estas características são bem

representadas por variáveis como *accZ_energy*, *mag_rms* e *accZ_abs_mean*, que ocuparam posições elevadas na análise de importância das características do *Random Forest* apresentada na Figura 5. Em contraste, os buracos induzem impactos breves e assimétricos, que podem não produzir estatísticas distintivas ao nível da janela [2]. A eficiência computacional dos modelos clássicos, com tempos de treino de 1 a 2 segundos para o *Random Forest* e o *Histogram Gradient Boosting*, torna-os atrativos para aplicações em que a deteção de lombas constitui o objetivo principal e os recursos computacionais são limitados [11], [17].

5.1.3. Desempenho Insuficiente das Arquiteturas Recorrentes

Os modelos LSTM e CNN+LSTM não superaram a 1D CNN, contrariando as expectativas baseadas em trabalhos anteriores que demonstram as vantagens das arquiteturas híbridas no reconhecimento de atividades [21], [37]. Vários fatores poderão explicar esta discrepância. Em primeiro lugar, o contexto temporal limitado da janela de 50 amostras, correspondente a 0.5 segundos a 100 Hz, poderá ser demasiado curto para beneficiar das capacidades de modelação temporal de longo alcance da LSTM [36]. Os eventos de perigo rodoviário são, tipicamente, instantâneos ou breves, abrangendo 10 a 30 amostras, e o contexto relevante poderá estar confinado à vizinhança imediata do evento, em vez de se encontrar distribuído ao longo de toda a janela [24]. Em segundo lugar, as LSTM requerem, geralmente, conjuntos de dados de treino maiores do que as CNN para aprenderem dinâmicas recorrentes eficazes. Embora o conjunto de treino tenha sido reequilibrado, poderá não ter fornecido exemplos suficientes de contextos diversos de buracos e lombas para treinar representações recorrentes robustas. Em terceiro lugar, o *dropout* recorrente, aplicado para prevenir o sobreajustamento, poderá ter prejudicado a capacidade da LSTM de aprender dependências de longo alcance, particularmente no caso dos eventos raros de buracos. Por fim, trabalhos recentes demonstraram que a modelação do contexto ao longo de múltiplas janelas consecutivas, ou seja, de dependências temporais entre janelas, pode melhorar substancialmente o desempenho [38]. Contudo, o presente estudo tratou cada janela de forma independente, limitando potencialmente a utilidade das arquiteturas recorrentes, que são mais adequadas à modelação de dependências sequenciais em escalas temporais mais longas. Globalmente, estes resultados sugerem que, em tarefas de classificação baseadas em eventos com janelas curtas, as

arquiteturas convolucionais poderão ser mais eficazes do que os modelos recorrentes ou híbridos, particularmente quando os dados de treino são limitados [19], [20].

5.2. Desequilíbrio entre Classes e Deteção de Classes Minoritárias

O acentuado desequilíbrio entre classes – 96,2% *none*, 2,1% *pothole* e 1,7% *speed_bump* – constituiu um desafio fundamental para todos os modelos, tendo a deteção de buracos revelado dificuldades particularmente significativas [21], [22].

5.2.1. Limitações da Estratégia de Reequilíbrio

A estratégia de reequilíbrio do conjunto de treino, que envolveu a sobre amostragem das classes minoritárias e a limitação da classe maioritária ao dobro do número máximo de amostras de perigo, melhorou o *recall* das classes minoritárias em comparação com o treino sobre a distribuição original desequilibrada, embora estes resultados não sejam apresentados. Contudo, persistiram diferenças substanciais de desempenho, particularmente na deteção de buracos. Várias estratégias alternativas poderão melhorar a deteção das classes minoritárias. As funções de perda ponderadas por classe podem atribuir penalizações de perda mais elevadas às classificações incorretas das classes minoritárias, aumentando a sensibilidade do modelo a eventos raros [21], [22], embora esta abordagem não tenha sido explorada sistematicamente no presente estudo. A *focal loss*, uma função de perda de entropia cruzada modificada que reduz o peso dos exemplos fáceis e concentra a aprendizagem nos exemplos difíceis e incorretamente classificados, tem demonstrado potencial na deteção de objetos com classes desequilibradas e poderá beneficiar a deteção de perigos raros [42]. Os métodos de sobre amostragem sintética, como o SMOTE, geram exemplos sintéticos das classes minoritárias por interpolação entre amostras existentes [23]. Contudo, a eficácia do SMOTE em dados de séries temporais permanece debatida, uma vez que as sequências interpoladas podem não preservar a coerência temporal. As abordagens de balanceamento por *ensemble*, como os métodos *Data Balanced Bagging Ensemble*, treinam múltiplos aprendizes base em subconjuntos equilibrados e agregam as respetivas predições. Ward et al. [21] demonstraram que o DBBE com aprendizes base Conv-LSTM alcançou mais de 90% de *recall* em dados de acelerómetros desequilibrados, sugerindo potencial para a deteção de perigos rodoviários. Por fim, formular a tarefa como um problema de deteção de anomalias, em que os modelos são

treinados principalmente com dados de estrada normal e utilizados para assinalar desvios, poderá ser mais eficaz do que a classificação multiclasse para eventos extremamente raros [35]. Abordagens baseadas em *one-class* SVM também têm demonstrado potencial na deteção de anomalias rodoviárias.

5.2.2. Considerações sobre as Métricas de Avaliação

A escolha das métricas de avaliação influencia substancialmente a seleção e a otimização dos modelos. Neste estudo, a pontuação F1 com média macro foi utilizada como métrica principal, uma vez que trata todas as classes de forma equitativa e é sensível ao desempenho das classes minoritárias [21]. Contudo, em sistemas operacionais de monitorização rodoviária, diferentes tipos de perigos podem ter pesos de importância distintos, dado que os buracos podem ser priorizados em relação às lombas para efeitos de planeamento da manutenção. Métricas alternativas poderão, portanto, refletir melhor as prioridades operacionais. A pontuação F1 ponderada atribui pesos específicos por classe com base na importância operacional ou no custo da classificação incorreta [22]. O *recall* com precisão fixa poderá ser mais adequado para aplicações críticas em termos de segurança, nas quais pode ser preferível detetar a maioria dos perigos, mantendo simultaneamente um limiar aceitável de precisão e limitando os falsos alarmes, em vez de otimizar apenas a pontuação F1. As métricas sensíveis ao custo modelam explicitamente os custos dos falsos negativos, como perigos não detetados, e dos falsos positivos, como inspeções desnecessárias, com o objetivo de otimizar a eficiência operacional [23].

5.3. Engenharia de Características VS Aprendizagem Automática de Características

O desempenho superior da 1D CNN na deteção de buracos evidencia as limitações da engenharia manual de características na representação de padrões de sinal complexos e subtis [18], [19], [14].

5.3.1. Limitações das Características Estatísticas

O conjunto de características de 58 dimensões, embora abrangente, poderá não captar adequadamente características críticas da aceleração induzida por buracos. As

características estatísticas agregam informação ao longo de toda a janela de 50 amostras, podendo diluir assinaturas breves associadas a buracos [24]. Os buracos podem induzir picos abruptos de aceleração com duração de apenas 5 a 10 amostras, os quais podem ser suavizados nas estatísticas ao nível da janela. Além disso, o conjunto de características incluiu apenas características no domínio temporal, omitindo representações no domínio da frequência, tais como coeficientes de Fourier, decomposições por onduletas ou características espectrais. Cong et al. [35] demonstraram que a decomposição por pacotes de onduletas melhorou substancialmente a deteção de buracos, sugerindo que as características no domínio da frequência captam informação complementar. Para além disso, as características estatísticas captam distribuições marginais, como a média e a variância, bem como correlações aos pares, mas não modelam interações de ordem superior ou relações não lineares entre eixos [14]. Assim, as assinaturas associadas a buracos poderão envolver acoplamentos complexos e não lineares entre as componentes de aceleração vertical e horizontal.

5.3.2. Vantagens das Representações Aprendidas

As redes neuronais convolucionais aprendem hierarquias de características específicas da tarefa através de retro propagação, otimizando as representações para o objetivo de classificação [19], [20]. Esta aprendizagem automática de características oferece várias vantagens. Em primeiro lugar, as características são aprendidas diretamente a partir dos dados, em vez de serem concebidas com base em pressupostos do domínio, permitindo potencialmente que o modelo descubra padrões inesperados [18], [41]. Em segundo lugar, múltiplas camadas convolucionais aprendem características em níveis crescentes de abstração, desde padrões de sinal de baixo nível até conceitos semânticos de alto nível [38]. Em terceiro lugar, os filtros convolucionais conseguem detetar padrões independentemente da sua posição dentro da janela, conferindo robustez à variabilidade temporal na ocorrência dos perigos. Contudo, as representações aprendidas sacrificam interpretabilidade em comparação com características manualmente definidas [14]. A análise da importância das características do *Random Forest* apresentada na Figura 5 fornece uma compreensão clara sobre quais características do sinal orientam a classificação, enquanto as ativações dos filtros das CNN são difíceis de interpretar sem técnicas especializadas de visualização.

5.4. Generalização entre Fontes de Dados Heterogéneas

A diferença substancial entre o desempenho em validação cruzada e o desempenho no teste, como a pontuação F1 macro do *Random Forest* de 0.9456 em comparação com 0.4498, indica uma generalização limitada entre ficheiros de origem [25]. Este desafio de generalização reflete a heterogeneidade do conjunto de dados combinado, que inclui múltiplas posições dos sensores, tais como *smartphones* em posições manuais ou montadas, sensores na suspensão do veículo colocados acima ou abaixo, e sensores montados no painel de instrumentos [28]. Inclui também diferentes taxas de amostragem, com dados a 100 Hz provenientes da IEEE DataPort e taxas variáveis da Kaggle e do PVS [27], [29]. Além disso, o conjunto de dados abrange diversos tipos de veículos, incluindo automóveis de passageiros e bicicletas [43], superfícies rodoviárias variadas, como asfalto e calçada, e diferentes características dos perigos, incluindo variações na profundidade e largura dos buracos, bem como na altura e no perfil das lombas [2].

5.4.1. Desvio de Domínio e Aprendizagem por Transferência

A divisão entre treino e teste baseada em grupos, embora previna a fuga de informação, cria um cenário desafiante de desvio de domínio, no qual os dados de teste têm origem em ficheiros que podem apresentar características sensoriais, dinâmicas veiculares e propriedades dos perigos diferentes das observadas nos dados de treino [25]. As abordagens de aprendizagem por transferência poderão contribuir para melhorar a generalização neste contexto. Uma estratégia possível consiste no pré-treino de modelos profundos em conjuntos de dados de acelerómetros extensos e diversificados, tais como referências de reconhecimento de atividades, seguido de ajuste fino em dados de perigos rodoviários, o que poderá melhorar a aprendizagem de características [41]. Técnicas de adaptação de domínio, incluindo treino adversarial de domínio ou normalização por lotes específica de domínio, poderão também reduzir a sensibilidade do modelo a características específicas de cada domínio [44]. Além disso, a aprendizagem multitarefa, na qual os modelos são treinados conjuntamente em tarefas relacionadas, como classificação da superfície rodoviária, deteção do tipo de veículo e deteção de perigos, poderá melhorar a robustez das características aprendidas.

5.4.2. Aumento de Dados

As técnicas de aumento de dados, amplamente utilizadas na classificação de imagens, poderão melhorar a generalização em dados de acelerómetro. Estas incluem a deformação temporal (*time warping*), que aplica alongamento ou compressão não linear às séries temporais de modo a simular diferentes velocidades do veículo [24], e a escala de magnitude, que ajusta as magnitudes da aceleração para simular diferentes sensibilidades dos sensores ou características da suspensão do veículo. A injeção de ruído também pode ser utilizada através da adição de ruído gaussiano para simular ruído sensorial e melhorar a robustez [22], enquanto o deslocamento aleatório das janelas (*window jittering*) altera aleatoriamente os limites das janelas para reduzir o sobreajustamento a alinhamentos temporais específicos. Contudo, é necessário assegurar que as amostras aumentadas permaneçam fisicamente plausíveis e não distorçam as assinaturas dos perigos.

5.5. Implicações Práticas para Sistemas de Monitorização Rodoviária

Os resultados deste estudo apresentam várias implicações para a conceção e implementação de sistemas de monitorização de perigos rodoviários baseados em acelerómetros.

5.5.1. Critérios de Seleção dos Modelos

Para aplicações centradas na deteção de buracos, a 1D CNN constitui a escolha mais adequada, oferecendo uma melhoria de uma ordem de grandeza na deteção de buracos, com uma pontuação F1 de 0.3566, em comparação com os modelos clássicos. O seu tempo de treino moderado, de 443 segundos, e a rápida inferência tornam-na viável para implementação em dispositivos de computação periférica (*edge devices*) ou em fluxos de processamento baseados na nuvem [19]. Para aplicações centradas na deteção de lombas, os *ensembles* clássicos baseados em árvores, como o *Random Forest* e o *Histogram Gradient Boosting*, oferecem desempenho competitivo, com pontuações F1 entre 0.3542 e 0.3590, exigindo simultaneamente um custo computacional substancialmente inferior, com tempos de treino de 1 a 2 segundos. Estes modelos são adequados para sistemas embebidos com recursos limitados [11], [17]. Para a deteção multiclasse equilibrada de perigos, a 1D CNN proporciona o melhor

desempenho global, alcançando uma pontuação F1 macro de 0.5108, devendo ser preferida quando a deteção de buracos e de lombas é igualmente importante.

5.5.2. Considerações de Implementação

O processamento em tempo real é suportado pela janela de 50 amostras com um passo de 25 amostras, permitindo a deteção de perigos em quase tempo real, com uma latência de 0,25 a 0,5 segundos a uma taxa de amostragem de 100 Hz. O tempo de predição da 1D CNN, que foi de 4.3 segundos para 47 706 janelas, corresponde a aproximadamente 0.09 ms por janela, possibilitando o processamento em tempo real em hardware moderno [19]. Contudo, a gestão de falsos positivos continua a ser importante, uma vez que a baixa precisão para as classes minoritárias utilizando a 1D CNN indica taxas substanciais de falsos positivos: para buracos, a precisão de 0.7419 implica cerca de 26 falsos alarmes em cada 100 alertas; para lombas, a precisão de 0.1365 implica cerca de 86 falsos alarmes em cada 100 alertas. Assim, os sistemas práticos poderão exigir agrupamento espacial, no qual as deteções são agregadas entre múltiplos veículos ou passagens repetidas para confirmar a presença do perigo [3], [2], limiarização de confiança, na qual os limiares de classificação são ajustados para equilibrar revocação e precisão com base nos requisitos operacionais [21], e fusão multimodal, na qual as deteções por acelerómetros são combinadas com GPS, câmaras ou outras modalidades sensoriais para reduzir os falsos positivos [31], [32]. Por fim, o *crowdsourcing* e a implementação em frotas através de aplicações para *smartphone* ou veículos de frota podem permitir uma monitorização rodoviária em larga escala e de baixo custo [18]. Os requisitos computacionais moderados da 1D CNN tornam-na adequada para inferência no próprio dispositivo em *smartphones* modernos.

5.5.3. Integração com Fluxos de Trabalho de Manutenção

Os perigos detetados devem ser integrados nos fluxos de trabalho de manutenção municipal para gerar valor. A priorização pode ser suportada pela severidade do perigo, inferida a partir da magnitude da aceleração, bem como pela localização através de coordenadas GPS e pela confiança da deteção, fatores que podem informar a tomada de decisão em matéria de manutenção [3], [2]. A validação também é necessária, uma vez que deteções de elevada confiança podem desencadear ordens de trabalho automatizadas, enquanto deteções de baixa confiança podem requerer validação

manual através de imagens ou inspeção no terreno [31]. Além disso, a monitorização longitudinal baseada em deteções repetidas ao longo do tempo pode acompanhar a evolução dos perigos, como o crescimento de buracos, e apoiar estratégias de manutenção preventiva.

5.6. Limitações e Ameaças à Validade

Várias limitações condicionam a interpretação e a generalização dos resultados.

5.6.1. Limitações dos Conjuntos de Dados

Devem ser consideradas várias limitações na interpretação dos resultados. Em primeiro lugar, a representação de buracos foi limitada, uma vez que o conjunto de teste continha apenas 980 janelas de buracos, correspondentes a 2,1% dos dados. Tal proporciona um poder estatístico limitado para estimar o desempenho na deteção de buracos, o que significa que os intervalos de confiança para as métricas desta classe são amplos e que o desempenho poderá variar substancialmente em conjuntos de dados de buracos maiores e mais diversificados [21]. Em segundo lugar, a qualidade das anotações poderá ser heterogénea, uma vez que os três conjuntos de dados utilizaram diferentes metodologias de anotação, incluindo rotulagem manual, correspondência de marcas temporais e deteção baseada em sensores, podendo introduzir inconsistências nas etiquetas de referência [25]. Em terceiro lugar, os conjuntos de dados não incluíam informação sobre a severidade dos perigos, fornecendo apenas etiquetas binárias de perigo sem gradações de severidade, tais como buracos superficiais versus profundos ou lombas baixas versus altas. Modelos sensíveis à severidade poderão apoiar melhor a priorização da manutenção [3], [2]. Por fim, a diversidade geográfica foi limitada, uma vez que os conjuntos de dados tiveram origem em regiões específicas e poderão não generalizar para condições rodoviárias, tipos de veículos ou comportamentos de condução de outras regiões.

5.6.2. Limitações Metodológicas

Devem também ser consideradas várias limitações metodológicas. A otimização sistemática de hiperparâmetros foi realizada apenas para o modelo *Random Forest*, enquanto os modelos de aprendizagem profunda e os restantes modelos clássicos foram avaliados utilizando hiperparâmetros predefinidos ou ajustados manualmente,

podendo tal ter subestimado o seu desempenho [16]. Além disso, as arquiteturas de aprendizagem profunda foram concebidas manualmente com base em trabalhos anteriores, o que significa que a pesquisa de arquiteturas neuronais ou uma exploração arquitetural mais extensiva poderá identificar configurações superiores [41], [38]. A avaliação baseou-se também numa única divisão treino-teste baseada em grupos, com validação cruzada realizada apenas no conjunto de treino; por conseguinte, múltiplas divisões treino-teste ou validação cruzada aninhada proporcionariam estimativas de desempenho mais robustas [25]. Por fim, a janela de 50 amostras com 50% de sobreposição foi selecionada com base em trabalhos anteriores, mas não foi otimizada sistematicamente, embora o tamanho da janela e a sobreposição possam afetar substancialmente tanto o desempenho como o custo computacional [24].

5.6.3. Limitações da Avaliação

Devem ser assinaladas várias limitações da avaliação. O conjunto de teste encontrava-se severamente desequilibrado, com 96,2% das amostras pertencentes à classe *none*, tornando a estimativa de desempenho desafiante, particularmente para as classes minoritárias com suporte limitado [21], [22]. Além disso, as métricas com média macro tratam todas as classes de forma equitativa, o que poderá não refletir prioridades operacionais nas quais certos tipos de perigo podem ser mais importantes. Por fim, os modelos foram avaliados em janelas independentes, sem considerar a coerência temporal, como deteções consecutivas do mesmo perigo. Assim, a suavização temporal ou a avaliação ao nível da sequência poderá proporcionar estimativas de desempenho mais realistas [38].

5.7. Comparação com Trabalhos Anteriores

Os resultados deste estudo estão alinhados com a investigação anterior sobre a deteção de perigos rodoviários baseada em acelerómetros, enquanto a expandem.

5.7.1. Consistência com a Literatura

Os resultados deste estudo são consistentes com várias conclusões reportadas em trabalhos anteriores. A superioridade da 1D CNN na deteção de buracos está alinhada com Varona et al. [18], que demonstraram que as CNN superam os classificadores tradicionais na distinção entre buracos, lombas e ações do condutor. De forma

semelhante, Ozoglu e Gökgöz [19] alcançaram aproximadamente 93% de exatidão de validação utilizando modelos CNN em dados de acelerómetros de smartphone. O forte desempenho dos *ensembles* baseados em árvores, particularmente o *Random Forest* e o *Gradient Boosting*, são também consistentes com Wu et al. [3], que alcançaram 88,5% de precisão e 75% de *recall* na deteção de buracos utilizando *Random Forest* com características nos domínios temporal e da frequência. Dey et al. [17] reportaram igualmente cerca de 92% de exatidão utilizando *Random Forest* com características de acelerómetros e magnetómetro para classificação da superfície rodoviária. Os padrões de importância das características observados neste estudo, especialmente a predominância das características de magnitude e de aceleração vertical apresentadas na Figura 5, são consistentes com Martínez et al. [2], que identificaram a aceleração vertical como o principal discriminador para buracos e lombas, e com Purnama e Gunawan [30], que enfatizaram características baseadas na magnitude para a classificação rodoviária baseada em vibração. Por fim, a dificuldade na deteção de classes minoritárias, particularmente no caso dos buracos, está alinhada com Rautiainen et al. [34], que verificaram que o *Random Forest* apresentava enviesamento para classes dominantes em conjuntos de dados inerciais desequilibrados, e com Ward et al. [21], que desenvolveram técnicas especializadas de balanceamento por ensemble para abordar este problema.

5.7.2. Divergência face a Trabalhos Anteriores

Apesar das convergências identificadas com a literatura, os resultados obtidos neste estudo também divergem de alguns trabalhos anteriores que reportam níveis de desempenho mais elevados na deteção de irregularidades rodoviárias [3], [17], [18], [19]. Esta divergência deve ser interpretada à luz de diferenças metodológicas relevantes, relacionadas com a natureza dos dados, a formulação do problema, a estratégia de validação e as métricas utilizadas. Esta observação encontra suporte na revisão sistemática conduzida pela autora no âmbito deste projeto, que identifica a heterogeneidade de métricas reportadas como uma das principais barreiras à comparação entre estudos, salientando que apenas uma pequena fração das publicações analisadas reporta a pontuação F1 como métrica primária de desempenho, recorrendo a maioria a exatidão, latência de alerta ou demonstrações de viabilidade [33].

Em primeiro lugar, vários estudos anteriores foram conduzidos em contextos mais controlados, com menor heterogeneidade de sensores, veículos, posições de aquisição ou condições rodoviárias [5], [6], [9], [12]. Nestes casos, os padrões de vibração associados a buracos ou lombas tendem a ser mais consistentes, facilitando a aprendizagem dos modelos. No presente estudo, pelo contrário, foram integrados múltiplos conjuntos de dados públicos, com diferentes condições de recolha, diferentes procedimentos de anotação, diferentes posições de sensores e diferentes distribuições de classes [27], [28], [29]. Esta opção aumenta a validade ecológica da avaliação, mas também torna a tarefa de classificação substancialmente mais exigente [25].

Em segundo lugar, parte da literatura reporta sobretudo métricas globais, como a exatidão, que podem produzir uma perceção otimista do desempenho em cenários de forte desequilíbrio entre classes [21], [22]. No presente estudo, a análise privilegiou métricas mais sensíveis ao desempenho das classes minoritárias, nomeadamente a pontuação F1 macro, a exatidão balanceada, a revocação por classe e a matriz de confusão. Esta opção metodológica torna a avaliação mais rigorosa, uma vez que penaliza modelos que classificam corretamente a classe maioritária, mas apresentam fraca capacidade de deteção de buracos ou lombas [21], [22].

Em terceiro lugar, alguns estudos anteriores abordam tarefas binárias, como a distinção entre estrada normal e anomalia, enquanto o presente trabalho considera uma classificação multiclasse, distinguindo entre superfície normal, buraco e lomba [3], [4], [10], [18]. Esta formulação é mais próxima de uma aplicação prática, mas também mais complexa, porque exige que o modelo não apenas detete a presença de uma irregularidade, mas também identifique corretamente o seu tipo.

A análise *bootstrap* reforça esta interpretação, ao demonstrar que as diferenças entre modelos devem ser avaliadas considerando também a incerteza associada às métricas de desempenho [39], [40]. Embora alguns modelos, como o *Random Forest* otimizado, apresentem elevada exatidão global, os intervalos de confiança das métricas por classe demonstram que o desempenho na deteção de buracos permanece muito limitado. Em contraste, a 1D CNN apresentou um desempenho mais robusto na classe *pothole*, com uma pontuação F1 de 0.3566 e intervalo de confiança a 95% entre 0.3234 e 0.3891, evidenciando maior capacidade para identificar eventos raros e subtis.

Assim, os resultados mais moderados obtidos neste estudo não devem ser interpretados apenas como menor desempenho absoluto dos modelos, mas como consequência de um protocolo experimental mais exigente, assente em dados heterogéneos, classes

desequilibradas, avaliação multiclasse e métricas adequadas à deteção de eventos raros [21], [22], [25]. Esta diferença reforça a importância de comparar estudos não apenas pelos valores finais das métricas, mas também pelas condições de recolha dos dados, pela estratégia de validação, pela distribuição das classes e pelo tipo de tarefa de classificação considerada. Neste contexto, abordagens futuras baseadas em adaptação de domínio poderão ser relevantes para melhorar a generalização entre diferentes fontes de dados sensoriais [44].

5.7.3. Contributos Originais

Este estudo expande trabalhos anteriores em diferentes vertentes. Em primeiro lugar, proporciona uma avaliação comparativa abrangente através da comparação sistemática de 9 modelos clássicos e 3 arquiteturas de aprendizagem profunda num conjunto de dados unificado, oferecendo a avaliação empírica mais abrangente, até à data, das abordagens de classificação de perigos rodoviários [18], [3], [19]. Em segundo lugar, a integração de três conjuntos de dados públicos distintos, abrangendo diferentes posições dos sensores, taxas de amostragem e tipos de perigos, proporciona uma avaliação da generalização mais realista do que estudos baseados num único conjunto de dados [25]. Em terceiro lugar, a análise detalhada das falhas, incluindo a análise da matriz de confusão apresentada na Figura 4 e as decomposições do desempenho por classe, revela modos de falha específicos, como a classificação incorreta de buracos como *none*, os quais podem orientar direções de investigação futuras [21], [22]. Por fim, a análise da importância das características do *Random Forest* apresentada na Figura 5 fornece evidência quantitativa da importância relativa de diferentes características do sinal, complementando o conhecimento qualitativo do domínio [14], [30].

5.8. Direções de Investigação Futura

Os resultados e as limitações deste estudo sugerem várias direções promissoras para investigação futura.

5.8.1. Arquiteturas Avançadas de Aprendizagem Profunda

Trabalhos futuros poderão explorar várias abordagens avançadas de modelação. Modelos baseados em atenção podem aprender a focar-se nas regiões temporais mais informativas dentro de cada janela, melhorando potencialmente a deteção de

assinaturas de perigo breves e localizadas [38]. Arquiteturas de modelação temporal multiescala, como a *DeepConvContext*, poderão também captar melhor tanto impactos instantâneos como respostas veiculares de maior duração. Além disso, arquiteturas *Transformer*, que utilizam mecanismos de auto atenção, têm demonstrado potencial na classificação de séries temporais e poderão superar CNN e LSTM na deteção de perigos rodoviários [22]. Por fim, abordagens de aprendizagem profunda por ensemble, combinando múltiplos modelos profundos, como 1D CNN, LSTM e CNN+LSTM, poderão melhorar a robustez e o desempenho global [21].

5.8.2. Engenharia de Características Aprimorada

Trabalhos futuros poderão também melhorar a representação das características através da incorporação de tipos adicionais de características. Características no domínio da frequência, tais como transformadas de Fourier, decomposições por onduletas ou características espectrais, poderão captar as características de frequência das vibrações induzidas por perigos [35], [30]. Características temporais que descrevem a dinâmica dentro de cada janela, incluindo autocorrelação, taxa de cruzamento por zero e temporização dos picos, poderão melhorar adicionalmente a discriminação [14]. Além disso, características específicas do domínio, derivadas de modelos de dinâmica veicular, tais como a resposta da suspensão e a interação pneu-estrada, poderão proporcionar representações mais interpretáveis e robustas [2].

5.8.3. Tratamento Aprimorado do Desequilíbrio entre Classes

Os projetos subsequentes terão a possibilidade de investigar táticas otimizadas para lidar com o desequilíbrio entre classes e a deteção de classes minoritárias. Métodos avançados de reamostragem, incluindo variantes do SMOTE concebidas para dados de séries temporais ou modelos generativos, como as GAN, poderiam ser utilizados para sintetizar exemplos realistas das classes minoritárias [23]. A aprendizagem sensível ao custo poderá também ser explorada de forma sistemática, através da otimização de pesos por classe ou de funções de perda, com o objetivo de equilibrar precisão e revocação de acordo com os requisitos operacionais [21], [22]. Além disso, a deteção de perigos poderia ser formulada como um problema de deteção de anomalias, no qual os modelos são treinados principalmente com dados de estrada normal e utilizados para assinalar desvios [35].

5.8.4. Fusão Multimodal de Sensores

Trabalhos futuros poderão também investigar a integração multimodal para melhorar a deteção de perigos rodoviários. A integração de câmaras, através de métodos como SSD ou YOLO, poderia ser combinada com deteções baseadas em acelerómetros para melhorar a precisão e fornecer confirmação visual [31], [32]. A integração de GPS e mapas poderá também apoiar o agrupamento espacial das deteções e permitir que as localizações dos perigos sejam acompanhadas ao longo do tempo através de coordenadas GPS e mapas da rede rodoviária [3], [2]. Além disso, a integração de outros sensores inerciais, como giroscópios e magnetómetros, poderá fornecer informação complementar sobre a orientação e o movimento do veículo [17].

5.8.5. Aprendizagem por Transferência e Adaptação de Domínio

Estudos posteriores poderão considerar estratégias de aprendizagem por transferência e adaptação de domínio com vista a melhorar a generalização. O pré-treino de modelos em grandes conjuntos de dados de reconhecimento de atividades, seguido de ajuste fino em dados de perigos rodoviários, poderá melhorar a aprendizagem de características [41]. O treino adversarial de domínio poderá também ser utilizado para desenvolver modelos que aprendam representações invariantes ao domínio, tornando-os mais robustos a variações na posição dos sensores, no tipo de veículo e nas condições da superfície rodoviária [44]. Além disso, abordagens de aprendizagem com poucos exemplos (few-shot learning) e de meta-aprendizagem poderão permitir uma adaptação rápida a novos tipos de perigos ou configurações de sensores quando apenas estão disponíveis dados etiquetados limitados.

5.8.6. Estudos de Implementação em Contexto Real

Investigações futuras poderão incluir estudos de implementação prática e validação. A validação em campo poderá envolver a implementação dos modelos em sistemas operacionais de monitorização rodoviária e a avaliação do seu desempenho em fluxos de dados reais com validação por dados de referência [2], [3]. Estudos com utilizadores poderão também avaliar a usabilidade e a aceitação de aplicações de deteção de perigos baseadas em *smartphone* entre condutores e pessoal responsável pela manutenção municipal [18]. Além disso, uma análise custo-benefício poderá quantificar o valor económico da deteção automatizada de perigos em termos de redução dos

custos de manutenção, melhoria da segurança rodoviária e prolongamento da vida útil do pavimento.

5.8.7. Exploração da Aplicação do Potencial do Edge-AI

Trabalhos futuros poderão também explorar o potencial do Edge-AI na deteção de perigos rodoviários em tempo real. Esta possibilidade está alinhada com os resultados obtidos, uma vez que a 1D CNN apresentou o melhor desempenho global e um tempo de inferência reduzido, tornando-a potencialmente adequada para execução em dispositivos próximos da fonte de dados, como *smartphones*, sistemas embebidos ou unidades de computação periférica instaladas em veículos.

A implementação em Edge-AI permitiria processar localmente as janelas de acelerómetro, reduzindo a necessidade de transmissão contínua de dados para a nuvem e diminuindo a latência entre a deteção e a comunicação do perigo. Esta abordagem seria particularmente relevante em aplicações municipais ou em frotas de veículos, nas quais os eventos detetados poderiam ser posteriormente agregados com informação GPS e enviados para uma plataforma central de monitorização.

No entanto, esta linha de trabalho exigiria validação adicional em contexto real, incluindo a avaliação do consumo energético, da latência por janela, da utilização de memória, da robustez em diferentes dispositivos e da gestão de falsos positivos. Técnicas de compressão de modelos, como quantização, poda ou destilação de conhecimento, poderão ser exploradas para reduzir o custo computacional, mantendo a capacidade de deteção das classes minoritárias.

Capítulo 6 – Conclusões

Esta dissertação apresentou um estudo comparativo abrangente entre abordagens de aprendizagem automática clássica e de aprendizagem profunda para a classificação de perigos rodoviários com recurso a dados de acelerómetro. A investigação abordou o desafio crítico da deteção automatizada de buracos e lombas a partir de dados sensoriais heterogéneos, com implicações para os sistemas de transporte inteligentes e para a manutenção rodoviária municipal.

6.1. Síntese dos Contributos

Os contributos principais deste trabalho são seis. Em primeiro lugar, proporciona uma avaliação empírica abrangente através da comparação sistemática de 9 modelos clássicos de aprendizagem automática e 3 arquiteturas de aprendizagem profunda num conjunto de dados unificado e multifonte, composto por 47 706 janelas de teste provenientes de três conjuntos de dados públicos. Em segundo lugar, demonstra a superioridade da aprendizagem profunda na deteção de buracos, uma vez que a 1D CNN alcançou uma melhoria de uma ordem de grandeza na pontuação F1 da classe *pothole* (0.3566) em comparação com o melhor valor de F1-pothole obtido entre os modelos clássicos (0.0376, Regressão Logística), representando um avanço na deteção de classes minoritárias em dados de perigos rodoviários severamente desequilibrados. Em terceiro lugar, identifica os pontos fortes dos modelos clássicos, demonstrando que os métodos de ensemble baseados em árvores, nomeadamente o *Random Forest* e o *Histogram-based Gradient Boosting*, alcançaram um desempenho competitivo na deteção de lombas, com pontuações F1 entre 0.3542 e 0.3590, exigindo simultaneamente um custo computacional substancialmente inferior, com tempos de treino de 1 a 2 segundos. Isto demonstra a sua relevância contínua para aplicações com recursos limitados. Em quarto lugar, o estudo fornece perspetivas sobre a importância das características através de uma análise quantitativa da importância das características do *Random Forest*, revelando a predominância das características de magnitude e de aceleração vertical e proporcionando validação empírica do conhecimento do domínio para futuros esforços de engenharia de características. Em quinto lugar, apresenta um enquadramento metodológico baseado num fluxo de pré-processamento robusto, que incorpora normalização de etiquetas, segmentação por janelas deslizantes com rotulagem baseada em prioridades, extração abrangente de características e divisão treino-teste baseada em grupos para prevenir a fuga de informação. Por fim, o estudo oferece uma análise detalhada dos modos de falha através da análise da matriz de confusão e da decomposição do desempenho por classe, revelando padrões específicos de falha, como a classificação incorreta de 99% dos buracos como estrada normal pelos modelos clássicos, e orientando direções futuras de investigação.

6.2. Principais Resultados

Os resultados experimentais permitiram identificar vários aspetos críticos. Em primeiro lugar, as arquiteturas convolucionais destacaram-se na deteção de eventos subtis e raros. A 1D CNN apresentou uma melhoria expressiva na deteção de buracos, alcançando uma pontuação F1 de 0.3566, uma precisão de 0.7419 e uma revocação de 0.2347. Este resultado demonstra que as características convolucionais aprendidas conseguem captar padrões de sinal específicos associados a buracos que não são adequadamente representados por características estatísticas manualmente definidas, com implicações significativas para aplicações críticas em termos de segurança, nas quais a deteção de eventos raros e subtis é essencial [18], [19].

Em segundo lugar, os *ensembles* clássicos baseados em árvores mantiveram-se competitivos na deteção de eventos estereotipados. O *Random Forest* e o *Gradient Boosting* alcançaram pontuações F1 para a classe *speed_bump* entre 0.3011 e 0.3590, comparáveis às dos modelos de aprendizagem profunda, cujas pontuações F1 variaram entre 0.2277 e 0.2947, exigindo simultaneamente tempos de treino inferiores em 1 a 2 ordens de grandeza. Isto demonstra que, para perigos com assinaturas consistentes e distintivas, os métodos clássicos oferecem um compromisso atrativo entre desempenho e eficiência [3], [11], [17].

Em terceiro lugar, as arquiteturas recorrentes não superaram as CNN na classificação baseada em janelas curtas. Os modelos LSTM e CNN+LSTM alcançaram pontuações F1 macro de 0.4245 e 0.4145, respetivamente, valores substancialmente inferiores à pontuação de 0.5108 obtida pela 1D CNN. Este resultado sugere que as dependências temporais para além da janela de 50 amostras não são críticas para a deteção de perigos ou não são aprendidas de forma eficaz pelas arquiteturas recorrentes com dados de treino limitados [36], [37].

Em quarto lugar, o acentuado desequilíbrio entre classes constituiu um desafio fundamental. Todos os modelos apresentaram dificuldades perante a distribuição extrema das classes, composta por 96,2% de amostras da classe *none*, 2,1% da classe *pothole* e 1,7% da classe *speed_bump*, sendo a deteção de buracos particularmente difícil. A diferença substancial entre o desempenho em treino e no teste, como a pontuação F1 macro do *Random Forest* de 0.9456 versus 0.4498, indica que as estratégias atuais de reequilíbrio são insuficientes e que são necessárias abordagens

mais sofisticadas, como a aprendizagem sensível ao custo e a deteção de anomalias [21], [22].

Em quinto lugar, as características de magnitude e de aceleração vertical foram as mais informativas. A análise da importância das características do *Random Forest* demonstrou que as características derivadas da magnitude, incluindo *mag_rms*, *mag_max* e *mag_energy*, juntamente com características de aceleração vertical, como *accZ_energy*, *accZ_abs_mean* e *accZ_rms*, dominaram as primeiras posições, proporcionando validação empírica do conhecimento do domínio relativo ao movimento veicular induzido por perigos rodoviários [3], [10], [30].

Em sexto lugar, a análise de incerteza por *bootstrap* confirmou a robustez dos principais resultados, demonstrando que a superioridade da 1D CNN em termos de F1 macro se manteve quando consideradas 5000 reamostragens das predições do conjunto de teste [39], [40]. Esta análise reforçou também a interpretação de que os modelos clássicos baseados em árvores, embora competitivos na deteção de lombas e eficientes do ponto de vista computacional, continuam a revelar limitações substanciais na deteção de buracos [3], [11], [17]. Assim, os resultados devem ser interpretados não apenas com base nas métricas pontuais, mas também considerando os respetivos intervalos de confiança e a incerteza associada ao desempenho por classe, especialmente em cenários de forte desequilíbrio entre classes [21], [22].

Por fim, a generalização entre fontes de dados heterogéneas permaneceu desafiante. A diferença substancial de desempenho entre os conjuntos de validação cruzada e de teste indica uma generalização limitada entre ficheiros de origem com diferentes posições de sensores, taxas de amostragem e tipos de veículos, evidenciando a necessidade de técnicas de adaptação de domínio e de conjuntos de treino maiores e mais diversificados [25], [44].

6.3. Implicações Práticas

Os resultados deste estudo apresentam implicações diretas para a conceção e implementação de sistemas de monitorização de perigos rodoviários baseados em acelerómetros. Para os responsáveis pelo desenvolvimento de sistemas, a *1D CNN* deve constituir a escolha predefinida para aplicações que exijam uma deteção robusta de buracos, uma vez que oferece desempenho superior com requisitos computacionais moderados, incluindo 443 segundos de tempo de treino e 0.09 ms por janela para

inferência. Os *ensembles* clássicos baseados em árvores permanecem viáveis para aplicações centradas na deteção de lombas ou para sistemas embebidos com recursos limitados [19], [17]. Para as autoridades municipais, os sistemas automatizados de deteção de perigos podem proporcionar uma monitorização rodoviária em larga escala e de baixo custo quando implementados através de aplicações para smartphone ou veículos de frota. Contudo, a baixa precisão nas classes minoritárias, particularmente nas lombas, com valores de precisão entre 0.1365 e 0.2612, exige agrupamento espacial e validação multimodal para reduzir falsos positivos antes de serem desencadeadas ações de manutenção [3], [2]. Para os investigadores, o acentuado desequilíbrio entre classes e os desafios de generalização identificados neste estudo evidenciam áreas críticas para trabalhos futuros, incluindo técnicas avançadas de reequilíbrio, métodos de adaptação de domínio e fusão multimodal de sensores [21], [22], [44].

6.4. Limitações

Existem várias limitações que condicionam o âmbito e a generalização deste trabalho. Em primeiro lugar, a representação de buracos foi limitada, uma vez que o conjunto de teste continha apenas 980 janelas de buracos, correspondentes a 2,1% dos dados, o que proporcionou poder estatístico limitado para estimar o desempenho na deteção de buracos e poderá ter subestimado as capacidades dos modelos em conjuntos de dados maiores e mais diversificados. Em segundo lugar, o estudo utilizou uma única divisão treino-teste baseada em grupos, com validação cruzada realizada apenas no conjunto de treino, o que significa que múltiplas divisões ou validação cruzada aninhada proporcionariam estimativas de desempenho mais robustas [25]. Em terceiro lugar, a otimização sistemática de hiperparâmetros foi realizada apenas para o modelo *Random Forest*, enquanto os modelos de aprendizagem profunda e os restantes modelos clássicos foram avaliados com hiperparâmetros predefinidos ou ajustados manualmente, podendo tal ter subestimado o seu desempenho [16]. Em quarto lugar, a janela de 50 amostras com 50% de sobreposição foi selecionada com base em trabalhos anteriores, mas não foi otimizada sistematicamente, sendo que configurações alternativas de janelas poderão produzir diferentes compromissos entre desempenho e eficiência [24]. Em quinto lugar, os conjuntos de dados forneceram etiquetas binárias de perigo sem gradações de severidade, limitando a utilidade dos modelos para a

priorização da manutenção [3], [2]. Por fim, os conjuntos de dados tiveram origem em regiões geográficas e períodos temporais específicos, pelo que a generalização para outros contextos, como diferentes climas, materiais rodoviários ou tipos de veículos, permanece incerta.

6.5. Trabalho Futuro

Os resultados e as limitações deste estudo sugerem várias direções prioritárias para investigação futura. As prioridades imediatas incluem a investigação de técnicas avançadas de reequilíbrio, tais como aprendizagem sensível ao custo, *focal loss* e métodos de balanceamento por *ensemble*, com vista a melhorar a deteção de classes minoritárias [21], [22]. Outra prioridade imediata consiste na incorporação de características no domínio da frequência, incluindo decomposições por onduletas e características espectrais, de modo a complementar as representações no domínio temporal [35], [30]. Além disso, a otimização sistemática de hiperparâmetros deverá ser realizada através de pesquisas abrangentes de hiperparâmetros para modelos de aprendizagem profunda e outros algoritmos clássicos [16]. As prioridades a médio prazo incluem a exploração de abordagens de modelação temporal multiescala, tais como mecanismos de atenção e arquiteturas multiescala como a *DeepConvContext*, de modo a captar melhor tanto impactos instantâneos como respostas de maior duração [38]. A aprendizagem por transferência deverá também ser investigada através do pré-treino em grandes conjuntos de dados de reconhecimento de atividades, seguido de ajuste fino em dados de perigos rodoviários [41]. Além disso, a fusão multimodal deverá ser explorada através da integração de deteções baseadas em acelerómetros com deteção baseada em câmara e dados GPS, de modo a melhorar a precisão e a localização espacial [31], [32]. As prioridades a longo prazo incluem estudos de implementação em contexto real, nos quais os modelos sejam implementados em sistemas operacionais de monitorização rodoviária e validados em campo através de verificação com dados de referência [2]. A adaptação de domínio deverá também ser desenvolvida para melhorar a generalização entre diferentes posições dos sensores, tipos de veículos e condições rodoviárias, utilizando abordagens como treino adversarial de domínio ou meta-aprendizagem [44]. Por fim, a estimativa da severidade deverá ser explorada através da extensão da classificação binária para regressão de severidade ou classificação ordinal, de modo a apoiar a priorização da manutenção [3].

6.6. Considerações Finais

Esta dissertação demonstrou que a aprendizagem profunda, especificamente as redes neuronais convolucionais 1D, oferece vantagens substanciais em relação à aprendizagem automática clássica na deteção de perigos rodoviários subtis e raros, como buracos, em dados de acelerómetros severamente desequilibrados. A 1D CNN alcançou uma melhoria de uma ordem de grandeza na pontuação F1 da classe pothole em comparação com o melhor valor de F1-pothole obtido entre os modelos clássicos, representando um avanço significativo na deteção automatizada de perigos rodoviários. Contudo, o estudo também revelou que os *ensembles* clássicos baseados em árvores permanecem competitivos na deteção de perigos mais estereotipados, como lombas, oferecendo compromissos atrativos entre desempenho e eficiência para aplicações com recursos limitados. O acentuado desequilíbrio entre classes e a generalização limitada entre fontes de dados heterogéneas constituem desafios fundamentais que exigem investigação contínua em técnicas avançadas de reequilíbrio, métodos de adaptação de domínio e fusão multimodal de sensores.

À medida que cidades em todo o mundo procuram modernizar a manutenção rodoviária através de sistemas de transporte inteligentes, a deteção de perigos baseada em acelerómetros oferece uma solução escalável e economicamente eficiente para a monitorização contínua das condições rodoviárias. Os resultados desta dissertação fornecem orientação empírica para a seleção de modelos, evidenciam considerações metodológicas críticas e identificam direções promissoras para investigação futura. Com os avanços contínuos nas arquiteturas de aprendizagem profunda, a disponibilização de conjuntos de dados de treino maiores e mais diversificados e a validação em contextos reais de implementação, a deteção automatizada de perigos rodoviários tem potencial para melhorar significativamente a segurança rodoviária, reduzir os custos de manutenção e aumentar a qualidade da infraestrutura de transporte. Este percurso é consistente com o roteiro de investigação proposto pela autora numa revisão sistemática complementar conduzida no âmbito deste projeto, que recomenda a normalização de protocolos de avaliação, a expansão de ensaios de campo multi-local e a adoção de arquiteturas de privacidade desde a conceção, incluindo aprendizagem federada e inferência local, como prioridades para colmatar o fosso entre o desempenho em protótipo e a operacionalidade real dos sistemas de monitorização rodoviária [33].

Referências

[1] N. Silva, V. Shah, J. Soares, e H. Rodrigues, «Road Anomalies Detection System Evaluation», *Sensors*, vol. 18, n.º 7, p. 1984, jul. 2018, doi: 10.3390/s18071984.

[2] F. Martinez, L. Carlos Gonzalez, e M. Ricardo Carlos, «Identifying Roadway Surface Disruptions Based on Accelerometer Patterns», *IEEE Lat. Am. Trans.*, vol. 12, n.º 3, pp. 455–461, mai. 2014, doi: 10.1109/TLA.2014.6827873.

[3] C. Wu *et al.*, «An Automated Machine-Learning Approach for Road Pothole Detection Using Smartphone Sensor Data», *Sensors*, vol. 20, n.º 19, p. 5564, jan. 2020, doi: 10.3390/s20195564.

[4] A. Mednis, G. Strazdins, R. Zviedris, G. Kanonirs, e L. Selavo, «Real time pothole detection using Android smartphones with accelerometers», em *2011 International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS)*, jun. 2011, pp. 1–6. doi: 10.1109/DCOSS.2011.5982206.

[5] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, e H. Balakrishnan, «The pothole patrol: using a mobile sensor network for road surface monitoring», em *Proceedings of the 6th international conference on Mobile systems, applications, and services*, em MobiSys '08. New York, NY, USA: Association for Computing Machinery, jun. 2008, pp. 29–39. doi: 10.1145/1378600.1378605.

[6] C.-W. Yi, Y.-T. Chuang, e C.-S. Nian, «Toward Crowdsourcing-Based Road Pavement Monitoring by Mobile Sensing Technologies», *IEEE Trans. Intell. Transp. Syst.*, vol. 16, n.º 4, pp. 1905–1917, ago. 2015, doi: 10.1109/TITS.2014.2378511.

[7] R. Bhoraskar, N. Vankadhara, B. Raman, e P. Kulkarni, «Wolverine: Traffic and road condition estimation using smartphone sensors», em *2012 Fourth International Conference on Communication Systems and Networks (COMSNETS 2012)*, jan. 2012, pp. 1–6. doi: 10.1109/COMSNETS.2012.6151382.

[8] P. Mohan, V. N. Padmanabhan, e R. Ramjee, «Nericell: rich monitoring of road and traffic conditions using mobile smartphones», em *Proceedings of the 6th ACM conference on Embedded network sensor systems*, em SenSys '08. New York, NY, USA: Association for Computing Machinery, nov. 2008, pp. 323–336. doi: 10.1145/1460412.1460444.

[9] A. Allouch, A. Koubâa, T. Abbes, e A. Ammar, «RoadSense: Smartphone Application to Estimate Road Conditions Using Accelerometer and Gyroscope», *IEEE Sens. J.*, vol. 17, n.º 13, pp. 4231–4238, jul. 2017, doi: 10.1109/JSEN.2017.2702739.

[10] R. Du, G. Qiu, K. Gao, L. Hu, e L. Liu, «Abnormal Road Surface Recognition Based on Smartphone Acceleration Sensor», *Sensors*, vol. 20, n.º 2, p. 451, jan. 2020, doi: 10.3390/s20020451.

[11] «Machine learning approach for predicting bumps on road». Acedido: 31 de maio de 2026. [Em linha]. Disponível em: https://ieeexplore.ieee.org/document/7456932

[12] A. Basavaraju, J. Du, F. Zhou, e J. Ji, «A Machine Learning Approach to Road Surface Anomaly Assessment Using Smartphone Sensors», *IEEE Sens. J.*, vol. 20, n.º 5, pp. 2635–2647, mar. 2020, doi: 10.1109/JSEN.2019.2952857.

[13] A. Anaissi, N. L. D. Khoa, T. Rakotoarivelo, M. M. Alamdari, e Y. Wang, «Smart pothole detection system using vehicle-mounted sensors and machine learning», *J. Civ. Struct. Health Monit.*, vol. 9, n.º 1, pp. 91–102, fev. 2019, doi: 10.1007/s13349-019-00323-0.

[14] E. Escobar-Linero, F. Luna-Perejón, L. Muñoz-Saavedra, J. L. Sevillano, e M. Domínguez-Morales, «On the feature extraction process in machine learning. An experimental study about guided versus non-guided process in falling detection

systems», *Eng. Appl. Artif. Intell.*, vol. 114, p. 105170, set. 2022, doi: 10.1016/j.engappai.2022.105170.
[15] F. Pedregosa *et al.*, «Scikit-learn: Machine Learning in Python», *J. Mach. Learn. Res.*, vol. 12, n.º 85, pp. 2825–2830, 2011.
[16] T. Chen e C. Guestrin, «XGBoost: A Scalable Tree Boosting System», em *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, em KDD '16. New York, NY, USA: Association for Computing Machinery, ago. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[17] M. R. Dey, U. Satapathy, P. Bhanse, B. Kr. Mohanta, e D. Jena, «MagTrack: Detecting Road Surface Condition using Smartphone Sensors and Machine Learning», em *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, out. 2019, pp. 2485–2489. doi: 10.1109/TENCON.2019.8929717.
[18] B. Varona, A. Monteserin, e A. Teyseyre, «A deep learning approach to automatic road surface monitoring and pothole detection», *Pers. Ubiquitous Comput.*, vol. 24, n.º 4, pp. 519–534, ago. 2020, doi: 10.1007/s00779-019-01234-z.
[19] F. Ozoglu e T. Gökgöz, «Detection of Road Potholes by Applying Convolutional Neural Network Method Based on Road Vibration Data», *Sensors*, vol. 23, n.º 22, p. 9023, jan. 2023, doi: 10.3390/s23229023.
[20] J. Wang e X. liu, «Activity Recognition using 1D convolution from Accelerometers Data», *J. Phys. Conf. Ser.*, vol. 1550, p. 032161, mai. 2020, doi: 10.1088/1742-6596/1550/3/032161.
[21] M. Ward, K. Malmsten, H. Salamy, e C.-H. Min, «Data Balanced Bagging Ensemble of Convolutional- LSTM Neural Networks for Time Series Data Classification with an Imbalanced Dataset», em *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, mai. 2021, pp. 1–5. doi: 10.1109/ISCAS51556.2021.9401389.
[22] H. Yhdego, C. Paolini, e M. Audette, «Toward Real-Time, Robust Wearable Sensor Fall Detection Using Deep Learning Methods: A Feasibility Study», *Appl. Sci.*, vol. 13, n.º 8, p. 4988, jan. 2023, doi: 10.3390/app13084988.
[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, e W. P. Kegelmeyer, «SMOTE: synthetic minority over-sampling technique», *J. Artif. Intell. Res.*, vol. 16, n.º 1, pp. 321–357, jun. 2002.
[24] M. Jaén-Vargas *et al.*, «Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models», *PeerJ Comput. Sci.*, vol. 8, p. e1052, 2022, doi: 10.7717/peerj-cs.1052.
[25] «A Comprehensive Study of Activity Recognition Using Accelerometers». Acedido: 31 de maio de 2026. [Em linha]. Disponível em: https://www.mdpi.com/2227-9709/5/2/27
[26] M. Abadi *et al.*, «TensorFlow: a system for large-scale machine learning», em *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, em OSDI'16. USA: USENIX Association, nov. 2016, pp. 265–283.
[27] R. K. Kaur e R. K. Kumar, «Dataset collected for rough and plane roads». IEEE DataPort. doi: 10.21227/EPGV-BX80.
[28] «PVS - Passive Vehicular Sensors Datasets». Acedido: 10 de maio de 2026. [Em linha]. Disponível em: https://www.kaggle.com/datasets/jefmenegazzo/pvs-passive-vehicular-sensors-datasets
[29] «Pothole, Road Quality Detection Sensor data». Acedido: 10 de maio de 2026. [Em linha]. Disponível em: https://www.kaggle.com/datasets/dextergoes/pothole-sensor-data

[30] Y. Purnama e F. Gunawan, «Vibration-Based Damaged Road Classification Using Artificial Neural Network», *TELKOMNIKA Telecommun. Comput. Electron. Control*, vol. 16, p. 2179, out. 2018, doi: 10.12928/telkomnika.v16i5.7574.
[31] S. Silvister *et al.*, «Deep Learning Approach to Detect Potholes in Real-Time using Smartphone», em *2019 IEEE Pune Section International Conference (PuneCon)*, dez. 2019, pp. 1–4. doi: 10.1109/PuneCon46936.2019.9105737.
[32] «Deep learning approach for road pothole detection using accelerometer and imagery data», Coventry University. Acedido: 1 de junho de 2026. [Em linha]. Disponível em: https://pureportal.coventry.ac.uk/en/studentTheses/deep-learning-approach-for-road-pothole-detection-using-accelerom/
[33] J. Moutinho, E. Zdravevski, F. J. Velez, I. M. Pires, e F. Madeira, «Low-Cost IoT-Based Systems for Road Hazard Monitoring: A Systematic Review», *Syst. Rev.*, 2026.
[34] H. Rautiainen, M. Alam, P. G. Blackwell, e A. Skarin, «Identification of reindeer fine-scale foraging behaviour using tri-axial accelerometer data», *Mov. Ecol.*, vol. 10, n.º 1, p. 40, set. 2022, doi: 10.1186/s40462-022-00339-0.
[35] F. Cong *et al.*, «Applying Wavelet Packet Decomposition and One-Class Support Vector Machine on Vehicle Acceleration Traces for Road Anomaly Detection», em *Advances in Neural Networks – ISNN 2013*, C. Guo, Z.-G. Hou, e Z. Zeng, Eds., Berlin, Heidelberg: Springer, 2013, pp. 291–299. doi: 10.1007/978-3-642-39065-4_36.
[36] B. U. Kempaiah, R. J. Mampilli, e K. S. Goutham, «A Deep Learning Approach for Speed Bump and Pothole Detection Using Sensor Data», em *Emerging Research in Computing, Information, Communication and Applications*, N. R. Shetty, L. M. Patnaik, H. C. Nagaraj, P. N. Hamsavath, e N. Nalini, Eds., Singapore: Springer, 2022, pp. 73–85. doi: 10.1007/978-981-16-1338-8_7.
[37] T. Patel, B. H. W. Guo, J. D. van der Walt, e Y. Zou, «Effective Motion Sensors and Deep Learning Techniques for Unmanned Ground Vehicle (UGV)-Based Automated Pavement Layer Change Detection in Road Construction», *Buildings*, vol. 13, n.º 1, p. 5, jan. 2023, doi: 10.3390/buildings13010005.
[38] M. Bock, M. Moeller, e K. Van Laerhoven, «DeepConvContext: A Multi-Scale Approach to Timeseries Classification in Human Activity Recognition», arXiv.org. Acedido: 31 de maio de 2026. [Em linha]. Disponível em: https://arxiv.org/abs/2505.20894v1
[39] B. Efron, «Bootstrap Methods: Another Look at the Jackknife», em *Breakthroughs in Statistics: Methodology and Distribution*, S. Kotz e N. L. Johnson, Eds., New York, NY: Springer, 1992, pp. 569–593. doi: 10.1007/978-1-4612-4380-9_41.
[40] B. Efron e R. J. Tibshirani, *An Introduction to the Bootstrap*. CRC Press, 1994.
[41] I. Trabelsi, J. Francoise, e Y. Bellik, «Sensor-based Activity Recognition using Deep Learning: A Comparative Study», em *Proceedings of the 8th International Conference on Movement and Computing*, em MOCO '22. New York, NY, USA: Association for Computing Machinery, jun. 2022, pp. 1–8. doi: 10.1145/3537972.3537996.
[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, e P. Dollár, «Focal Loss for Dense Object Detection», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, n.º 2, pp. 318–327, fev. 2020, doi: 10.1109/TPAMI.2018.2858826.
[43] B. Setiawan, V. Kryssanov, e U. Serdült, *Monitoring Road Surface Conditions with Cyclist's Smartphone Sensors*. 2020.
[44] Y. Ganin *et al.*, «Domain-adversarial training of neural networks», *J. Mach. Learn. Res.*, vol. 17, n.º 1, pp. 2096–2030, jan. 2016.

Anexos

Anexo A – Comparação dos Conjuntos de Dados de Origem

A Tabela 6 apresenta uma comparação dos três conjuntos de dados públicos utilizados neste estudo, antes da sua integração no fluxo experimental unificado. Os valores correspondem aos dados brutos de cada fonte, tal como disponibilizados pelos respetivos autores.

Tabela 6 - Comparação dos conjuntos de dados de origem

Dataset	**Nº de amostras**	**Taxa de Amostragem**	**Distribuição de Classes**
IEEE DataPort (Kaur & Kumar, 2022) [27]	24 296	100 Hz	Classe 0 (normal): 1 483 (6,10%) Classe 1 (buraco): 22 813 (93,90%)
Kaggle — Pothole, Road Quality Detection Sensor Data (dextergoes) [29]	9 859 (5 viagens; 96 eventos de buraco anotados)	≈ 5 Hz	96 eventos de buraco anotados sobre as 9 859 amostras de sensor (eventos discretos, não etiqueta por amostra)
PVS — Passive Vehicular Sensors (Menegazzo & von Wangenheim, 2020) [28]	6 485 430 (9 sub-datasets × 2 lados × 3 posições de sensor = 54 ficheiros)	100 Hz	none: 6 382 776 (98,42%) speed_bump: 102 654 (1,58%)

Anexo B – Resultados Completos da Otimização de Hiperparâmetros do *Random Forest*

A Tabela B.1 apresenta as 40 configurações de hiperparâmetros avaliadas através de pesquisa aleatória (RandomizedSearchCV) com validação cruzada de 5 partições, ordenadas por desempenho (pontuação F1 macro média em validação cruzada). A configuração com a 1ª classificação destacada, corresponde à configuração ótima selecionada para o modelo *Random Forest* final, alcançando uma pontuação F1 macro de 0.9456 em validação cruzada e de 0.4498 no conjunto de teste independente.

Tabela 7 - Resultados da pesquisa aleatória de hiperparâmetros do Random Forest (40 configurações)

#	n_estim.	min_split	min_leaf	max_samples	max_features	max_depth	criterion	class_weight	F1 macro (CV)	Desvio-padrão
1	500	5	1	0.8	log2	30	entropy	-	0.9456	0.0031
2	500	2	1	0.8	-	20	entropy	-	0.9439	0.0027
3	500	10	2	1	sqrt	30	entropy	-	0.9434	0.0044
4	250	2	1	0.8	sqrt	15	gini	-	0.9417	0.0020
5	800	5	2	0.6	-	-	entropy	balanced	0.9396	0.0037
6	500	2	2	1	-	15	entropy	balanced	0.9392	0.0037
7	500	5	5	0.8	-	-	entropy	balanced	0.9372	0.0030
8	500	10	1	0.6	-	30	gini	-	0.9371	0.0058
9	100	2	5	0.8	-	20	entropy	balanced	0.9358	0.0024
10	100	10	2	0.6	log2	30	entropy	-	0.9350	0.0043
11	100	5	5	1	log2	20	gini	-	0.9346	0.0031
12	250	2	10	1	-	20	entropy	-	0.9326	0.0039
13	100	10	10	1	sqrt	30	entropy	balanced	0.9312	0.0038
14	800	20	10	1	-	30	entropy	balanced	0.9301	0.0034
15	800	20	10	0.8	-	-	entropy	-	0.9293	0.0036

#	n_estim.	min_split	min_leaf	max_samples	max_features	max_depth	criterion	class_weight	F1 macro (CV)	Desvio-padrão
16	500	2	10	-	sqrt	30	gini	-	0.9278	0.0031
17	800	10	10	0.8	log2	-	entropy	-	0.9259	0.0035
18	100	20	10	0.8	log2	15	gini	balanced	0.9255	0.0034
19	250	20	10	0.8	sqrt	-	gini	-	0.9235	0.0034
20	250	20	10	0.8	-	15	gini	-	0.9234	0.0033
21	800	10	10	0.8	log2	15	gini	-	0.9228	0.0018
22	500	2	1	0.8	-	10	gini	balanced	0.9197	0.0027
23	100	5	2	-	-	10	gini	balanced	0.9190	0.0043
24	800	5	2	0.8	-	10	entropy	balanced	0.9181	0.0040
25	800	2	5	0.6	-	10	gini	balanced	0.9172	0.0037
26	100	10	1	-	log2	10	gini	balanced	0.9161	0.0021
27	250	5	2	0.8	sqrt	10	gini	-	0.9158	0.0033
28	100	5	2	-	log2	10	entropy	-	0.9096	0.0024
29	800	10	5	0.6	log2	10	gini	-	0.9087	0.0030
30	800	2	10	0.8	log2	10	entropy	balanced	0.9084	0.0027
31	100	20	10	1	sqrt	10	entropy	-	0.9045	0.0032
32	250	10	10	0.6	sqrt	10	entropy	-	0.9041	0.0033
33	800	2	1	0.8	-	5	gini	balanced	0.8571	0.0080
34	800	10	10	-	-	5	gini	balanced	0.8564	0.0085
35	250	20	1	-	log2	5	gini	balanced	0.8555	0.0046
36	250	10	10	1	sqrt	5	gini	balanced	0.8506	0.0062
37	100	20	2	0.6	sqrt	5	gini	-	0.8484	0.0033

#	n_estim.	min_split	min_leaf	max_samples	max_features	max_depth	criterion	class_weight	F1 macro (CV)	Desvio-padrão
38	250	2	1	-	sqrt	5	gini	-	0.8456	0.0019
39	800	2	10	1	log2	5	gini	-	0.8396	0.0016
40	500	20	1	-	sqrt	5	entropy	-	0.8353	0.0019