INSTITUTO TECNOLÓGICO VALE

Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração (PROFICAM)
Escola de Minas, Universidade Federal de Ouro Preto (UFOP)
Associação Instituto Tecnológico Vale (ITV)

Dissertação

UM SISTEMA DE MÚLTIPLOS CLASSIFICADORES PARA DETECÇÃO DE DEFEITOS EM DORMENTES DE AÇO

Leonardo Pessoa Freitas e Silva

Ouro Preto
Minas Gerais, Brasil
2022

Leonardo Pessoa Freitas e Silva

UM SISTEMA DE MÚLTIPLOS CLASSIFICADORES PARA DETECÇÃO DE DEFEITOS EM DORMENTES DE AÇO

Dissertação apresentada ao Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração da Universidade Federal de Ouro Preto e do Instituto Tecnológico Vale, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Controle e Automação.

Orientador: Prof. Agnaldo José da Rocha Reis, D.Sc.
Coorientador: Prof. Glauco Ferreira Gazel Yared, D.Sc.

Ouro Preto
2022

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

F866u Freitas e Silva, Leonardo Pessoa.
Um sistema de múltiplos classificadores para detecção de defeitos em dormentes de aço. [manuscrito] / Leonardo Pessoa Freitas e Silva. - 2022.
114 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Agnaldo José da Rocha Reis.
Coorientador: Prof. Dr. Glauco Ferreira Gazel Yared.
Dissertação (Mestrado Profissional). Universidade Federal de Ouro Preto. Programa de Mestrado Profissional em Instrumentação, Controle e Automação de Processos de Mineração. Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração.
Área de Concentração: Engenharia de Controle e Automação de Processos Minerais.

1. Aprendizado do computador. 2. Ferrovias - Dormentes - Dormentes de Aço. 3. Ferrovias - Trilhos - Defeitos. 4. Ferrovias - Medidas de segurança. 5. Sistemas de Múltiplos Classificadores (SMC). I. Reis, Agnaldo José da Rocha. II. Yared, Glauco Ferreira Gazel. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 681.5:622.2

Bibliotecário(a) Responsável: Maristela Sanches Lima Mesquita - CRB-1716

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
ESCOLA DE MINAS
PROGR. POS GRAD. PROF. INST. CONT. E AUT. PROCESSOS DE MIN.

FOLHA DE APROVAÇÃO

Leonardo Pessoa Freitas e Silva

Um sistema de multiplos classificadores para detecção de defeitos em dormentes de aço

Dissertação apresentada ao Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração (PROFICAM), Convênio Universidade Federal de Ouro Preto/Associação Instituto Tecnológico Vale - UFOP/ITV, como requisito parcial para obtenção do título de Mestre em Engenharia de Controle e Automação na área de concentração em Instrumentação, Controle e Automação de Processos de Mineração.

Aprovada em 06 de julho de 2022

Membros da banca

Doutor - Agnaldo José da Rocha Reis - Orientador (Universidade Federal de Ouro Preto)
Doutor - Glauco Ferreira Gazel Yared - Coorientador (Universidade Federal de Ouro Preto)
Doutor - Eduardo josé da Silva Luz -(Universidade Federal de Ouro Preto)
Doutor - Jodelson Sabino - Vale SA
Doutor - Guilherme de Alencar Barreto - Univesidade Federal do Ceará

Agnaldo José da Rocha Reis, orientador do trabalho, aprovou a versão final e autorizou seu depósito no Repositório Institucional da UFOP em 23/09/2022

Documento assinado eletronicamente por **Bruno Nazário Coelho**, **COORDENADOR(A) DE CURSO DE PÓS-GRADUAÇÃO EM INST. CONTROLE AUTOMAÇÃO DE PROCESSOS DE MINERAÇÃO**, em 29/09/2022, às 14:23, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0405507** e o código CRC **B4483232**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: (31)3552-7352 - www.ufop.br

Agradecimentos

A Deus, que iluminou o meu caminho nos momentos difíceis e deu forças para me manter firme ao longo dessa caminhada.

Aos meus pais Vicente de Paula (In Memorian) e Maria Inês, agradeço profundamente pelo apoio, amor e carinho.

Aos professores, Glauco e Agnaldo, pela orientação e os ensinamentos valiosos concedidos ao longo do desenvolvimento da pesquisa. Os aprendizados nesses dois anos foram fundamentais para a minha evolução pessoal e profissional.

Aos alunos de iniciação científica (Guilherme, Nathan, Vitor Martins, Vitor Hugo, Milene e Felipe) do Instituto de Ciências Exatas e Aplicadas (ICEA/UFOP) que, de alguma forma, deixaram as suas importantes contribuições para o projeto.

Aos membros da banca, Prof. Guilherme de Alencar Barreto, Prof. Eduardo José da Silva Luz e Jodelson Sabino, que se prontificaram a participar e opinar em busca de melhoria para o trabalho.

À Vale SA, o meu agradecimento pelo apoio financeiro e por toda a estrutura disponibilizada para o desenvolvimento deste projeto de pesquisa, que se encontra vinculado à Cátedra Under Rail.

Aos membros interno da Vale SA, Renato Lataliza Vasconcelos, Luciano Cassaro e Luciano Oliveira, também sou grato por todo suporte e conhecimentos compartilhados de todo o processo de estudo.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (CAPES), Código de Financiamento 001; do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG); e da Vale SA.

“Cada sonho que você deixa pra trás, é um pedaço do seu futuro que deixa de existir”
(Steve Jobs)

Resumo

Resumo da Dissertação apresentada ao Programa de Pós Graduação em Instrumentação, Controle e Automação de Processos de Mineração como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UM SISTEMA DE MÚLTIPLOS CLASSIFICADORES PARA DETECÇÃO DE DEFEITOS EM DORMENTES DE AÇO

Leonardo Pessoa Freitas e Silva

Julho/2022

Orientadores: Agnaldo José da Rocha Reis
Glauco Ferreira Gazel Yared

Os sistemas ferroviários são importantes para a logística do transporte de cargas e de pessoas em muitos países, contribuindo para uma melhoria nos seus indicadores econômicos. Assim, com o intuito de garantir a confiabilidade e a segurança do transporte ferroviário, torna-se cada vez mais importante o monitoramento das condições da via permanente e a realização de manutenções planejadas. No que diz respeito aos dormentes, eles devem suportar os dispositivos de fixação dos trilhos e a capacidade estrutural de transmitir as esforços dos trilhos ao lastro. Qualquer ruptura de um determinado dormente causará uma sobrecarga nos dormentes adjacentes, acelerando a fadiga da estrutura desses componentes, contribuindo para a ocorrência de novos defeitos e, finalmente, afetando a bitola da via. Especificamente com relação aos dormentes de aço, ainda não existe uma solução automática para avaliar sua condição estrutural. Neste contexto, propõe-se um novo método para detecção de defeitos em dormentes de aço a partir de sinais geométricos de via permanente, baseado em processamento de sinais e aprendizado de máquina. Cinco classificadores com diferentes características de aprendizagem foram treinados: Redes Neurais Artificiais, Modelos de Mistura Gaussianas, Modelos de Markov Ocultos, Máquina de Vetores de Suporte e AdaBoost. Além disso, um sistema de múltiplos classificadores foi implementado para melhorar a acurácia da classificação. A metodologia proposta neste trabalho demonstrou eficácia na detecção de defeitos em dormentes de aço com Taxa de Acerto acima de 80% e Taxa de Falso Positivo abaixo de 40%, na maioria dos casos.

Palavras-chave: Aprendizado de máquina, dormentes de aço, detecção de defeitos, análise de padrões, segurança ferroviária, sistema de múltiplos classificadores.

Macrotema: Logística; **Linha de Pesquisa:** Tecnologias da Informação, Comunicação e Automação Industrial; **Tema:** Inspeção Automática de Ativos; **Área Relacionada da Vale:** Ferrovias.

Abstract

Abstract of Dissertation presented to the Graduate Program on Instrumentation, Control and Automation of Mining Process as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

MULTIPLE CLASSIFIER SYSTEM FOR STEEL SLEEPERS DEFECT DETECTION

Leonardo Pessoa Freitas e Silva

July/2022

Advisors: Agnaldo José da Rocha Reis
Glauco Ferreira Gazel Yared

The railroad transport system is essential for trading activities in several countries and plays an important role to improve their economic indicators. To ensure the reliability and safety of rail transport, it is becoming increasingly important to monitor the conditions of the railway and to execute planned maintenance. With regard to sleepers, they must support the rail fastening devices and the structural capacity to transmit forces from the rails to the ballast. The occurred damages in the form of cracks in the sleepers can introduce dangerous situations depending on the daily traffic load and the crack type. Any rupture of a given sleeper will cause an overload on adjacent sleepers, accelerating the structure fatigue of such components, contributing to the occurrence of new cracks, and finally affecting that track gauge. Specifically with regard to steel sleepers, there is still no automatic solution to assess their structural condition. In this context, one proposes a new method for detecting defects in steel sleepers from the permanent way geometric signals, based on signal processing and machine learning. Five classifiers with different learning characteristics were trained: Artificial Neural Networks, Gaussian Mixture Models, Hidden Markov Models, Support Vector Machine and AdaBoost. In addition, a multiple classifier system was implemented to improve classification accuracy. The new methodology proposed in this work has demonstrated effectiveness in steel sleepers defect detection with Hit Rate above 80% and an False Positive Rate below 40%, in most cases.

Keywords: Machine learning, steel sleepers, defect detection, pattern analysis, railway safety, multiple classifier system.

Macrotheme: Logistics; **Research Line:** Information Technologies, Communication and Industrial Automation; **Theme:** Automatic Asset Inspection; **Related Area of Vale:** Railways.

Lista de Siglas e Abreviaturas

ADB	*AdaBoost* ou *Adaptive Boosting*
ANOVA	*One-way analysis of variance*
ANTT	Agência Nacional de Transportes Terrestres
CC	Carro Controle
CP	Conselheiro Pena
CRISP-DM	*Cross Industry Standard Process for Data Mining*
DFT	*Discrete Fourier Transform*
DWT	*Discrete Wavelet Transform*
EH	Entre-*Housing*
EFVM	Estrada de Ferro Vitória-Minas
EM	*Expectation Maximization*
FCA	Ferrovia Centro-Atlântica
FDR	*Fisher's Discriminant Ratio*
FFT	*Fast Fourier Transform*
FICO	Ferrovia de Integração Centro-Oeste
FIOL	Ferrovia de Integração Oeste-Leste
FLD	*Fisher Linear Discriminant*
FT	*Fourier Transform*
GMM	*Gaussian Mixture Models*
GV	Governador Valadares
HMM	*Hidden Markov Models*
MC	Matriz de Confusão
MLP	*Multilayer Perceptron*
MR	Mário Carvalho
NFN	Número de Falsos Negativos

NFP	Número de Falsos Positivos
NVN	Número de Verdadeiros Negativos
NVP	Número de Verdadeiros Positivos
PCA	*Principal Component Analyses*
PDF	*Probability Density Functions*
RBM	*Radial Basis Function*
RNA	Redes Neurais Artificiais
SI	Sistema Internacional de Unidades
SVM	*Support Vectors Machines*
TA	Taxa de Acerto
TED	Taxa de Esforço Desnecessário
WT	*Wavelet Transform*
ZCR	*Zero Crossing Rate*

Lista de Figuras

Figura 2.1 Fases do modelo de referência CRISP-DM. 22
Figura 2.2 Componentes da via permanente. 24
Figura 2.3 Ilustração do formato dos dormente de Aço. 26
Figura 2.4 Sintomas da perda da capacidade estrutural dos dormentes de aço (trinca). . 27
Figura 2.5 Sintomas da perda da capacidade estrutural dos dormentes de aço (corrosão). 27
Figura 2.6 Característica das fraturas em dormentes de aço. 28
Figura 2.7 Carro Controle Plasser EM-80H. 28
Figura 2.8 Técnica de medição da bitola. 30
Figura 2.9 Alargamento de bitola. 30
Figura 2.10 Estreitamento de bitola. 30
Figura 2.11 Sinal de bitola para dois elementos da EFVM. 31
Figura 2.12 Ilustração do empeno ou torção. 31
Figura 2.13 Sinal de empeno para dois elementos da EFVM. 32
Figura 2.14 Ilustração de uma linha desnivelada longitudinalmente. 32
Figura 2.15 Sinal de nivelamento para uma mesma curva da EFVM. 33
Figura 2.16 Sinal de nivelamento para uma mesma tangente da EFVM. 33
Figura 2.17 Ilustração da medição de uma superelevação. 34
Figura 2.18 Sinal de superelevação para duas curvas da EFVM. 34
Figura 2.19 Sinal de superelevação para uma tangente da EFVM. 34
Figura 2.20 Ilustração de desalinhamento em uma tangente. 35
Figura 2.21 Sinal de alinhamento para uma mesma curva da EFVM. 35
Figura 2.22 Sinal de alinhamento para uma mesma tangente da EFVM. 36
Figura 2.23 Diferença entre os dados da planilha de prospecção e da base rotulada. . . 37
Figura 2.24 Exemplo de parametrização espacial para uma janela de tamanho 32. . . . 44
Figura 2.25 Exemplo de configuração da RNA. 57
Figura 2.26 Espaço de características. 58
Figura 2.27 Representação de hiperplanos de separação de classes. 59
Figura 2.28 Mapeamento do espaço de entrada via função *Kernel*. 60
Figura 2.29 Bagging Ensemble. 62
Figura 2.30 Stacking Ensemble. 62
Figura 2.31 Boosting Ensemble. 63

Figura 2.32 Modelo de Markov com 6 estados. 64
Figura 2.33 Modelo da MC utilizada na pesquisa. 67

Figura 3.1 Determinação da melhor janela de dados. 73
Figura 3.2 Determinação do melhor método de extração de características. 73
Figura 3.3 Determinação do melhor método de seleção de características. 74
Figura 3.4 Frequência de ocorrência dos sinais geométricos. 77
Figura 3.5 Desempenho médio obtido para a melhor configuração (curvas). 77
Figura 3.6 Desempenho médio obtido para a melhor configuração (tangentes). 78
Figura 3.7 Matriz de Confusão para a RNA (Curvas - Conselheiro Pena). 82
Figura 3.8 Matriz de Confusão para a RNA (Curvas - Governador Valadares). 82
Figura 3.9 Matriz de Confusão para a RNA (Curvas - Mário Carvalho). 83
Figura 3.10 Matriz de Confusão para a RNA (Tangentes - Conselheiro Pena). 83
Figura 3.11 Matriz de Confusão para a RNA (Tangentes - Governador Valadares). . . . 84
Figura 3.12 Matriz de Confusão para a RNA (Tangentes - Mário Carvalho). 84

Figura 1 Opções de escolha para o usuário. 99
Figura 2 Matriz de Confusão para a SVM (Curvas - Conselheiro Pena). 101
Figura 3 Matriz de Confusão para a SVM (Curvas - Governador Valadares). 102
Figura 4 Matriz de Confusão para a SVM (Curvas - Mário Carvalho). 102
Figura 5 Matriz de Confusão para a SVM (Tangentes - Conselheiro Pena). 103
Figura 6 Matriz de Confusão para a SVM (Tangentes - Governador Valadares). . . . 103
Figura 7 Matriz de Confusão para a SVM (Tangentes - Mário Carvalho). 104
Figura 8 Matriz de Confusão para a ADB (Curvas - Conselheiro Pena). 104
Figura 9 Matriz de Confusão para a ADB (Curvas - Governador Valadares). 105
Figura 10 Matriz de Confusão para a ADB (Curvas - Mário Carvalho). 105
Figura 11 Matriz de Confusão para a ADB (Tangentes - Conselheiro Pena). 106
Figura 12 Matriz de Confusão para a ADB (Tangentes - Governador Valadares). . . . 106
Figura 13 Matriz de Confusão para a ADB (Tangentes - Mário Carvalho). 107
Figura 14 Matriz de Confusão para a GMM (Curvas - Conselheiro Pena). 107
Figura 15 Matriz de Confusão para a GMM (Curvas - Governador Valadares). 108
Figura 16 Matriz de Confusão para a GMM (Curvas - Mário Carvalho). 108
Figura 17 Matriz de Confusão para a GMM (Tangentes - Conselheiro Pena). 109
Figura 18 Matriz de Confusão para a GMM (Tangentes - Governador Valadares). . . . 109
Figura 19 Matriz de Confusão para a GMM (Tangentes - Mário Carvalho). 110
Figura 20 Matriz de Confusão para a HMM (Curvas - Conselheiro Pena). 110
Figura 21 Matriz de Confusão para a HMM (Curvas - Governador Valadares). 111
Figura 22 Matriz de Confusão para a HMM (Curvas - Mário Carvalho). 111
Figura 23 Matriz de Confusão para a HMM (Tangentes - Conselheiro Pena). 112
Figura 24 Matriz de Confusão para a HMM (Tangentes - Governador Valadares). . . . 112

Figura 25 Matriz de Confusão para a HMM (Tangentes - Mário Carvalho). 113

Lista de Tabelas

Tabela 2.1 Segmento da base rotulada para um dos elementos da EFVM. 37
Tabela 2.2 Famílias *wavelets* mais comuns. 46
Tabela 2.3 Conjunto de dados para exemplo de análise da FDR. 49
Tabela 2.4 Conjunto de dados para exemplo de análise da PCA. 52
Tabela 2.5 Algoritmos de aprendizado e sua respectiva função no *Matlab*. 56

Tabela 3.1 Configurações analisadas. 70
Tabela 3.2 Estrutura da base de dados do primeiro experimento (curvas). 71
Tabela 3.3 Estrutura da base de dados do primeiro experimento (tangentes). 71
Tabela 3.4 Melhor desempenho geral (considerando todas as técnicas utilizadas) para as curvas de cada supervisão/linha, com base nas duas métricas apresentadas. . . . 72
Tabela 3.5 Melhor desempenho para as curvas de cada supervisão/linha, fixando os melhores parâmetros (variando apenas os classificadores), com base nas duas métricas apresentadas. 74
Tabela 3.6 Melhor desempenho para as tangentes de cada supervisão/linha, fixando os melhores parâmetros (variando apenas os classificadores), com base nas duas métricas apresentadas. 75
Tabela 3.7 Combinações de sinais com os melhores resultados para as curvas. 76
Tabela 3.8 Combinações de sinais com os melhores resultados para as tangentes. . . . 76
Tabela 3.9 Segmento de um arquivo de diagnóstico gerado para inspeção em campo. . 78
Tabela 3.10 Quantidade de elementos e instâncias por configuração (curvas). 79
Tabela 3.11 Quantidade de elementos e instâncias por configuração (tangentes). 79
Tabela 3.12 Resultados para a supervisão de CP. 80
Tabela 3.13 Resultados para a supervisão de GV. 80
Tabela 3.14 Resultados para a supervisão de MR. 80
Tabela 3.15 Resultado da técnica de votação por maioria. 85
Tabela 3.16 Exemplo da técnica de grupos por votação. 86
Tabela 3.17 Resultados dos grupos de cada uma das supervisões/linha para as curvas. . 87
Tabela 3.18 Resultados dos grupos de cada uma das supervisões/linha para as tangentes. 87
Tabela 3.19 Arquivo de diagnóstico gerado para inspeção em campo (*ensemble*). . . . 89

Sumário

1 Introdução **17**

1.1 Motivação . . . 18

1.2 Trabalhos Relacionados . . . 19

1.3 Objetivos . . . 20

1.3.1 Objetivo Geral . . . 20

1.3.2 Objetivos Específicos . . . 20

1.4 Organização do Texto . . . 21

2 O Método Proposto **22**

2.1 Metodologia CRISP-DM . . . 22

2.2 Entendendo o Problema . . . 24

2.2.1 Conceitos e Características da Estrutura . . . 24

2.2.2 Dormentes de Aço e as Características de Defeito . . . 25

2.2.3 Carro Controle . . . 28

2.3 Compreensão dos Dados . . . 29

2.3.1 Sinais da Geometria Espacial da Via Permanente . . . 29

2.3.1.1 Bitola . . . 30

2.3.1.2 Empeno . . . 31

2.3.1.3 Nivelamento Longitudinal . . . 32

2.3.1.4 Nivelamento Transversal . . . 33

2.3.1.5 Alinhamento . . . 35

2.3.2 Localização dos Dormentes de Aço Danificados . . . 36

2.3.3 Construção da Base de Dados . . . 38

2.3.3.1 *Data set* 1 . . . 39

2.3.3.2 *Data set* 2 . . . 39

2.4 Preparação dos Dados . . . 40

2.4.1 Filtro de Média-Móvel Não-Causal . . . 41

2.4.2 Janelamento . . . 41

2.4.3 Extração de Características . . . 42

2.4.3.1 Atributos Espaciais . . . 43

2.4.3.2 Transformada de Fourier . . . 44
2.4.3.3 Transformada *Wavelet* . . . 45
2.4.4 Seleção de Características e Redução de Dimensionalidade . . . 47
2.4.4.1 A Maldição da Dimensionalidade . . . 47
2.4.4.2 Razão Discriminante de Fisher . . . 48
2.4.4.3 Análise de Componentes Principais . . . 50
2.5 Modelagem . . . 54
2.5.1 Redes Neurais Artificiais . . . 55
2.5.1.1 Algoritmos de Aprendizado . . . 55
2.5.1.2 Redes *Perceptron* Multicamadas . . . 55
2.5.2 Máquinas de Vetores de Suporte . . . 57
2.5.3 Modelo de Mistura de Gaussianas . . . 60
2.5.4 Métodos de *ensemble* e o *Adaboost* . . . 61
2.5.5 Modelos Ocultos de Markov . . . 64
2.6 Avaliação . . . 66
2.6.1 Matriz de Confusão . . . 66
2.6.2 Métricas Principais . . . 68

3 Análise dos Resultados **70**
3.1 Primeira Fase do Experimento . . . 70
3.1.1 Conjunto dos Sinais . . . 75
3.1.2 Efeito da Máxima Taxa de Esforço Desnecessário Aceitável . . . 77
3.1.3 Geração do Diagnóstico . . . 78
3.2 Segunda Fase do Experimento . . . 79
3.2.1 Resultados Individuais . . . 79
3.2.2 *Ensembles* . . . 84
3.2.2.1 Votação por Maioria . . . 85
3.2.2.2 Grupos por Votação . . . 86
3.2.3 Geração do Diagnóstico (*ensemble*) . . . 88
3.3 Discussão dos Resultados . . . 89

4 Conclusão **91**
4.1 Contribuições . . . 92

5 Recomendações para Trabalhos Futuros **93**

Referências Bibliográficas **94**

Apêndices **98**

1. Introdução

O sistema ferroviário é essencial para as atividades comerciais e transporte de pessoas em vários países, desempenhando um papel importante na melhoria dos seus indicadores econômicos. Essas atividades estão associadas a custos operacionais mais baixos em comparação com atividades rodoviárias e aéreas, além de ser mais eficaz em termos de utilização de energia e emissão de carbono (KHAN *et al.*, 2018). As ferrovias operadas pela Vale SA estão entre os grandes diferenciais competitivos da empresa, sendo o Brasil, país onde estão localizados os seus maiores sistemas de mineração. A principal dessas ferrovias, a Estrada de Ferro Vitória-Minas (EFVM), foco do presente trabalho, tem origem na região metropolitana de Belo Horizonte, onde faz conexão com a Ferrovia Centro-Atlântica (FCA) e alcança o porto de Tubarão, em Vitória-ES, após um percurso de 905 km. Embora seja uma ferrovia em bitola métrica (bitola igual a 1.000 mm), possui grande capacidade e eficiência no transporte de cargas, sendo que o principal produto transportado é o minério de ferro proveniente de Minas Gerais e destinado à exportação. Além disso, operam-se trens de passageiros de longa distância entre dois trechos importantes do país (Vitória-Minas).

Tendo em vista a importância do sistema ferroviário tanto para questões comerciais quanto para o transporte de passageiros, o Governo Federal do Brasil planeja realizar investimentos para ampliação da malha ferroviária nos próximos anos. O objetivo é fazer com que as ferrovias se tornem uma opção logística para escoamento de uma maior quantidade de produtos brasileiros, reduzindo os custos de transportes e, consequentemente, deixando os preços dos produtos brasileiros mais competitivos, tanto no mercado interno como no externo. O ministério da infraestrutura estima um investimento nos próximos anos de R$14 bilhões para a construção da Ferrovia de Integração Oeste-Leste (FIOL), Ferrovia de Integração Centro-Oeste (FICO) e da Ferrogrão, que deverá ligar os Estados do Mato Grosso e Pará (BRASIL, 2020).

A via permanente é responsável por absorver e dissipar o impacto da carga do material rodante. Especificamente no que diz respeito aos dormentes, eles devem suportar os dispositivos de fixação dos trilhos e a capacidade estrutural de transmitir as forças dos trilhos ao lastro. Os danos causados à sua estrutura podem implicar em situações perigosas dependendo da carga diária de tráfego e da criticidade do defeito. Alguns dos sintomas críticos causados pelo aparecimento desses defeitos é a abertura da bitola da via permanente e linha desnivelada.

Convencionalmente, os dormentes das ferrovias são inspecionados manualmente, realizados por inspetores que percorrem a via permanente observando aspectos visuais da ferrovia, em busca de possíveis problemas. No entanto, essas atividades são uma tarefa de manutenção extremamente demorada e trabalhosa e expõem a saúde e segurança dos inspetores a riscos. Alguns exemplos dessas situações são: caminhar longas distâncias na chuva ou sob temperaturas extremas; o risco de cair ao caminhar em superfícies irregulares; inspeção em áreas de risco ou de difícil acesso. Além desses problemas, a avaliação visual é subjetiva, dependendo da percepção de cada inspetor. Além disso, em algumas áreas, os dormentes podem estar em

situações não visíveis, cobertos por lastro, dificultando a inspeção. No caso da Vale SA, além deste monitoramento visual, um dos principais meios para a coleta de dados e informações é a partir dos sensores do Carro Controle (CC), que realiza inspeções bimestrais ao longo da ferrovia. O CC é um veículo ferroviário autopropulsado, capaz de realizar diversas medições, sendo parte delas tipicamente relacionadas com as características geométricas da via permanente, a partir das quais pode-se realizar inferências sobre determinadas condições estruturais da ferrovia. Além disso, o CC realiza a filmagem da rodovia para a verificação de possíveis materiais no trecho, condição da sinalização da via e da vegetação.

1.1. Motivação

As tecnologias para análise das condições estruturais de componentes ferroviários têm atraído muita atenção da academia nos últimos anos, proporcionando diversos estudos nesse seguimento. Krummenacher *et al.* (2017) introduziram dois métodos de aprendizado de máquina para detecção de defeitos em rodas de trens ferroviários. Ng *et al.* (2019) apresentaram um sistema para verificar as relações entre os defeitos da superfície do trilho e seus correspondentes sinais de aceleração da caixa de eixo. Ng *et al.* (2018) investigaram a influência da distância entre os dormentes para o crescimento da corrugação dos trilhos.

A EFVM possui dormentes de aço na composição de sua estrutura ao longo de quase todo o seu percurso. Um fator relevante, que vale tanto para os dormentes quanto para os outros componentes da superestrutura é que alguns parâmetros de desempenho da ferrovia precisam ser verificados e garantidos periodicamente de acordo com a Agência Nacional de Transportes Terrestres (ANTT) (ANTT, 2018). Especificamente com relação aos dormentes, eles devem ser mantidos de forma a garantirem a bitola, suportando os dispositivos de fixação dos trilhos e a capacidade estrutural para transmitir esforços dos trilhos para o lastro. As condições estruturais dos dormentes de ferrovia devem ser monitoradas, dando uma atenção especial a trincas e fraturas. Trincados não visíveis podem piorar e podem se tornar uma fratura completa. Qualquer ruptura de um determinado dormente pode causar uma sobrecarga nos dormentes adjacentes, acelerando a fadiga da estrutura de tais componentes e contribuindo para a ocorrência de novas trincas e, finalmente, comprometendo as condições da via permanente.

Atualmente, a Vale SA possui um sistema que apresenta de forma gráfica os dados (sinal de bitola, alinhamento, nivelamento, superelevação etc.) provenientes do CC relativos aos sinais da geometria espacial da ferrovia. Estes dados, por sua vez, são analisados considerando alguns limiares que definem as condições de normalidade dos elementos da via, de acordo com as normas de segurança. Nesse sentido, não é uma tarefa simples correlacionar as análises de forma isolada dos sinais a um problema em um elemento específico. Além disso, as análises que são realizadas pelos dados do CC, normalmente apontam para um problema mais crítico da via permanente, em que os elementos podem estar em um nível avançado do defeito. Dessa forma, considerando que essa verificação ainda é feita de forma pontual pela equipe responsável, busca-

se desenvolver um novo processo para a análise dos defeitos, que seja sistematizado e preciso. Portanto, a pesquisa se concentrou na implementação de um novo método que, a partir dos sinais geométricos coletados pelos sensores do CC já existente, seja capaz de diagnosticar trechos da EFVM contendo dormentes defeituosos tanto em seu estágio inicial quanto mais avançado.

1.2. Trabalhos Relacionados

Os dormentes de aço são empregados em vários continentes, mas aparecem de forma mais expressiva na Oceania, Europa e América do Sul. No entanto, a grande maioria dos dormentes utilizados no mundo são feitos de concreto ou madeira (FERDOUS e MANALO, 2014). Provavelmente, por esse motivo, até agora não exista uma solução industrial robusta para avaliar a condição estrutural dos dormentes de aço.

Diversas abordagens foram propostas na literatura sobre o problema de detecção de dormentes defeituosos, com diferentes casos de uso. Os métodos utilizados nos trabalhos apresentados geralmente contêm técnicas automatizadas baseadas em processamento de imagens, reconhecimento de padrões ou híbridos. Especificamente no que diz respeito aos dormentes de madeira, Yella *et al.* (2009) avaliaram o monitoramento da condição dos dormentes de uma ferrovia, extraindo características de imagem e usando uma estratégia de mesclar diferentes classificadores supervisionados (*Perceptron* multicamadas, rede de função de base radial, máquina de vetores de suporte, modelos de mistura de gaussiana e quantização de vetores de aprendizagem). Os resultados obtidos pela fusão dos classificadores demonstram uma taxa de acerto de 92% no caso em estudo.

Na proposta desenvolvida por Franca e Vassallo (2020), foi possível demonstrar a capacidade de um sistema em trabalhar com imagens reais de ferrovias, classificando os dormentes em dois tipos (madeira ou aço), além de detectar defeitos nos dormentes de madeira. A transformada de Haar e "imagem integral" (do inglês *Integral Image*) foram utilizadas, assim como outras técnicas de processamento de imagens, como detecção de bordas e cálculo de entropia, juntamente com aspectos da topologia ferroviária.

Para os dormentes de concreto, Delforouzi *et al.* (2017) introduziram um método baseado em visão computacional para detecção de trincas usando um sistema integrado incluindo solução de *hardware* e *software*. *Template Matching* e métodos para encontrar deslocamentos de imagem são as principais abordagens usadas para detecção de dormentes, além de algumas técnicas de limiarização binária. Em outro trabalho relacionado, Clark *et al.* (2017) defendem o uso de sensores de emissão acústica para detectar problemas estruturais em dormentes de concreto. Os autores apresentaram investigações experimentais para detectar trincas localizadas no centro. Os testes (flexão de três pontos) foram realizados em laboratório com quatro dormentes de concreto. Os resultados mostram que a tecnologia de sensoriamento de emissão acústica é eficaz na detecção de eventos iniciais da trinca. A pesquisa evidenciou que o salto de energia induzido por essas trincas correlaciona-se bem com outros parâmetros variáveis.

Estudos na Universidade de Wollongong, na Austrália, envolveram respostas de vibração em uma tentativa de desenvolver uma ferramenta não destrutiva para monitoramento da saúde de dormentes de concreto protendido em trilhos ferroviários (KAEWUNRUEN e REMENNIKOV, 2008). Foi investigado o efeito dinâmico das fissuras nas assinaturas de vibração dos dormentes de concreto protendido da ferrovia. A análise modal foi utilizada para avaliar as mudanças modais nas características de vibração das travessas de concreto protendido na faixa de frequência entre 0 e 1.600 Hz. A partir dos experimentos realizados em laboratório, ficou evidente que as fissuras desenvolvem atritos internos entre cimento, concreto e agregado, de modo que os coeficientes de amortecimento tendem a aumentar com maior ocorrência de fissuras.

Com relação aos dormentes de aço, uma pesquisa desenvolvida na Universidade Federal de Ouro Preto (YARED *et al.*, 2019) apresentou um método para detecção de dormentes de aço trincados na Estrada de Ferro Vitória-Minas, baseado na medição de vibração. Os sinais de vibração dos dormentes de aço foram adquiridos após a aplicação de um impacto impulsivo com uma marreta. Em seguida, foi utilizado um algoritmo de pré-processamento e técnicas de reconhecimento de padrões para diagnosticar o estado de saúde dos dormentes de aço. Os resultados apontaram um desempenho (acerto na detecção dos dormentes trincados) de 85%.

1.3. Objetivos

1.3.1. Objetivo Geral

Objetiva-se com este trabalho o desenvolvimento de um novo método para detecção de defeitos em dormentes de aço a partir de sinais geométricos de via permanente, baseado em processamento de sinais e aprendizado de máquina.

1.3.2. Objetivos Específicos

- Encontrar a melhor estratégia para modelar os dados disponíveis para diferentes tipos de elementos (curvas e tangentes), supervisões e linhas;
- Determinar o tamanho da janela espacial de dados para efeito de processamento;
- Definir a técnica de extração de características mais indicada;
- Avaliar a melhor abordagem para contornar a dificuldade associada à elevada dimensionalidade do problema em questão;
- Determinar os conjuntos dos sinais de geometria da via permanente mais relevantes para o problema;
- Avaliar o desempenho de classificadores de padrões em termos da taxa de acerto na identificação dos dormentes danificados e da taxa de esforço desnecessário;

- Realizar uma análise comparativa entre os desempenhos apresentados pelas técnicas de múltiplos classificadores implementadas;
- Construir um sistema de diagnóstico que produza um relatório contendo os intervalos de defeito previstos.

1.4. Organização do Texto

Este trabalho está dividido em mais 4 capítulos, além desta introdução. No capítulo 2 é apresentada a fundamentação teórica sobre a metodologia utilizada na pesquisa e sobre as ferramentas utilizadas em cada uma de suas etapas (pré-processamento, extração e seleção de características e classificação). Este capítulo fornece os fundamentos básicos para o entendimento e acompanhamento do trabalho, assim como as principais informações com relação às técnicas implementadas durante o desenvolvimento. Os resultados finais são assunto do capítulo 4, em que são apresentadas análises quantitativas e qualitativas do desempenho obtido para os sistemas desenvolvidos. Por fim, apresenta-se no capítulo 5 as conclusões e no capítulo 6 as recomendações para trabalhos futuros.

2. O Método Proposto

Este capítulo aborda as principais características e apresenta as etapas da metodologia utilizada nesta pesquisa, começando pela compreensão do negócio e seguindo por todas as outras fases, com exceção da última (implementação). Serão apresentados os principais conceitos das técnicas e ferramentas escolhidas para cada uma de suas etapas. São elas: técnicas de pré-processamento, ferramentas para seleção e extração de características, além de diversos classificadores. Também será exposto a hipótese sobre o problema do aumento da dimensionalidade dos dados, que está diretamente relacionado com as abordagens dessa pesquisa.

2.1. Metodologia CRISP-DM

Atendendo aos objetivos e ao problema proposto, a metodologia que se revela mais adequada para esta pesquisa é a *Cross Industry Standard Process for Data Mining* (CRISP-DM), usualmente utilizada em problemas que envolvem mineração de dados (do inglês, *data mining*). Basicamente, ela é um modelo de processo que fornece uma estrutura para a realização de projetos de mineração de dados, independente do setor da indústria e da tecnologia utilizada. A metodologia consiste em um conjunto de 6 fases e processos que são flexíveis.

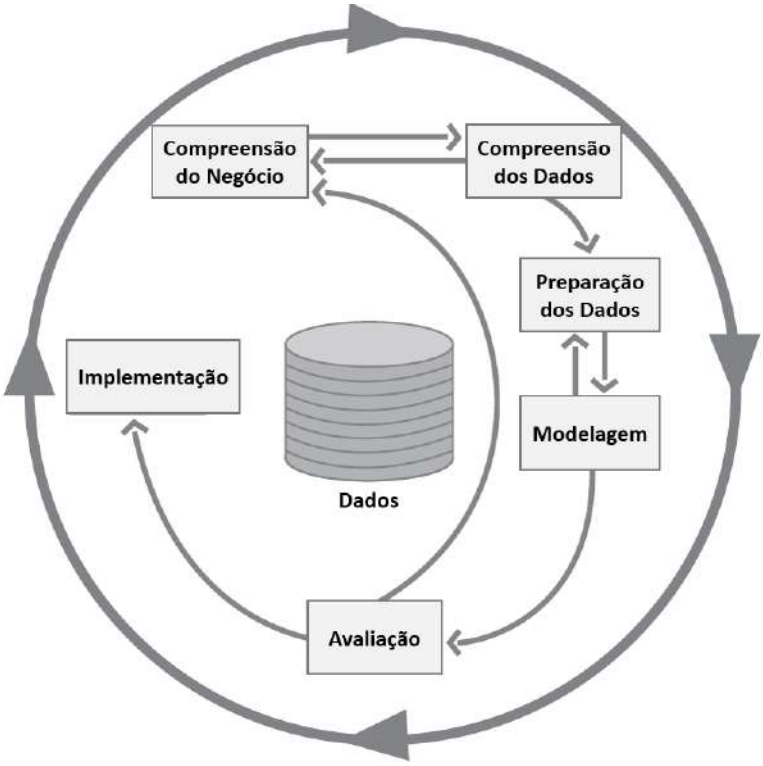


Figura 2.1: Fases do modelo de referência CRISP-DM.
Fonte: Adaptado de Chapman *et al.* (2000).

O modelo de referência para a mineração de dados fornece uma visão geral do ciclo de vida de um projeto. Ele contém as fases, suas respectivas tarefas e os relacionamentos entre elas. Apresenta-se na Figura 2.1, a relação entre as fases do modelo de referência CRISP-DM. As setas indicam as dependências mais importantes e frequentes entre as fases. Normalmente, mover-se para frente e para trás entre as diferentes fases é necessário. Nesse sentido, a partir do resultado de cada fase, determina-se qual fase, ou tarefa específica de uma fase, deve ser executada em seguida (CHAPMAN *et al.*, 2000). O círculo externo simboliza a natureza cíclica da mineração de dados em si. As lições aprendidas durante o processo e da solução implantada podem desencadear novas questões de negócios, muitas vezes mais focadas. A seguir, descreve-se brevemente o objetivo de cada fase:

- **Compreensão do Negócio:** Conhecer e compreender o problema a ser resolvido é de suma importância neste processo. Esta fase inicial se concentra na compreensão dos objetivos e requisitos do projeto a partir de uma perspectiva de negócios, buscando-se converter esse conhecimento em uma definição de problema de mineração de dados e um plano inicial projetado para atingir os objetivos.

- **Compreensão dos Dados:** Se a solução do problema de negócios é o objetivo, os dados compreendem a matéria-prima disponível a partir da qual a solução será construída. Essa fase começa com a coleta inicial dos dados e prossegue com as atividades que permitem a familiarização com os dados, identificação dos problemas de qualidade dos dados, verificação do volume dos dados, entre outras análises.

- **Preparação dos Dados:** Essa fase abrange todas as atividades necessárias para construir o conjunto de dados final (que serão inseridos na(s) ferramenta(s) de modelagem) a partir dos dados brutos iniciais. As tarefas incluem técnicas de pré-processamento, extração e seleção de características ou quaisquer outras técnicas necessárias de preparo dos dados para a fase seguinte.

- **Modelagem:** Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são ajustados para valores ótimos. Neste ponto, é possível que seja necessário o retorno à atividade de preparação dos dados, visto que existem várias técnicas para o mesmo tipo de problema e algumas delas têm requisitos específicos sobre a forma dos dados.

- **Avaliação:** A saída da fase anterior constrói o que é a base para esta fase, os modelos. Antes de prosseguir para a implantação final do modelo, é importante avaliá-lo e revisar as etapas executadas para criá-lo, garantindo que o modelo atinja adequadamente os objetivos de negócios. A avaliação vai checar se o modelo elaborado condiz com as expectativas da organização e do que foi definido anteriormente na fase inicial do processo.

- **Implementação:** A criação do modelo geralmente não é o fim do projeto. Essa fase consiste em um conjunto de ações que conduzam à utilização dos resultados oriundos das técnicas/soluções aplicadas no negócio. Mesmo que o objetivo do modelo seja aumentar o conhecimento dos dados, o conhecimento adquirido precisará ser organizado e apresentado de forma que o cliente possa usá-lo. É importante que o cliente entenda quais ações precisam ser realizadas para realmente fazer uso dos modelos criados.

2.2. Entendendo o Problema

A EFVM está entre as principais ferrovias do mundo, com alguns dos melhores índices de produtividade. São mais de 135 milhões de toneladas de carga transportada, cerca de 40% da carga ferroviária brasileira, com um tráfego diário equivalente a aproximadamente 70 navios cargueiros. Transporta mais de 60 diferentes produtos, como aço, carvão, calcário, granito, contêineres, ferro-gusa, produtos agrícolas, madeira e celulose. Com 905 quilômetros de extensão, interligada à FCA, a EFVM tem como principal destino de suas cargas o Porto de Tubarão, em Vitória-ES. Além disso, ela realiza ainda o transporte de passageiros, sendo a única ferrovia de cargas a operar trens diários para esse serviço nos dois sentidos (VLI, 2017).

2.2.1. Conceitos e Características da Estrutura

A via permanente é o ponto chave do transporte ferroviário, dado que é responsável por orientar a passagem dos trens de maneira estável e segura. De forma geral, todo o conjunto que constitui o sistema (superestrutura e infraestrutura) pode ser dividido em três principais subsistemas: o veículo, a via permanente e o solo (KOUROUSSIS *et al.*, 2015). Além disso, os subsistemas ainda podem ser divididos em vários elementos, como ilustrado na Figura 2.2.

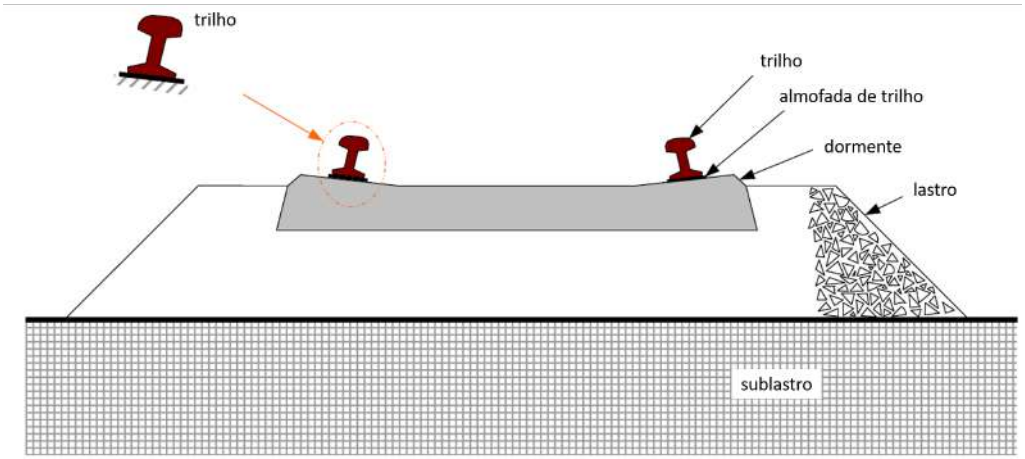


Figura 2.2: Componentes da via permanente.
Fonte: Adaptado de Kaewunruen e Remennikov (2008).

Cada um dos conjuntos ou elementos desempenha um papel característico na geração da dinâmica da ferrovia, sendo descritos a seguir.

1. Com relação ao veículo, as massas das suspensões (primária e secundária), da carroceria, do truque e do conjunto de rodas, desempenham um papel importante nos modos de vibração do veículo.

2. Relacionados à via permanente tem-se três elementos principais: o lastro, os dormentes e os trilhos (BRINA, 1988). O lastro é usado para facilitar a drenagem da água e distribuir a carga dos tirantes/dormentes da ferrovia, sem distorção por assentamento. O dormente é outro elemento constitutivo importante da pista possuindo dois papéis principais: manter a bitola do trilho e transferir as cargas dos trilhos para o lastro. Para os trilhos, vários perfis e tipos estão disponíveis no transporte ferroviário, de acordo com a forma, o peso e a natureza da via. Além disso, a geometria varia de acordo com a aplicação e o país. O papel dos trilhos é absorver uma parte das vibrações e permitir que a roda o atravesse, sem danificar o dormente.

3. Em relação ao solo, a resposta dinâmica das fundações sujeitas a cargas depende de alguns parâmetros-chave (perfis do solo, forma e geometria da fundação, interação entre fundações adjacentes, etc.) com amplificações e/ou atenuações em função da frequência de excitação.

De forma geral, a superestrutura é responsável por absorver os esforços dos vagões e aqueles provocados pela própria estrutura, e transferir para a infraestrutura, garantindo, assim, a passagem estável e segura do trem. Os principais esforços experimentados pela via permanente são: longitudinais, transversais, verticais e os impactos devido a velocidade e defeito do carro. Além disso, os defeitos apresentados na superestrutura podem ser classificados em duas categorias: geométrica e estrutural. Por ser o foco da presente pesquisa, os defeitos geométricos serão abordados com um maior aprofundamento.

O defeito geométrico é caracterizado pela diferença entre o parâmetro real (medido) e o definido em projeto (medida absoluta) ou mesmo a partir de uma base predefinida sob a própria via (medida relativa). Nesse sentido, existirá um desvio geométrico caso os parâmetros estiverem fora dos limites estabelecidos, ou seja, ultrapassando a tolerância predeterminada para cada ferrovia.

2.2.2. Dormentes de Aço e as Características de Defeito

O dormente é um dos elementos mais importantes do sistema ferroviário. Sua função é transferir e distribuir as cargas transportadas dos trilhos para o lastro, fixar transversalmente os trilhos para manter a largura da bitola correta e resistir às ações de corte e abrasão das placas de rolamento e do material do lastro. Os dormentes também resistem ao movimento lateral e longitudinal do sistema ferroviário (ZHAO *et al.*, 2007).

Quanto ao material, os dormentes utilizados atualmente podem ser feitos de madeira, concreto armado ou protendido, de aço, entre outros. Os dormentes de madeira, por um lado apresentam menor custo do que outros tipos de dormentes, são de fácil manuseio, se adequam ao lastro e podem ser usados em vias de qualquer bitola e que não apresentem manutenções rigorosas. Por outro lado, apresenta menor vida útil, menor estabilidade lateral e longitudinal e, além disso, a escassez das madeiras de boa qualidade e o reflorestamento deficiente, acarreta em seu crescente encarecimento. Os dormentes de concreto dão uma maior estabilidade, contribuindo com a economia de lastro e, devido ao seu peso, aumentam a resistência transversal da via. Além disso, possuem pouca sensibilidade aos agentes atmosféricos e uma maior durabilidade. Entretanto, os dormentes de concreto apresentam um alto custo inicial e não podem ser utilizados fora da bitola projetada. Outra desvantagem desse sistema é a dificuldade de manuseio causadas pelo fator peso.

Os dormentes de aço, foco do presente trabalho, são empregados em muitos continentes, como Oceania, Europa e América do Sul. No Brasil, a utilização dos dormentes de aço remonta à década de 80, com testes realizados na EFVM que, desde então busca-se implantar os dormentes de aço em praticamente toda a via permanente. Em sua maioria, os dormentes de aço são perfis U laminados, dotados de extremidades curvadas, com o intuito de formar garras que afundam no lastro a fim de se opor ao movimento transversal da via (Figura 2.3). Entretanto, novos modelos têm sido desenvolvidos na busca por melhorar ainda mais as características dos dormentes de aço. Dentre eles, os dormentes em Y, quando comparados ao modelo convencional, possuem uma maior redução de peso e ganho de resistência contra movimentos transversais devido à quantidade de lastro acumulado em sua parte central como consequência de seu desenho semelhante à letra Y (KAEWUNRUEN *et al.*, 2017).

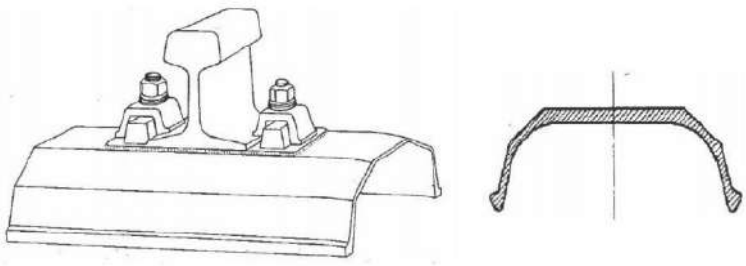


Figura 2.3: Ilustração do formato dos dormente de Aço.
Fonte: Retirado de Brina (1988).

Sobre suas vantagens, os dormentes de aço garantem uma distribuição de carga eficiente para que os danos ao lastro sejam atenuados. Um dormente de aço pesa menos que o de madeira, o que o torna fácil de manusear, além de ter uma expectativa de vida conhecida por ser superior a 50 anos. No entanto, os dormentes de aço estão sendo usados apenas em vias com menos tráfego e são considerados adequados apenas onde as velocidades são de 160 km/h ou menos. Como desvantagem tem-se o elevado custo inicial, tendências a fissuras e corrosões e a dificuldade da socaria em virtude do seu formato. Além disso, outra desvantagem desse tipo de dormente é a poluição sonara causada pela passagem dos veículos.

Ainda sobre a viabilidade da utilização dos dormentes de aço, estudos apontaram que o projeto de substituição dos dormentes de madeira da EFVM pelos de aço, na década passada, evitou que mais de 100 mil árvores/ano fossem derrubadas (VALE, 2013). Economicamente falando, quando comparado aos de concreto, os dormentes de aço permitem uma substituição ágil. Enquanto os dormentes de concreto exigem que se paralise a via por um dia, os de aço são instalados em questão de horas.

Os diferentes tipos de defeitos associados aos dormentes de aço podem estar relacionados com vários sintomas, comprometendo o próprio dormente ou mesmo as condições da via permanente. Dentre esses defeitos mencionados, tem-se a perda de capacidade estrutural do dormente, caracterizados com os sintomas de corrosão e trinca. Estes sintomas, muitas vezes, não são facilmente percebidos devido ao local onde frequentemente aparecem. Apresentam-se nas figuras 2.5 e 2.4 os dormentes com perda de capacidade estrutural (corrosão e trinca). Outro defeito, agora em um nível mais crítico, é a fratura, cujos sintomas podem ser: abertura da bitola e linha desnivelada. Considera-se ruptura quando existe uma fratura completa no dormente. Visualmente, a fratura e a ruptura são facilmente percebidas pois a aba do dormente fica elevada, assim como pode ser visto nas figuras 2.6a e 2.6a.

Figura 2.4: Sintomas da perda da capacidade estrutural dos dormentes de aço (trinca).
Fonte: O autor.

Figura 2.5: Sintomas da perda da capacidade estrutural dos dormentes de aço (corrosão).
Fonte: O autor.

(a) Primeiro exemplo. (b) Segundo exemplo.

Figura 2.6: Característica das fraturas em dormentes de aço.
Fonte: O autor.

2.2.3. Carro Controle

A queda do nível de desempenho da via deve ocorrer progressivamente. Assim, com o intuito de preservar a infra e a superestrutura, é necessário que ações de manutenção sejam executadas no sentido de conter a queda de desempenho provocada principalmente pela elevada quantidade de toneladas por eixo transportada e pelas condições climáticas da via permanente.

A Vale SA possui diversos mecanismos para o levantamento de parâmetros associados ao desgaste e desempenho dos elementos que constituem a super e infraestrutura. Dentre estes, o Carro Controle tem sido utilizado para a leitura de características geométricas da via permanente, juntamente com filmagens da superestrutura, além de outros aspectos relevantes. Tais informações são importantes para a avaliação da saúde da linha ferroviária, a qual pode ser executada por meio da análise dos dados coletados de tal modo a se estabelecer uma correspondência entre as observações realizadas e as condições dos componentes.

Figura 2.7: Carro Controle Plasser EM-80H.
Fonte: Retirado de Plasser (2022).

O modelo do CC utilizado pela Vale SA é o EM-80 (similar ao da Figura 2.7), o qual é capaz de realizar a leitura de 99 tipos de parâmetros diferentes a cada 25 cm percorridos. Dessa forma, considerando a velocidade padrão de 80 km/h, e que o carro percorra 1500 km a cada

inspeção, o veículo tem uma capacidade plena de leitura de 594 milhões de dados a 80 km/h. Além disso, o CC realiza a filmagem da rodovia para a verificação de possíveis materiais no trecho, condição da sinalização da via e da vegetação, além das medições de gabarito da via permanente. A Vale SA mantém um padrão de medições bimestrais ao longo do ano.

Vários instrumentos de medição estão presentes no CC para a coleta das diferentes informações mencionadas: Leitor de gabarito da via (*TunelLaser*), medição de aceleração vertical e horizontal, leitor de perfil do trilho e bitola (KLD), leitor de bitola em dois pontos (OGMS), imagens dos trilhos, dormentes e fixações (*RailCheck*), imagens panorâmicas da via (RailScan) e a unidade de medição inercial (IMU). A partir das informações provenientes destes dispositivos é possível encontrar os sinais utilizados nessa pesquisa (alinhamento, nivelamento, bitola, empeno e superelevação).

2.3. Compreensão dos Dados

Os dados coletados durante as medições são separados de acordo com alguns critérios, tais como o tipo do elemento, a supervisão, a Entre-*Housing* (EH) nas quais o elemento está inserido, e a linha férrea em que se localiza o elemento (a EFVM possui duas linhas férreas nos trechos investigados neste trabalho). As supervisões estão divididas por regiões e recebem o nome de algumas cidades que fazem parte do percurso da EFVM, como por exemplo, Nova Era, Mário Carvalho, Governador Valadares e etc. As Entre-*Housing*s são trechos entre dois pontos dentro de uma supervisão, por exemplo, 13/14, 31/32, 56/57 e etc. Por fim, os elementos são caracterizados pelo formato dos trechos (curvas ou tangentes). Em resumo, a EFVM é dividida por supervisões; cada supervisão possui várias EHs dentro da sua delimitação; entre duas EHs existem elementos curvas e tangentes que são enumerados de 1 até o último valor de contagem dentro desse trecho; entre duas outras EHs a numeração dos elementos se inicia novamente. Dessa forma, um arquivo com os dados coletados poderiam conter, por exemplo, informações da curva 3, linha 1, EH 23/24 de Governador Valadares, ou da tangente 5, linha 2, EH 13/14 de Conselheiro Pena.

2.3.1. Sinais da Geometria Espacial da Via Permanente

A compreensão das informações extraídas a partir dos dados coletados pelo CC depende de alguns conceitos fundamentais. Primeiramente, deve-se destacar que a leitura dos parâmetros é realizada segundo uma base de medição, onde o trilho pode ser visto ao longo de três planos: longitudinal, transversal e horizontal. Nas seções subsequentes, serão apresentados os conceitos de algumas medidas que caracterizam a geometria da via permanente, assim como será exposto o significado físico e o que tais medidas podem trazer de informação.

2.3.1.1. Bitola

Pelo fato de definir a base do rolamento dos veículos da ferrovia, a bitola pode ser considerada como o parâmetro de maior importância na definição das características geométricas da via. De acordo com a NBR 16387 ABNT (2020), denomina-se bitola a distância perpendicular aos trilhos da via, medida entre as faces internas dos boletos quando sem carregamento lateral. Apresenta-se na Figura 2.8 o ponto correto de medição, feita a 16 mm do topo do boleto.

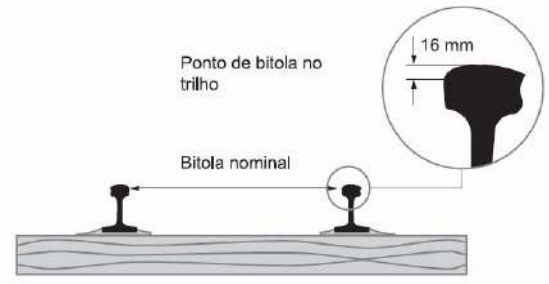


Figura 2.8: Técnica de medição da bitola.
Fonte: Retirado de ABNT (2020).

Quando a medida do valor da bitola ultrapassa os limites de tolerância estabelecidos para a mesma, então pode-se dizer que existe um defeito que pode ser positivo ou negativo. Em outras palavras, a bitola pode ser avaliada sob dois aspectos: alargamento (Figura 2.9) e estreitamento (Figura 2.10).

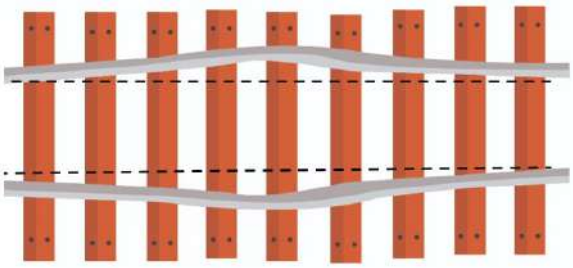


Figura 2.9: Alargamento de bitola.
Fonte: O autor.

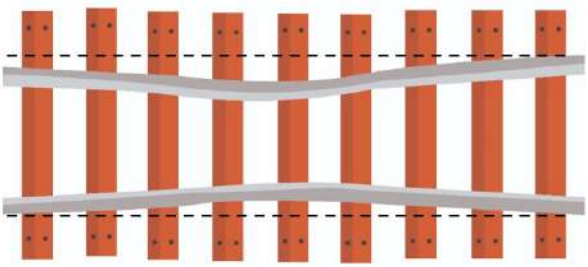
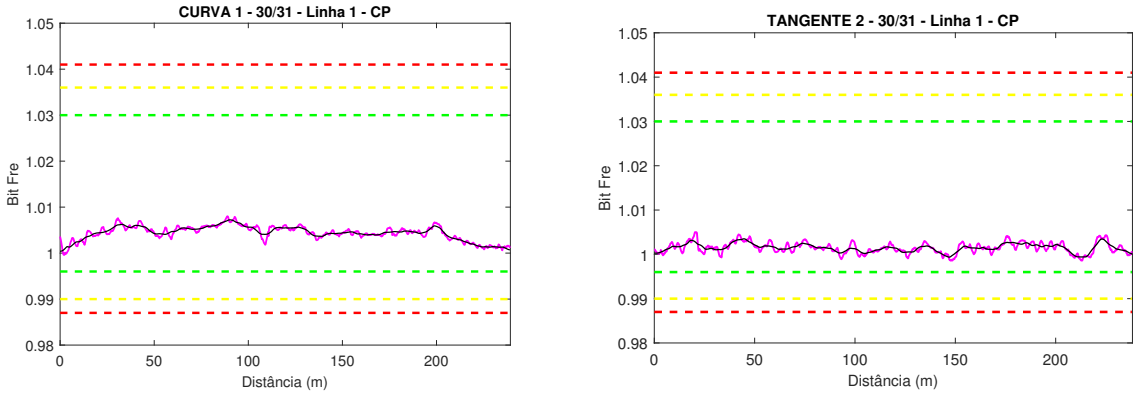


Figura 2.10: Estreitamento de bitola.
Fonte: O autor.

A seguir, apresenta-se nas figuras 2.11a e 2.11b dois gráficos contendo exemplos de sinais de bitola; o primeiro para uma curva e o segundo para uma tangente. Como a ferrovia possui uma bitola métrica, é possível perceber as oscilações do sinal em torno desse valor. Vale ressaltar que tanto os gráficos apresentados nesta seção quanto os das seções subsequentes são dados reais obtidos pela 3ª inspeção do CC na EFVM ao longo do ano de 2020. As linhas pontilhadas indicam os limites aceitáveis de acordo com a norma. No caso da bitola, como pode-se observar nas figuras 2.11a e 2.11b, existe uma tolerância maior de alargamento do que de estreitamento.



(a) Curva. (b) Tangente.

Figura 2.11: Sinal de bitola para dois elementos da EFVM.
Fonte: O autor.

2.3.1.2. Empeno

O conceito de empeno é importante para indicar uma possível instabilidade do vagão. Segundo Coimbra (2008), para entender melhor o conceito de empeno, consideram-se quatro pontos (A, B, C e D) sobre a superfície de rolamento dos trilhos, dois em cada trilho, formando um retângulo, assim como ilustrado na Figura 2.12. Define-se empeno como a distância vertical (y) de um dos pontos selecionados (B' ou D') ao plano formado pelo retângulo.

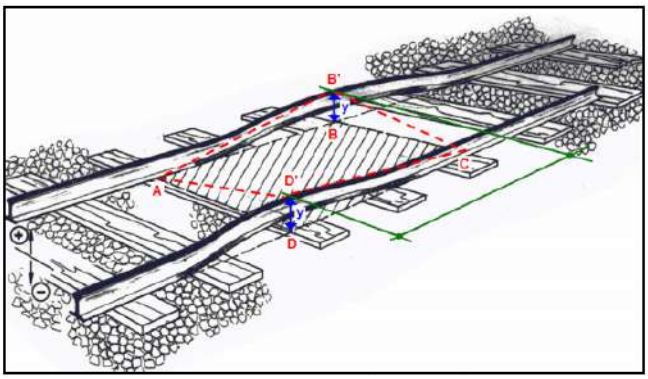
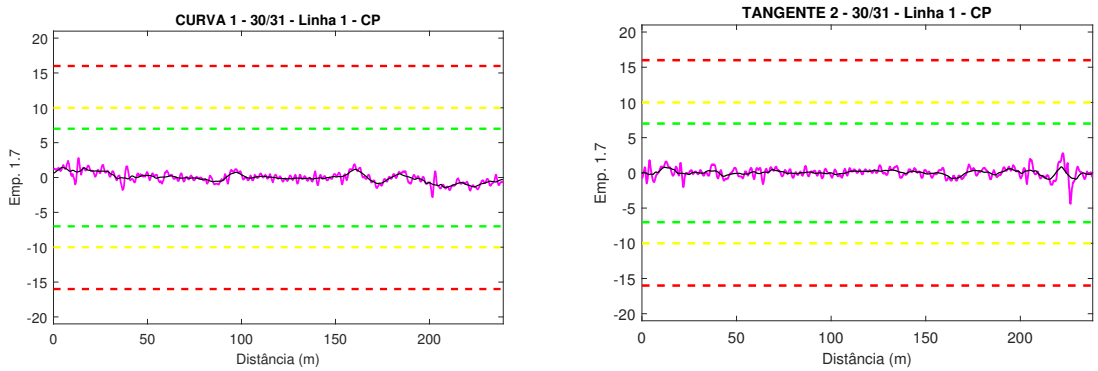


Figura 2.12: Ilustração do empeno ou torção.
Fonte: Retirado de Coimbra (2008).

Apresenta-se nas figuras 2.13a e 2.13b o sinal de empeno para uma curva e uma tangente, respectivamente. Como pode ser visto, para ambos os casos, o sinal oscila em torno do ponto de amplitude nula.



(a) Curva. (b) Tangente.

Figura 2.13: Sinal de empeno para dois elementos da EFVM.
Fonte: O autor.

2.3.1.3. Nivelamento Longitudinal

Alguns fatores como o excesso de carga transportada pelos veículos, ou mesmo o mau acondicionamento da carga, podem causar sobre-esforços verticais do trilho que aceleram o processo de degradação do parâmetro nivelamento. Este desvio pode ser no sentido longitudinal ou transversal da via permanente e é medido separadamente em cada trilho.

Para verificar o desnivelamento longitudinal da via, deve-se comparar o nivelamento da linha férrea medindo a deformação vertical (X) de um ponto qualquer na superfície de rolamento de um trilho em relação ao segmento de reta formado pelo plano horizontal original, assim como ilustrado na Figura 2.14. Segundo a ABNT NBR 16387 ABNT (2020), o desnivelamento do perfil longitudinal é a medida de flecha vertical em uma mesma fila de trilhos, medida no topo de trilho, no centro de uma corda de 20 m.

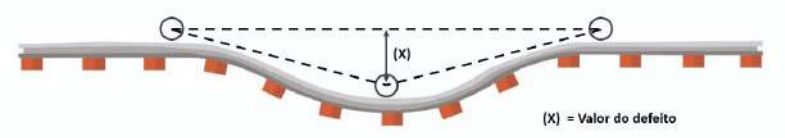
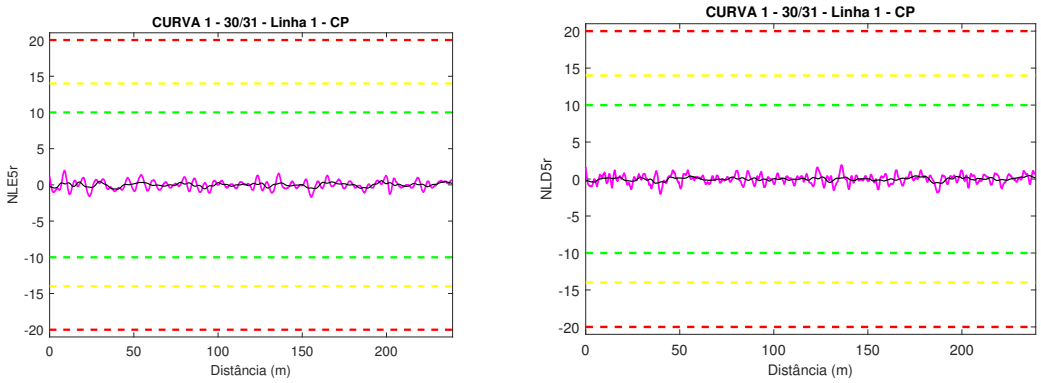


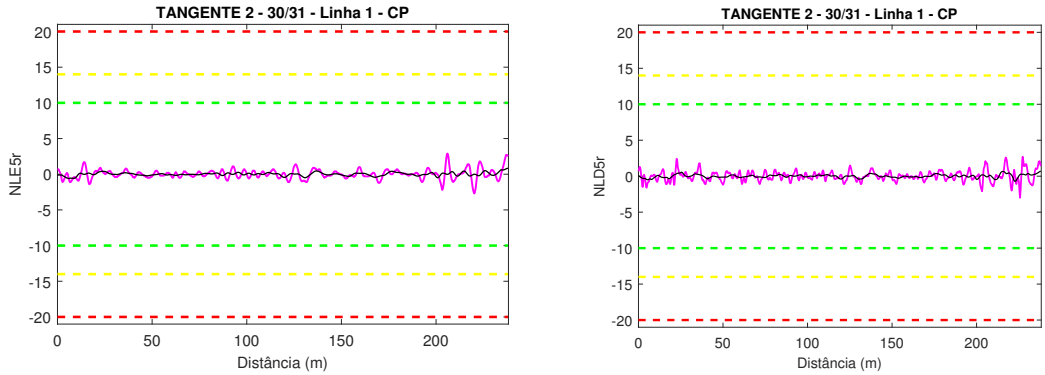
Figura 2.14: Ilustração de uma linha desnivelada longitudinalmente.
Fonte: O autor.

Apresenta-se nas figuras 2.15a e 2.15b os sinais de nivelamento longitudinal, esquerdo e direito, para uma curva. As figuras 2.16a e 2.16b indicam o mesmo sinal, mas para uma tangente. Os sinais de nivelamento também devem oscilar em torno do ponto de amplitude nula.



(a) Nivelamento esquerdo. (b) Nivelamento direito.

Figura 2.15: Sinal de nivelamento para uma mesma curva da EFVM.
Fonte: O autor.



(a) Nivelamento esquerdo. (b) Nivelamento direito.

Figura 2.16: Sinal de nivelamento para uma mesma tangente da EFVM.
Fonte: O autor.

2.3.1.4. Nivelamento Transversal

Antes de compreender o conceito do nivelamento transversal é importante definir um outro conceito, o da superelevação. A superelevação nada mais é do que a inclinação da seção transversal de uma pista de rolamento (em relação ao eixo da estrada), com o objetivo de contrabalancear a força centrífuga e os esforços laterais, de modo a auxiliar o veículo a realizar a curva de maneira mais confortável e segura. O desnivelamento transversal pode ocorrer tanto na tangente quanto na curva. Na tangente, que não tem uma superelevação, sua amplitude é simplesmente a diferença de nível entre os dois trilhos no plano horizontal. Na curva horizontal, que tem uma superelevação, o desnivelamento é a diferença de nível entre os dois trilhos menos a superelevação de projeto da curva. Apresenta-se na Figura 2.17, uma ilustração de medição das superelevações (S_A e S_B) em dois pontos (A e B).

Também é possível identificar uma outra medida, a variação do nivelamento transversal Δ entre os dois pontos. Ela é, na verdade, a taxa de variação de cota entre o topo dos trilhos medida em duas secções transversais (ABNT, 2020).Tendo em vista o que foi citado anteriormente, é possível imaginar que, pela forma de medição, esse sinal terá um comportamento

diferente para curvas e tangentes. A superelevação terá uma amplitude maior no meio curva, ao passo que oscila em torno de zero para o elemento tangente.

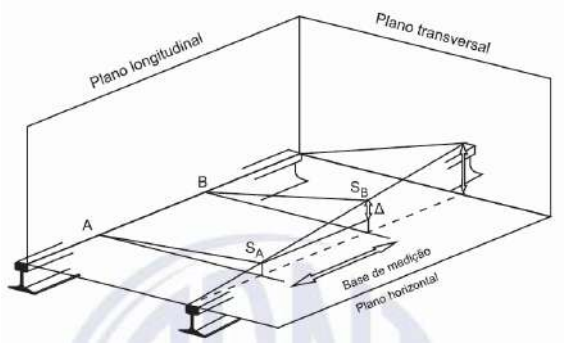
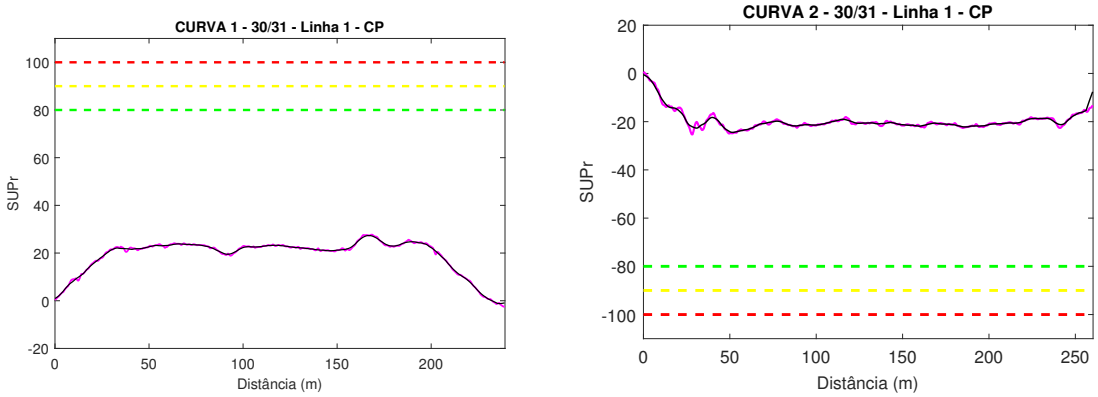


Figura 2.17: Ilustração da medição de uma superelevação.
Fonte: Retirado de ABNT (2020).

Essa diferença, conforme supracitado, pode ser verificada pelas figuras 2.18a, 2.18b e 2.19 que, respectivamente, apresentam os sinais de superelevação para uma curva percorrida no sentido horário e anti-horário e para uma tangente.



(a) Sentido horário. (b) Sentido anti-horário.

Figura 2.18: Sinal de superelevação para duas curvas da EFVM.
Fonte: O autor.

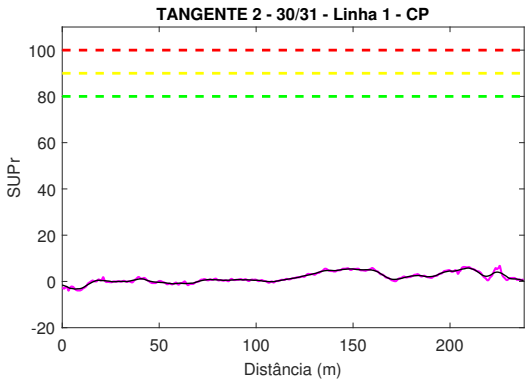


Figura 2.19: Sinal de superelevação para uma tangente da EFVM.
Fonte: O autor.

2.3.1.5. Alinhamento

Se uma corda estendida em dois pontos laterais do boleto de um trilho de um trecho em tangente evidencia uma flecha, conforme indicado pela Figura 2.20, significa que a linha está desalinhada. A amplitude do desalinhamento corresponde ao tamanho da flecha (X) formada por uma corda (de tamanho 10 ou 20 metros) estendida entre 2 pontos. De acordo com a ABNT NBR 16387 ABNT (2020), o alinhamento em vias de bitola métrica, como é o caso da EFVM, consiste na variação máxima de flecha horizontal entre pontos adjacentes, medida a cada 2,5 m no centro de corda (10 m), tanto em curvas quanto em tangentes.

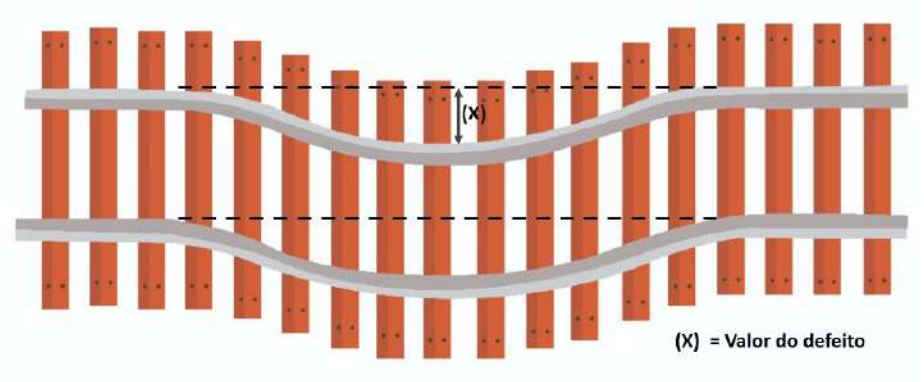
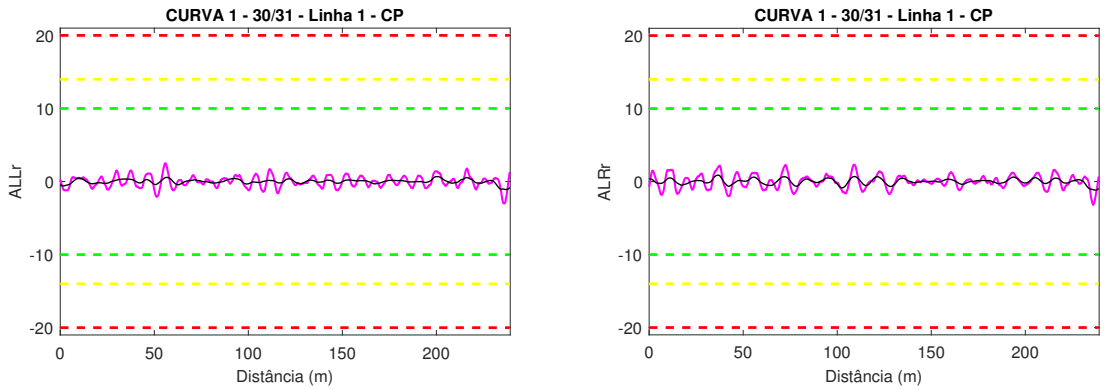


Figura 2.20: Ilustração de desalinhamento em uma tangente.
Fonte: O autor.

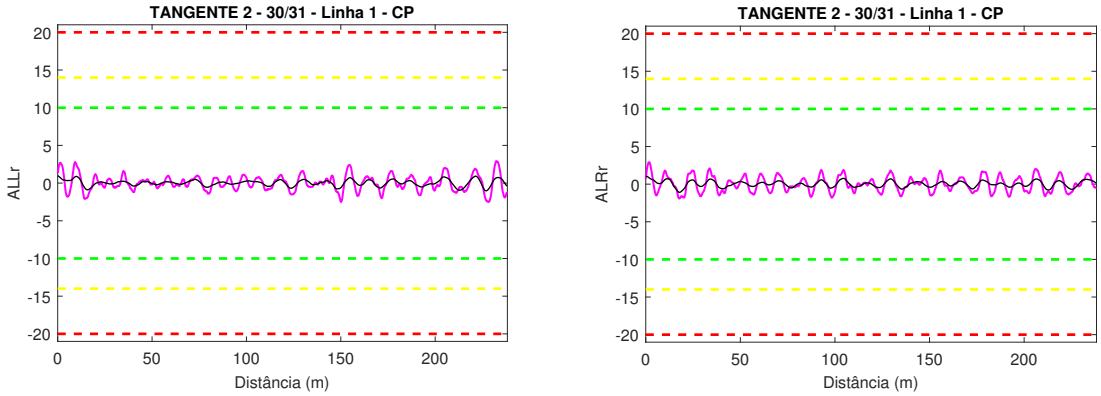
Apresentam-se nas figuras 2.21a e 2.21b os sinais de alinhamento esquerdo e direito para uma curva. As figuras 2.22a e 2.22b indicam o mesmo sinal, mas para uma tangente. Como pode-se observar, tanto os sinais das curvas quanto tangentes oscilam em torno da amplitude nula, como era de se esperar.



(a) Alinhamento esquerdo.

(b) Alinhamento direito.

Figura 2.21: Sinal de alinhamento para uma mesma curva da EFVM.
Fonte: O autor.



(a) Alinhamento esquerdo.

(b) Alinhamento direito.

Figura 2.22: Sinal de alinhamento para uma mesma tangente da EFVM.
Fonte: O autor.

2.3.2. Localização dos Dormentes de Aço Danificados

Os dados disponibilizados pela Vale SA para a identificação dos rótulos de cada dormente, estão presentes em um único arquivo, referenciado neste trabalho como **planilha de prospecção**, contendo todos os elementos da EFVM. Vale ressaltar que o processo de rotulação dos dormentes acontece de forma manual pelos inspetores da Vale SA. A partir dos dados, realizou-se uma busca e, ao mesmo tempo, uma filtragem daqueles elementos incluídos no experimento, a fim de gerar uma nova base de dados contendo os rótulos apenas desses elementos, a qual, neste trabalho, será chamada apenas de **base rotulada**. Esse processo foi realizado por meio de uma rotina implementada no ambiente de programação do *Matlab* e visou facilitar o desenvolvimento do algoritmo de leitura dos dados que seria realizado posteriormente.

A base de dados rotulada traz informações importantes sobre a condição estrutural de cada dormente de aço presente na via permanente, de modo que seja possível determinar a posição daqueles que se encontram trincados ou fraturados. A Tabela 2.1, por exemplo, ilustra um segmento de planilha contendo informações sobre os rótulos dos dormentes de um elemento específico (Curva 2 - Conselheiro Pena - EH 30/31 - Linha 1).

Os arquivos de rótulos da base de dados são utilizados para a determinação da posição dos dormentes de aço danificados ao longo da extensão do elemento, ou seja, assume-se que N dormentes se encontram igualmente espaçados entre si de uma distância D dentro da extensão L do elemento. Dessa forma, o k-ésimo dormente deverá estar localizado a uma distância X_i contabilizada a partir da localização inicial do elemento, de acordo com a Equação 2.1:

$$X_i = k\frac{L}{N} \tag{2.1}$$

Dessa forma, é realizada a determinação das posições dos dormentes danificados, contabilizada a partir da localização inicial de cada elemento de acordo com as informações disponíveis na planilha de prospecção. Um fato crucial para essa localização é que cada inspeção realizada pelo CC, para um mesmo elemento, pode sofrer alterações quanto ao quilômetro ini-

cial e final da inspeção. Nesse sentido, haverá uma diferença entre o início e fim do elemento indicado pela planilha de prospecção e a indicada pela inspeção do CC. Um caso particular dessa diferença supracitada pode ser analisado pela Figura 2.23.

Tabela 2.1: Segmento da base rotulada para um dos elementos da EFVM.

Indice	Rótulo	Dist. Inicial L	Dist. Final	Extensão
1	Bom	187009	187242	233
2	Bom	187009	187242	233
3	Bom	187009	187242	233
4	Trincado	187009	187242	233
5	Trincado	187009	187242	233
6	Bom	187009	187242	233
7	Bom	187009	187242	233
8	Bom	187009	187242	233
9	Fraturado	187009	187242	233
10	Fraturado	187009	187242	233
11	Bom	187009	187242	233
12	Bom	187009	187242	233
13	Bom	187009	187242	233
14	Bom	187009	187242	233
15	Bom	187009	187242	233
16	Bom	187009	187242	233

Fonte: O autor.

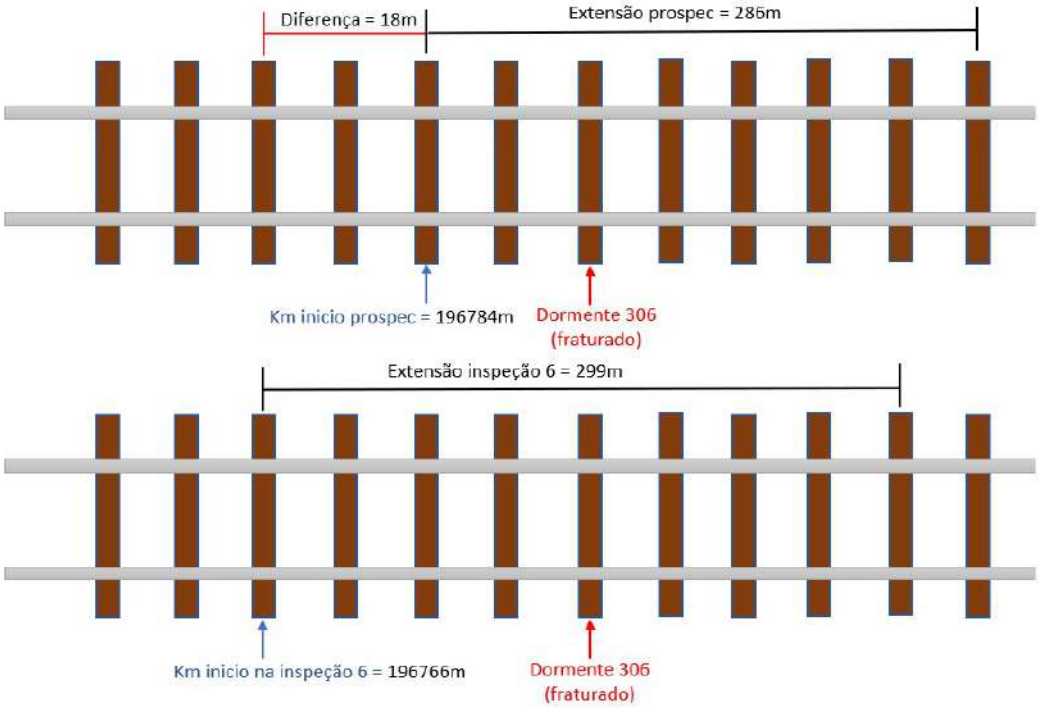


Figura 2.23: Diferença entre os dados da planilha de prospecção e da base rotulada.
Fonte: O autor.

É possível perceber que a extensão do elemento, indicada pela planilha de prospecção, é de 268 metros, enquanto a extensão indicada pela inspeção do CC é de 299 metros. Adicionalmente, pode existir uma diferença do quilômetro inicial fixo indicado pela planilha de prospecção e o quilômetro inicial em que o CC começou coletar os dados para aquele elemento específico. Por esses motivos, para todos os elementos, são realizados ajustes dessa diferença para encontrar a posição correta do defeito indicado pela planilha de prospecção.

Para encontrar, por exemplo, a posição do dormente (306) defeituoso indicada na Figura 2.23, primeiramente é realizada a diferença do quilômetro inicial da prospecção pelo quilômetro inicial da inspeção, que nesse caso resulta em uma diferença de 18 metros. Posteriormente, aplicam-se os dados do elemento à Equação 2.1 para obter a posição do defeito de acordo com a planilha de rótulos (sem ajuste). Ao final, a posição encontrada pela equação é somada pela diferença dos quilômetros iniciais calculada inicialmente, encontrando, assim, a posição de defeito ajustada.

$$X_i = 306 * \frac{286}{480} = 360 * 0,5958 = 214,5$$

$$X_i(Ajustado) = X_i + 18 = 232,5$$

As informações acerca das posições dos dormentes danificados são fundamentais para a extração do conjunto de dados correspondentes aos trechos da EFVM em que este problema ocorre. Isso é válido também para a posição dos dormentes saudáveis, de forma que aconteça a extração correta do conjunto de dados daquelas posições.

2.3.3. Construção da Base de Dados

O CC realiza a aquisição de dados durante inspeções bimestrais, produzindo um conjunto de arquivos nos quais, em geral, encontram-se identificados o ano, a inspeção (de 1 a 6), o trecho (Entre-*Housing*) e o sentido (ida ou volta). Cada arquivo traz informações acerca dos quilômetros inicial e final de cada elemento (curva ou tangente) e o número da linha correspondente ao sentido da ferrovia percorrido pelo CC. Além disso, eles contém diversas informações relacionadas às características geométricas da via permanente, a posição espacial onde as amostras foram lidas, dentre outras. Um fato importante que deve ser destacado é que o período de amostragem espacial dos dados é de 25 cm e, portanto, menor do que a distância média entre dormentes adjacentes, que varia tipicamente entre 55 cm e 60 cm.

Os dados coletados pelas inspeções do CC, para os elementos da EFVM, são disponibilizados pela Vale SA em uma série de arquivos. Esses arquivos, por sua vez, possuem uma enorme quantidade de dados provenientes do CC, mas nem todos de interesse para a pesquisa. Portanto, busca-se encontrar, dentre toda a base disponível, apenas os dados daqueles elementos que fazem parte do experimento desta pesquisa e, além disso, buscar apenas as informações relevantes. Dessa forma, criou-se um algoritmo em ambiente *Matlab* capaz de gerar novos ar-

quivos (planilhas) contendo os dados individuais e de interesse para cada um dos elementos em estudo. Essa nova base será referenciada neste trabalho apenas como **base tratada**.

A base tratada, portanto, é formada apenas pelos elementos cujos dados se encontram íntegros em todas as inspeções realizadas no ano. De fato, essa etapa de tratamento é importante e possui impacto direto nas atividades subsequentes, uma vez que é necessário garantir que os elementos utilizados nos experimentos possuam todas as informações consistentes.

2.3.3.1. *Data set* 1

O trabalho foi dividido em duas fases, sendo que para a primeira delas utilizou-se os dados do ano de 2019, por estarem disponíveis no momento. Inicialmente, foram destacados 528 elementos para a investigação, dos quais optou-se por considerar como possíveis candidatos a serem incluídos nos experimentos os 320 elementos que apresentaram todos os arquivos e dados completos. Além disso, foram definidas três supervisões para efeito de inclusão nos protocolos experimentais: Conselheiro Pena (CP), Governador Valadares (GV) e Mário Carvalho (MR). Assim, considerando-se todos os elementos pertencentes a estas três supervisões e que se encontravam presentes na relação dos 320 possíveis candidatos, foram selecionados ao todo 201 elementos, dos quais 113 são curvas e 88 são tangentes, totalizando uma extensão de 68043 metros (33865 metros em curvas e 34178 metros em tangentes) da EFVM.

As curvas investigadas possuem um total de 57480 dormentes instalados, dos quais 862 são danificados (1,5%), enquanto as tangentes possuem um total de 52894 dormentes instalados, dos quais 215 são danificados (0,41%). Portanto, considerando os elementos da EFVM investigados neste projeto, tem-se um total de 110374 dormentes instalados, dos quais 1077 apresentam trincas ou fraturas (0,98%).

No protocolo experimental, optou-se pela utilização dos dados provenientes da 3ª, 4ª e 5ª inspeção para efeito de treinamento/validação dos modelos, uma vez que estas são realizadas entre o outono e a primavera, período no qual, geralmente ocorrem temperaturas médias mais baixas e uma menor pluviosidade, resultando em um menor impacto climático nas características de operação da via permanente. Os dados da 6ª inspeção de 2019 foram utilizados na etapa de teste dos modelos. Durante a escolha dos elementos tanto para o *data set* 1 quanto para o *data set* 2, foram verificados o comportamento dos dados para as 4 inspeções, para que nenhum evento anormal tivesse ocorrendo em uma delas e, consequentemente, criasse um enviesamento. Para todos os experimentos nesta pesquisa, o desempenho mostrado é calculado em um conjunto de teste que não foi usado para treinamento ou seleção de modelo/parâmetro.

2.3.3.2. *Data set* 2

Para a segunda fase, com a disponibilidade dos dados do ano de 2020, optou-se pela utilização dos novos dados a fim de validar a sistematização do processo. Inicialmente foram destacados 555 elementos para a investigação, dos quais optou-se por considerar como possíveis

candidatos a serem incluídos nos experimentos, os 362 que apresentaram todos os arquivos e dados completos ao longo das inspeções realizadas. Assim como no procedimento anterior, foram escolhidas apenas os elementos das supervisões de CP, GV e MR, totalizando 225 elementos, dos quais 132 são curvas e 93 são tangentes. A extensão total em que estes elementos estão compreendidos é de 78151 metros (39461 em curvas e 38690 em tangentes) da EFVM.

As curvas investigadas possuem um total de 67556 dormentes instalados, dos quais 705 são danificados (1,04%), enquanto as tangentes possuem um total de 60848 dormentes instalados, dos quais 275 são danificados (0,45%). Portanto, considerando os elementos da EFVM investigados neste projeto, tem-se um total de 128404 dormentes instalados, dos quais 980 apresentam trincas ou fraturas (0,76%).

Assim como no Data set 1, optou-se pela utilização dos dados provenientes da 3ª, 4ª e 5ª inspeção para efeito de treinamento/validação dos modelos. Da mesma forma, os dados da 6ª inspeção de 2020 foram utilizados na etapa de teste dos modelos.

2.4. Preparação dos Dados

Na fase de preparação ou representação dos dados, busca-se aplicar as ferramentas necessária para extrair informações dos dados e prepará-los para a fase seguinte (modelagem). Os sinais espaciais associados à geometria da via permanente, dentre outros, possuem informações intrínsecas que, a princípio, caracterizam o estado de operação de componentes da infraestrutura e da superestrutura ferroviária de forma interligada. Assim, as técnicas de extração de características buscam extrair informações capazes de revelar aspectos de interesse, como por exemplo os padrões de comportamento relacionados com danos estruturais.

Os avanços nas tecnologias de armazenamento de dados tem contribuído para fazer com que existam enormes volumes de dados disponíveis a todos. São exemplos dessas tecnologias, os dispositivos com maior capacidade de armazenamento e mais baratos, além de sistemas de gerenciamento de banco de dados mais eficientes, da tecnologia de *Data Warehousing* e do *World Wide Web*. O que impressiona também é a distância crescente entre a geração de dados e a capacidade de analisá-los e compreendê-los. À medida que o número de dados aumenta, a proporção dos dados que é compreendida e analisada pelas pessoas diminui.

A análise desenvolvida sobre a via permanente depende das condições dos dados, uma vez que os métodos empregados para treinamento e classificação não podem ser baseados em informações inconsistentes. Frequentemente, os dados apresentam diversos problemas tais como dados ruidosos (valores incorretos para os atributos), grande desproporção entre o número de exemplos de cada classe, grande quantidade de valores desconhecidos, entre outros; os quais podem impactar de forma negativa no desempenho do sistema de classificação.

No caso deste trabalho, instabilidades e erros inesperados durante a leitura dos sensores do carro controle podem produzir inconsistências, como por exemplo dados não numéricos ou até amplitudes incompatíveis com o restante das leituras em um momento específico. Com base

nas informações expostas, justifica-se a preparação dos dados, para que as etapas posteriores, principalmente a de modelagem, não tenha a sua capacidade de generalização comprometida. Além disso, técnicas como o filtro média-móvel não-causal e janelamento foram aplicadas durante o processo de preparação dos dados.

2.4.1. Filtro de Média-Móvel Não-Causal

No contexto do pré-processamento, o filtro média-móvel é empregado principalmente com a finalidade de permitir a subtração do valor médio da medida de superelevação das curvas, de modo a se obterem apenas as oscilações. Além disso, também tem o intuito de aplainar picos registrados no sinal, viabilizando o uso de operações matemáticas como derivadas e outras medidas calculadas durante a etapa de extração de características, uma vez que variações drásticas podem causar inconsistências nos cálculos.

Um sistema de média-móvel não-causal é definido como aquele cuja saída correspondente a n-ésima amostra é calculada a partir da média entre os M_1 valores que a precedem e os M_2 valores que a sucedem, incluindo também a própria n-ésima amostra. Essa média, portanto, é calculada a partir de $M_1 + M_2 + 1$ valores. Um sistema não-causal é definido como aquele que não depende apenas de valores passados e presentes de um sinal, porém também de valores futuros, conforme descrito na seguinte equação (OPPENHEIM *et al.*, 2010):

$$y[n] = \frac{1}{M_1 + M_2 + 1} \sum_{k=-M_1}^{M_2} x[n-k] \tag{2.2}$$

2.4.2. Janelamento

As principais características geométricas da via permanente, que se encontram presentes nos dados coletados pelo CC e que foram utilizadas neste trabalho, quais sejam, a bitola, superelevação, empeno, nivelamento e alinhamento, serão referenciados, doravante, como sinais. Cada um destes sinais apresenta um comportamento que varia ao longo da extensão do elemento, inclusive podendo indicar a presença de diferentes tipos de problema, incluindo aqueles relacionados aos dormentes de aço, que são o objeto de estudo desta pesquisa. Assim, torna-se necessário a utilização de janelas de dados para efeito de análise dos sinais de interesse.

O tamanho da janela de dados deve ser obtido de modo a se estabelecer o melhor compromisso entre a quantidade de informações presentes nas amostras contidas na janela e a resolução espacial. Em geral, quanto menor for o tamanho da janela, maior será a resolução espacial, porém conterá menos informação. As janelas maiores possuem mais informações e menos resolução espacial. Para este propósito, empregou-se a janela retangular, que representa um truncamento simples a partir das amostras inicial e final contidas na mesma (OPPENHEIM *et al.*, 2010). Isso implica que todos as amostras localizadas no interior da janela são conside-

radas com o mesmo peso, o que corresponde a utilização dos valores originais das medições.

Cada janela espacial é utilizada para a determinação de um vetor de parâmetros na etapa de extração de características. O n-ésimo vetor de parâmetros corresponde ao n-ésimo deslocamento da janela de dados, considerando-se uma janela deslizante com incremento unitário. Além disso, a quantidade de amostras contidas na janela também impacta na quantidade de parâmetros que serão extraídos, podendo contribuir para um aumento demasiado da dimensionalidade do problema, dependendo da técnica de parametrização utilizada. Assim, neste trabalho foram avaliadas janelas de dados contendo 32, 64 e 128 amostras, sendo o deslocamento da janela unitário, ou seja, de apenas uma amostra. A escolha do maior tamanho da janela de dados (128) foi baseada na extensão de alguns dos elementos da EFVM, para que o tamanho da janela não fosse maior que do sinal. A partir disso, foram selecionados as outras duas janelas: uma segunda com a metade do tamanho da primeira (64) e uma terceira com a metade do tamanho da segunda (32).

2.4.3. Extração de Características

Após a compreensão dos dados, uma sequência de análises podem ser realizadas para revelar características que as medições originais não apresentam de forma explícita, porém que podem contribuir para o estabelecimento dos padrões associados aos danos estruturais investigados. Estas análises podem ser no domínio do tempo ou no domínio da frequência. Os métodos no domínio do tempo são geralmente sensíveis à falhas de natureza impulsiva. Dessa forma, a principal aplicação dessas estratégias consiste na identificação da ocorrência de eventos de curta duração e sua taxa de repetição. Estes métodos, apesar de alertar para o surgimento e o desenvolvimento de uma falha, muitas vezes não permitem um diagnóstico preciso ou mesmo não localizam o defeito (MESQUITA *et al.*, 2002). Por outro lado, os métodos no domínio da frequência assumem que os sinais analisados possuem componentes com características periódicas e, portanto, um defeito qualquer pode produzir um sinal periódico na frequência característica do defeito (SANTIAGO e PEDERIVA, 2003).

Em linhas gerais, os algoritmos de extração criam novas características a partir de transformações ou combinações do conjunto original. Neste trabalho, três abordagens foram utilizadas: análises baseadas em parâmetros estatísticos e espaciais, a transformada de Fourier (FT, do inglês, *Fourier Transform*), a transformada *wavelet*, (WT, do inglês, *Wavelet Transform*). É importante reforçar que os sinais utilizados na pesquisa (Seção 2.3.1) estão em um domínio espacial, ou seja, sua amplitude varia de acordo com o deslocamento no espaço e não no tempo. Portanto, para toda a fundamentação teórica sobre as técnicas temporais de extração de características detalhadas nesta seção pode ser também aplicada no presente problema, substituindo-se a variável tempo por espaço. As transformadas não são limitadas a funções temporais, contudo para fins de convenção, o domínio original é comumente referido como domínio do tempo. Analogamente, considerando o contexto da pesquisa, quando fala-se

sobre a frequência, deve-se entender frequência espacial. A frequência espacial é uma característica de qualquer estrutura que é periódica ao longo da posição no espaço. Ela mede a frequência com que os componentes senoidais da estrutura se repetem por unidade de distância. No Sistema Internacional de Unidades (SI) a frequência espacial é dada em ciclos por metro.

2.4.3.1. Atributos Espaciais

Na abordagem baseada em atributos estatísticos e espaciais, as seguintes informações foram extraídas a partir do sinal contido em cada janela de dados: energia, variância, ZCR, diferença entre as amplitudes máxima e mínima (variação da amplitude), derivada primeira e segunda e função de autocorrelação dos valores da medida de geometria da via permanente (nivelamento, empeno, superelevação, bitola e alinhamento) contidos dentro da janela de dados. Por convenção, considerando a característica espacial dos sinais estudados, essas medidas serão tratadas nesta pesquisa como atributos espaciais. Apresenta-se a seguir uma breve explicação sobre cada um dos atributos utilizados.

- **Derivadas:** o uso das derivadas implica em uma análise direta sobre a taxa de variação do sinal, de modo que a derivada primeira indica a variação dos valores da variável de interesse, enquanto a segunda derivada mede a taxa de variação da própria variação desta função (STEWART, 2013). Por exemplo, a derivada de segunda ordem da posição de um objeto em relação ao tempo é a aceleração instantânea deste objeto. Essa aceleração, por sua vez, mede a taxa de variação da velocidade deste mesmo objeto. Matematicamente, define-se uma derivada discreta como:

$$f'(n) = f(n+1) - f(n) \tag{2.3}$$

  Para a obtenção da segunda derivada, basta aplicar a Equação 2.3 sobre o resultado obtido na primeira derivada.

- **Taxa de Cruzamento por Zero:** busca indicar o quanto um sinal cruza a linha entre valores positivos e negativos. Para calcular essa taxa, basta comparar cada amostra com a seguinte e avaliar se seu sinal mudou. Caso afirmativo, isso indica que naquele ponto houve cruzamento por zero.

- **Diferença de Amplitude Máxima e Mínima:** é a diferença entre o maior e o menor valor amostral para cada intervalo definido por uma janela de tamanho M.

- **Energia:** de acordo com Oppenheim *et al.* (2010), é a soma dos quadrados das amplitudes das amostras dentro da janela de tamanho M, conforme a Equação 2.4, e por isso apresenta grande crescimento quando contém valores com amplitude elevada.

$$\sum_{k=-M/2}^{M/2} |x[n]|^2 \tag{2.4}$$

- **Função de Autocorrelação:** realiza a correlação de uma função consigo mesma, avaliando a autossimilaridade para diferentes deslocamentos nas comparações, conforme definido pela Equação 2.5 (OPPENHEIM *et al.*, 2010).

$$\varphi[n] = \sum_{k=-\infty}^{\infty} x[k]x[n-k] \tag{2.5}$$

A fim de exemplificar, considera-se uma determinada janela de dados com tamanho 32. Considerando a extração de características para os 7 sinais utilizados, serão produzidos 266 atributos, conforme ilustrado na Figura 2.24. No caso de utilização da maior janela (contendo 128 amostras) serão produzidos 938 atributos. Neste exemplo, as linhas representam os setes sinais da geometria da via permanente, na ordem: Bitola (Bit Fre), Superelevação (SUPr), Empeno (Emp. 1.7), Nivelamento Esquerdo (NLE5r), Nivelamento Direito (NLD5r), Alinhamento Esquerdo (ALLr) e Alinhamento Direito (ALRr).

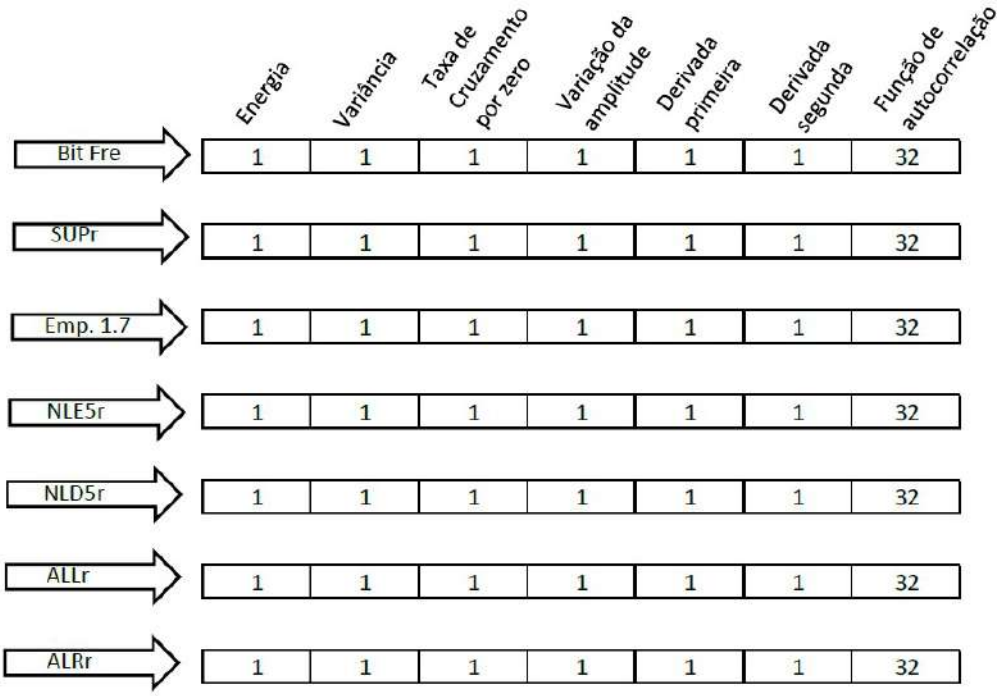


Figura 2.24: Exemplo de parametrização espacial para uma janela de tamanho 32.
Fonte: O autor.

2.4.3.2. Transformada de Fourier

A transformada de Fourier é uma ferramenta matemática que permite representar e estudar certos sinais com respeito a sua periodicidade. Ela tem sido largamente aceita como sendo um método extremamente confiável no diagnóstico de falhas para sinais estacionários (SANTIAGO e PEDERIVA, 2003), estabelecendo a representação de um sinal por uma soma infinita de termos em senos e cossenos. A análise espectral, comumente associada à FT, é extremamente útil, porque a frequência contida no sinal é de grande importância.

A análise espectral utilizada no trabalho busca realizar uma representação das medidas de geometria da via permanente originais, porém alterando seu domínio natural, o espaço, para o domínio espectral. Em aplicações práticas, a utilização da transformada clássica (empregando soluções analíticas) não é comum, pelo fato dos sinais de interesse muitas vezes não possuírem expressões analíticas para descrevê-los. Para isso, é empregada uma transformação matemática chamada de transformada discreta de Fourier (DFT, do inglês, *Discrete Fourier Transform*), uma ferramenta extremamente valiosa para análise na frequência de sinais no tempo discreto. A DFT permite uma representação discreta no domínio da frequência para sinais no tempo discreto. Como a representação na forma de sequências numéricas é natural para computadores digitais, a DFT é uma ferramenta muito poderosa, pois permite manipular informações no domínio da frequência da mesma forma que podemos manipular as sequências originais. Normalmente, utiliza-se de um algoritmo conhecido como transformada rápida de Fourier (FFT, do inglês, *Fast Fourier Transform*), o qual não é um tipo diferente de transformada e sim uma técnica que possibilita calcular a DFT de forma mais rápida e econômica, proporcionando um menor esforço computacional (OLIVEIRA, 2007).

A DFT pode ser definida como um somatório do produto de exponenciais complexas com a função original, espaçadas igualmente e localizadas em um intervalo limitado do sinal (DINIZ *et al.*, 2014).

$$X(e^{j(\frac{2\pi}{N})k}) = \sum_{k=0}^{N-1} x[n]e^{-j(\frac{2\pi}{N})kn} \tag{2.6}$$

Nessa segunda abordagem, calcula-se a DFT para o sinal contido no interior da janela por meio do algoritmo da FFT, seguido da determinação do módulo da amplitude, o que tipicamente produz, para cada sinal, uma quantidade de parâmetros que é metade do tamanho da janela utilizada. Assim, a janela com 32 amostras produzirá 112 parâmetros (7 sinais × 16 módulos de amplitude), enquanto a janela contendo 128 amostras produzirá 448 parâmetros (7 sinais × 64 módulos de amplitude).

2.4.3.3. Transformada *Wavelet*

A fim de contornar o problema de perda da identidade temporal, diversas soluções tem sido desenvolvidas, dentre elas, a transformada *wavelet* vem sendo bastante utilizada. A análise tempo-frequência tem em vista conciliar a unilateralidade existente quando se utiliza domínios puramente espaciais ou espectrais por meio da representação de informações em ambos os domínios. Nesse sentido, Mallat (1998) define que as transformadas *wavelet* atuam não sobre uma linha de domínio temporal ou espacial, mas sobre um plano. Segundo Cox (2004), a WT pode analisar variações espectrais com diferentes resoluções tempo-frequência; ou seja, ela permite variar o tamanho da janela de análise de acordo com a frequência do sinal. De maneira geral, a *wavelet* pode ser manipulada de dois modos: movendo-se para várias posições sobre o sinal e dilatando-se ou comprimindo-se.

A *wavelet* é uma onda curta de natureza oscilatória e energia finita, capaz de interpretar o sinal como versões deslocadas e escalonadas de uma *wavelet* original, denominada de *wavelet*-mãe (LACERDA *et al.*, 2011). As *wavelets* possuem diversas saídas, representando, para cada sinal, diferentes níveis de transformadas, os quais dependem do tamanho do vetor de dados de entrada e determinam a resolução espacial e espectral. Quando o objetivo é analisar sinais digitais, utiliza-se a transformada *wavelet discreta* (DWT, do inglês, *Discrete Wavelet Transform*), definida por:

$$DWT_f^{\psi}(j,k) = \frac{1}{\sqrt{a_0^j}} \sum_{n=-\infty}^{\infty} f(n)\psi \left[\frac{n - a_0^j k b_0}{a_0^j} \right] \tag{2.7}$$

Existem diversas famílias *wavelets*, cada qual possuindo uma função específica ψ, chamada de *kernel*, conforme indicado na Equação (2.7). Apresenta-se na Tabela 2.2, as famílias *wavelets* mais comumente utilizadas. No presente trabalho, decidiu-se testar os membros das famílias Daubechies, Biortogonal e Symlet para parametrização dos dados.

Tabela 2.2: Famílias *wavelets* mais comuns.

Família Wavelet	**Abreviação**	**Definição**
Haar	Haar	$\psi(x) = \begin{cases} 1, 0 \leq x < \frac{1}{2} \\ -1, \frac{1}{2} \leq x < 1 \\ 0, \text{caso contrário.} \end{cases}$
Symlets	Sym	Sem forma analítica
Daubechies	Db	Sem forma analítica, exceto Db1 (Haar)
Coiflets	Coif	Sem forma analítica
Biorthogonal	Bior	Sem forma anlítica
Meyer	Meyr	Sem forma anlítica
Discrete Meyer	Dmey	Sem forma analítica
Gaussian	Gaus	$\psi_n(x) = C_n \frac{d^n}{dt^n}(e^{-x^2})$
Mexican Hat	Mexh	$\psi(x) = (1-x^2)e^{\frac{x^2}{2}}$

Fonte: O autor.

Nessa terceira abordagem, cada janela de dados produz um determinado número de parâmetros, que depende da família e do nível de decomposição escolhido. As execuções iniciais utilizando a WT apontaram para um consumo de tempo muito elevado, em virtude do elevado número de possibilidades, fugindo do escopo desta pesquisa. Nesse sentido, após a execução de testes preliminares e com os primeiros resultados, optou-se por utilizar os parâmetros de aproximação da família *Symlet* (sym3), com apenas um nível de decomposição, pelo fato de apresentar o melhor resultado até o momento, dentre as variações testadas. Neste caso, para uma janela de tamanho 32, o cálculo da DWT fornece 126 parâmetros (7 sinais × 18 coeficientes de aproximação), enquanto para janela com 128 amostras, esta família (sym3) fornece 66

coeficientes de aproximação para cada um dos sete sinais considerados, o que resulta em um total de 462 parâmetros.

2.4.4. Seleção de Características e Redução de Dimensionalidade

Frequentemente, os problemas de classificação enfrentam a necessidade de tratar o aspecto da elevada dimensionalidade dos dados, com o intuito de garantir que o conjunto de dados disponível seja suficiente para permitir o ajuste dos parâmetros intrínsecos dos modelos classificadores na etapa de treinamento e também reduzir o custo computacional, que podem inclusive inviabilizar as etapas subsequentes. Existem algumas abordagens que possibilitam contornar esse tipo de problema, destacando-se as técnicas de seleção de características (atributos) e de redução de dimensionalidade. Os métodos de seleção de características visam determinar os parâmetros que são mais relevantes para o processo de classificação, principalmente do ponto de vista da discriminabilidade entre as classes, enquanto os métodos de redução de dimensionalidade visam realizar transformações nos parâmetros de entrada de modo a reduzir a sua dimensão e ao mesmo tempo tentando preservar as informações mais relevantes (com maior variância). Nas seções 2.4.4.2 e 2.4.4.3 serão apresentados dois métodos que utilizam as técnicas citadas, a razão discriminante de Fisher (FDR, do inglês, *Fisher's Discriminant Ratio*) e a análise de componentes principais (PCA, do inglês, *Principal Component Analyses*)

Conforme mencionado na seção anterior, as abordagens utilizadas na etapa de extração de características produzem uma grande quantidade de atributos, mesmo quando se considera a janela com 32 amostras. Em um espaço de alta dimensão, normalmente uma grande quantidade de dados de treinamento é necessária para garantir que haja várias amostras com cada combinação de valores. No caso desta pesquisa, o fator limitante é a quantidade de dormentes defeituosos presentes na base de dados, que corresponde a aproximadamente 1% do total de dormentes. Para o treinamento, são selecionados os dados referentes aos dormentes saudáveis na mesma proporção dos dados referentes aos dormentes defeituosos, fazendo com que o número de instâncias não seja consideravelmente grande.

Para exemplificar, considera-se a utilização da janela de dados de tamanho 32 para a parametrização espacial, o que resulta em um total de 266 atributos. Ao considerar uma supervisão/linha específica (por exemplo, Conselheiro Pena/linha 2) tem-se um total de 1778 amostras; um número muito baixo pela quantidade de parâmetros. Nesse sentido, para o maior tamanho de janela (128), torna-se um problema ainda maior.

2.4.4.1. A Maldição da Dimensionalidade

Atualmente, os grandes avanços em dispositivos de tecnologia, tanto para a coleta quanto para armazenamento de dados, faz com que a análise moderna, cada vez mais, tenha que lidar com enormes quantidades desses dados. Esse aumento não é encontrado apenas no número de amostras coletadas, por exemplo, ao longo do tempo, mas também no número de atributos, ou

características, que são medidos simultaneamente em um processo. Em muitas situações, os dados são reunidos em vetores cuja dimensão corresponde ao número de medições simultâneas sobre o processo. Quando a dimensão cresce, fala-se em dados de alta dimensão, pois cada amostra pode ser representada como um ponto ou vetor em um espaço de alta dimensão.

No processo de preparação dos dados, é importante se preocupar com a relação entre a quantidade de amostras (instâncias) e o número de características, pois estão diretamente relacionados ao desempenho do classificador. Embora o aumento do número de características possa levar a uma melhoria no desempenho, na prática, além de um certo ponto, adicionar novas características pode, na verdade, levar a uma redução no desempenho do sistema de classificação. Este fenômeno é denominado como "maldição da dimensionalidade" (do inglês, *curse of dimensionality*) Bishop *et al.* (1995), o que leva ao "fenômeno do pico" (do inglês, *peaking phenomenon*) Jain e Chandrasekaran (1982) no projeto de classificadores. Se o número de amostras de treinamento usadas para projetar o classificador for pequeno em relação ao número de características, o desempenho do classificador pode realmente degradar. Nesta pesquisa, há uma preocupação com a quantidade de instâncias, uma vez que há uma pequena quantidade de dados referentes aos dormentes de aço defeituosos, comparado à quantidade de dados referentes aos dormentes de aço saudáveis.

2.4.4.2. Razão Discriminante de Fisher

O discriminante linear de Fisher (FLD, do inglês, *Fisher Linear Discriminant*) é um abordagem eficiente para seleção de características e contribui, consequentemente, para a redução de dimensionalidade no reconhecimento estatístico de padrões (WEBB, 2003). O objetivo da análise de Fisher é realizar a redução de dimensionalidade preservando ao máximo a informação discriminatória da classe (SHARMA *et al.*, 2016). A abordagem adotada por Fisher foi encontrar uma combinação linear das variáveis que separa as duas classes. Esse critério proposto ficou conhecido como razão discriminante de Fisher.

A FDR pode ser utilizada para quantificar a capacidade de distinção de um conjunto de dados em relação a outro. Assim, considerando-se o caso de apenas duas classes, para efeito de simplificação da explicação, pode-se determinar a FDR a partir da Equação (2.8):

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{2.8}$$

A Equação (2.8) mostra que, idealmente, as médias das duas classes μ_1 e μ_2 devem estar distantes entre si, maximizando o numerador, enquanto cada classe deve possuir a menor variabilidade possível σ_1 e σ_2, minimizando o denominador. Portanto, quanto maior for o valor da FDR para um determinado parâmetro, maior será a contribuição deste parâmetro para a discriminabilidade do problema de classificação. Expandindo agora para M classes, tem-se a Equação (2.9) (THEODORIDIS e KOUTROUMBAS, 2010).

$$FDR = \sum_{i}^{M} \sum_{j \neq i}^{M} \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{2.9}$$

em que i e j representam os parâmetros. Dessa forma é possível quantificar a contribuição de cada um dos parâmetros para a discriminação das classes, sendo que aqueles que apresentarem maior capacidade terão os valores da FDR mais altos.

A seguir, apresenta-se um exemplo de aplicação da FDR para três parâmetros analisados: energia, variância e taxa de cruzamento por zeros (ZCR, do inglês, *Zero Crossing Rate*). A Tabela 2.3 mostra um conjunto de 20 amostras (metade associado aos dados saudáveis e a outra metade aos defeituosos) dos três parâmetros analisados. Para cada parâmetro, calcula-se o valor da FDR a partir da Equação 2.8.

Tabela 2.3: Conjunto de dados para exemplo de análise da FDR.

Defeituosos			**Saudáveis**		
Energia	**Variância**	**ZCR**	**Energia**	**Variância**	**ZCR**
450,42	3,5442	0,0781	1,4400	0,0112	0,0078
467,07	3,6716	0,0781	2,0800	0,0161	0,0078
479,54	3,7658	0,0781	2,3300	0,0179	0,0078
486,74	3,8191	0,0781	2,3300	0,0179	0,0156
488,98	3,8352	0,0781	2,4200	0,0187	0,0234
488,59	3,8327	0,0781	2,6700	0,0208	0,0234
488,87	3,8374	0,0859	3,3100	0,0260	0,0234
494,59	3,8865	0,0781	4,1200	0,0324	0,0234
508,94	4,0036	0,0781	5,1200	0,0402	0,0234
531,17	4,1815	0,0781	6,1200	0,0479	0,0234

Fonte: O autor.

Energia

- $\mu_{saudáveis} = 492,7211$
- $\mu_{defeituosos} = 3,3888$
- $\sigma^2_{saudáveis} = 332,2812$
- $\sigma^2_{defeituosos} = 2,0509$

$$FDR = \frac{(492,7211 - 3,3888)^2}{332,2812 + 2,0509} = 716,1924 \tag{2.10}$$

Variância

- $\mu_{saudáveis} = 3,8703$

- $\mu_{defeituosos} = 0,0264$
- $\sigma^2_{saudáveis} = 0,0214$
- $\sigma^2_{defeituosos} = 0,0001$

$$FDR = \frac{(3,8703 - 0,0264)^2}{0,0214 + 0,0001} = 687,2357 \tag{2.11}$$

ZCR

- $\mu_{saudáveis} = 0,0789$
- $\mu_{defeituosos} = 0,0190$
- $\sigma^2_{saudáveis} = 0,000006$
- $\sigma^2_{defeituosos} = 0,00004$

$$FDR = \frac{(0,0789 - 0,0190)^2}{0,000006 + 0,00004} = 78,0002 \tag{2.12}$$

De acordo com os resultados, o maior valor da FDR encontrado foi para a energia, seguido da variância e por último o da taxa de cruzamento por zeros. Como foi mencionado anteriormente, quanto maior o valor da FDR, maior é a contribuição deste parâmetro para a discriminação das classes. Nesse sentido, nos casos em que existe uma grande quantidade de parâmetros, como é o caso desta pesquisa, é possível selecionar aqueles que possuem o maior peso, na quantidade necessária para cada aplicação.

Para essa pesquisa, especificamente, foram variadas empiricamente a quantidade de parâmetros FDR selecionados para o treinamento do modelo na ordem de 2 a 24 (2, 4, 6, ... , 24). Isso significa que na primeira iteração, apenas os 2 melhores parâmetros FDR são selecionados; na segunda iteração, são selecionados os 4 melhores parâmetros e assim sucessivamente.

2.4.4.3. Análise de Componentes Principais

A PCA é um dos métodos mais populares para extração de características e redução de dimensionalidade no reconhecimento de padrões (THEODORIDIS e KOUTROUMBAS, 2010). A PCA obtém uma transformação linear de um vetor de entrada de alta dimensão em um de baixa dimensão cujos componentes não são correlacionados. Como a PCA usa os recursos mais expressivos (autovetores com os maiores autovalores), ele aproxima efetivamente os dados por um subespaço linear usando o critério de erro quadrático médio (JAIN *et al.*, 2000).

Em outras palavras, a PCA consiste em realizar uma transformação ortogonal com o intuito de cobrir a maior variabilidade dos dados possível, com um número menor de dimensões.

Portanto, o número de componentes principais é sempre menor ou igual ao número de dimensões dos dados originais. A PCA é calculada de modo que a primeira componente do novo sistema de coordenadas tenha a maior variância de todas, e a segunda componente tenha a segunda maior variância, e assim sucessivamente (SMITH, 2002).

A partir do cômputo de uma matriz em que as linhas representam as observações e as colunas representam os parâmetros, o primeiro passo para o cálculo da PCA é subtrair as médias de cada atributo, a partir da Equação (2.13).

$$X_{novo} = X_i - \bar{x} \tag{2.13}$$

O segundo passo é calcular a matriz de covariâncias entre os parâmetros, a partir da Equação 2.14.

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y}) \tag{2.14}$$

Em seguida, calcula-se os autovalores λ e autovetores v associados a matriz de covariância A. Os autovalores podem ser calculados encontrando as raízes da Equação 2.15.

$$det(\mathbf{A} - \lambda\mathbf{I}) = 0 \tag{2.15}$$

Feito isso, calcula-se os autovetores associados a cada autovalor, de acordo com a Equação 2.16.

$$\mathbf{Av} = \lambda\mathbf{v} \tag{2.16}$$

Na PCA, os maiores coeficientes de autovalor correspondem aos autovetores com maior variabilidade de informação. Portanto, o novo conjunto de dados é construído a partir dos autovetores que se deseja utilizar, e que representam a maior parte da variabilidade dos dados originais. Ao considerar originalmente um conjunto com n dimensões de dados, deve-se calcular n autovetores e autovalores. Ao escolher apenas os primeiros p autovetores, o conjunto de dados final terá p dimensões.

Na sequência, pode-se utilizar a matriz de autovetores (Φ) para realizar uma transformação linear sobre o conjunto de dados original $X_{m_x n}$.

$$Y = \Phi^T X \tag{2.17}$$

Cada linha da matriz Y corresponde a uma componente principal. Pode-se demonstrar que as p primeiras componentes principais concentram grande parte da variância dos dados e, em tese, podem ser utilizadas para a reconstrução aproximada dos dados originais por meio da Equação 2.18.

$$X_{m_x n} = \Phi_{m_x p} Y_{p_x n} \tag{2.18}$$

É possível calcular uma estimativa do erro quadrático médio obtido quando se utilizam

apenas as p primeiras componentes principais durante a reconstrução.

$$MSE = \frac{\sum_{i=p+1}^{m} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \tag{2.19}$$

Além disso, para estabelecer um valor mínimo aceitável da fração de variância total representada pelos p primeiros componentes, utiliza-se a Equação 2.20.

$$Frac.\lambda_p = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \tag{2.20}$$

A seguir, apresenta-se um exemplo de aplicação da PCA para os mesmos três atributos analisados no exemplo da FDR. Os dados foram gerados aleatoriamente dentro de um intervalo especificado para cada atributo. Inicialmente, encontra-se para cada instância, a média subtraída, como apresentado na Tabela 2.4. Na sequência, computa-se a matriz de covariância.

Tabela 2.4: Conjunto de dados para exemplo de análise da PCA.

Energia	**Variância**	**ZCR**	$E_i - \bar{e}$	$V_i - \bar{v}$	$Z_i - \bar{z}$
12,6213	2,0588	1,7108	-1,4822	-0,0747	0,6809
12,8283	2,0584	0,1608	-1,2758	-0,0751	-0,8691
12,7584	2,3932	1,4263	-1,3451	0,2596	0,3964
14,6723	1,9966	1,2918	0,5687	-0,1369	0,2619
15,6455	2,6707	0,4508	1,5419	0,5371	-0,5791
14,0075	1,8424	1,4227	-0,0960	-0,2911	0,3928
13,0506	1,8794	1,7011	-1,0529	-0,2546	0,6712
14,9539	2,2117	0,4861	0,8503	0,0781	-0,5438
12,5046	2,7226	0,3443	-1,5989	0,5890	-0,6856
15,8476	2,3523	0,9962	1,7440	0,2187	-0,0337
13,6673	2,3867	1,0558	-0,4362	0,2531	0,0259
14,8744	2,1148	0,6369	0,7708	-0,0187	-0,393
14,2674	2,7611	1,0507	0,1638	0,6275	0,0208
14,7928	1,7182	0,4078	0,6892	-0,4153	-0,6221
15,6108	2,6307	0,1313	1,5072	0,4971	-0,8986
13,256	1,3604	1,8788	-0,8475	-0,7731	0,8489
13,8177	1,6731	0,677	-0,2858	-0,4604	-0,3529
15,108	2,1865	1,1948	1,0042	0,0529	0,1649
12,6202	2,1495	1,8623	-1,4833	0,0159	0,8324
15,167	1,5041	1,7116	1,0634	-0,6294	0,6817

Fonte: O autor.

$$cov(E,E) = \frac{1}{n-1}\sum_{i=1}^{n}(E_i - \bar{e})(E_i - \bar{e}) = \frac{24,4716}{19} = 1,2879$$

$$cov(V,V) = \frac{1}{n-1}\sum_{i=1}^{n}(V_i - \bar{v})(V_i - \bar{v}) = \frac{3,0232}{19} = 0,1591$$

$$cov(Z,Z) = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{z})(Z_i - \bar{z}) = \frac{6,5317}{19} = 0,3437$$

$$cov(E,V) = \frac{1}{n-1}\sum_{i=1}^{n}(E_i - \bar{e})(V_i - \bar{v}) = \frac{0,9974}{19} = 0,0524$$

$$cov(E,Z) = \frac{1}{n-1}\sum_{i=1}^{n}(E_i - \bar{e})(Z_i - \bar{z}) = \frac{-4,4043}{19} = -0,2318$$

$$cov(V,Z) = \frac{1}{n-1}\sum_{i=1}^{n}(V_i - \bar{v})(Z_i - \bar{z}) = \frac{-2,0305}{19} = -0,1068$$

A matriz de covariância resultante é dada por:

$$cov = \begin{pmatrix} 1,2879 & 0,0524 & -0,2318 \\ 0,0524 & 0,1591 & -0,1068 \\ -0,2318 & -0,1068 & 0,3437 \end{pmatrix}$$

Uma vez que a matriz de covariância é quadrada, é possível calcular os autovetores (λ) e autovalores (Φ) para esta matriz, respectivamente pelas equações 2.15 e 2.16. Estes são bastante importantes, pois fornecem informações úteis sobre os dados.

$$\lambda = \begin{pmatrix} 1,3466 \\ 0,3364 \\ 0,1078 \end{pmatrix}$$

$$\Phi = \begin{pmatrix} 0,9708 & 0,2347 & 0,049 \\ 0,0637 & -0,4498 & 0,8908 \\ -0,2311 & 0,8616 & 0,4517 \end{pmatrix}$$

Em geral, uma vez que os autovetores são encontrados a partir da matriz de covariância, o próximo passo é ordená-los por autovalor, do maior para o menor. Isso fornece os componentes em ordem de importância. Dado o conjunto de dados do exemplo e o fato de existir somente 3 autovetores, tem-se 3 opções neste caso. Pode-se formar um vetor de características com os 3 autovetores ou optar por deixar de fora uma ou as duas componentes menos significativas. Por fim, as componentes principais podem ser encontradas por meio da transformação linear apresentada pela Equação 2.17.

No caso do exemplo, ao utilizar somente uma componente principal tem-se a fração de variância alcançando 75,19%. Essa verificação pode ser feita por meio da Equação 2.20.

$$Frac.\lambda_1 = \frac{\sum_{i=1}^{p}\lambda_i}{\sum_{i=1}^{m}\lambda_i} = \frac{1,3466}{1,7908} = 75,19\%$$

Para essa pesquisa, especificamente, os vetores de características (atributos) foram formados seguindo a ordem decrescente de importância das componentes principais, de tal modo a apresentarem tamanhos variando, assim como na FDR, de 2 a 24 (2, 4, 6, ... , 24).

2.5. Modelagem

De acordo com Theodoridis e Koutroumbas (2010), os classificadores são modelos matemáticos que incorporam um conjunto de regras capazes de estabelecer um mapeamento entre um dado padrão de entrada e uma determinada classe. O processo de obtenção de um classificador consiste em três estágios: treinamento, validação e inferência. O primeiro emprega uma base de dados conhecida a priori, que pode ou não ser rotulada, e realiza o ajuste dos parâmetros intrínsecos do modelo a partir das classes desejadas na saída quando os rótulos se encontram disponíveis (supervisionado), ou a partir de outros métodos que não necessitam destas informações (não-supervisionado). Na etapa de validação pode-se avaliar a taxa de acerto que o classificador deve apresentar durante a sua operação, indicando a precisão esperada durante a separação de classes. A etapa de inferência consiste em gerar a aplicação com base nas melhores configurações definidas nas etapas anteriores, sendo possível aplicar o sistema para quaisquer novos dados desconhecidos, diferentes daqueles existentes na base de dados utilizadas nas etapas anteriores.

O treinamento supervisionado consiste na utilização de dados de entrada rotulados, associados às possíveis classes existentes em um problema, para o ajuste dos parâmetros intrínsecos do modelo, ou seja, cada vetor de parâmetros está associado a uma classe previamente conhecida. Neste caso, os valores de erro observados na saída do modelo se encontram disponíveis e, portanto, podem ser utilizados como medidas objetivas a serem minimizadas durante o treinamento, como se tem por exemplo o erro quadrático médio. A extração dos dados rotulados deve ocorrer de forma que se obtenha um número aproximado de dados correspondentes a cada classe do problema. Neste ponto, deve-se destacar que, neste trabalho, a quantidade de dados saudáveis supera significativamente a quantidade de dados associados aos trechos em que existem dormentes de aço danificados. Por essa razão, realiza-se uma seleção aleatória dos dados saudáveis, tentando manter iguais proporções entre os diferentes elementos presentes na base, de modo que ao final do processo se obtenha iguais quantidades de dados provenientes de trechos com dormentes saudáveis e danificados.

Por outro lado, no treinamento não-supervisionado, o ajuste dos parâmetros intrínsecos do modelo é realizado de modo a satisfazer algum critério objetivo, como por exemplo a EM, sem conhecer a priori a correspondência entre os dados de entrada do problema e as possíveis classes existentes e, consequentemente, sem qualquer medida de erro cometido na saída. Assim, neste trabalho foram considerados cinco classificadores, sendo 4 treinados a partir de dados rotulados (RNA, SVM, GMM e ADB) e apenas um outro treinado sem a utilização de dados rotulados HMM. Em relação à complexidade de cada classificador, não houve uma análise sistemática para encontrar a melhor configuração no ajuste dos parâmetros. Nesse sentido, para todos os classificadores a complexidade foi empiricamente variada.

2.5.1. Redes Neurais Artificiais

Nos últimos tempos, as Redes Neurais Artificiais (RNA) tornaram-se um modelo popular e útil para classificação, reconhecimento de padrões, previsão e clusterização. As RNA são um tipo de modelo para aprendizado de máquina e se tornaram relativamente competitivas em relação aos modelos convencionais de regressão e estatística.

As RNA surgiram com base na ideia de simular o cérebro humano em seu funcionamento e em suas potencialidades. Nesse sentido, define-se as RNA como sistemas paralelos e distribuídos compostos por unidades de processamento simples (neurônios) que calculam determinadas funções matemáticas (normalmente não lineares). Tais unidades são dispostas em uma ou mais camadas e interligadas por um grande número de conexões, geralmente unidirecionais (MORAIS, 2010). Na maioria dos modelos, estas conexões estão associadas a pesos, os quais armazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida por cada neurônio da rede. Em um paralelo com o cérebro humano, os pesos na versão artificial são análogos às forças das sinapses que ligam os neurônios anteriores aos dendritos deste neurônio.

2.5.1.1. Algoritmos de Aprendizado

Os algoritmos de aprendizado podem ser considerados como um conjunto de procedimentos definidos capazes de adaptar os parâmetros de uma rede neural para que a mesma possa aprender uma determinada função. Em outras palavras, os parâmetros livres de uma rede neural são adaptados através de um processo de estimulação pelo ambiente no qual ela está inserida (HAYKIN, 2007). Existem vários algoritmos de aprendizado e, basicamente, eles se diferem pela maneira na qual o ajuste dos pesos é realizado. Diversos métodos para treinamento de rede foram desenvolvidos mas, de forma geral, o processo de aprendizado implica na seguinte sequência de eventos:

1. A rede neural é estimulada pelo ambiente;
2. Ela sofre modificações em seus parâmetros livres como resultado desta estimulação;
3. A rede neural responde de uma nova maneira ao ambiente, devido às modificações ocorridas na sua estrutura interna.

2.5.1.2. Redes *Perceptron* Multicamadas

As RNA possuem diversas arquiteturas, dentre as quais se pode citar a rede *perceptron* multicamadas (MLP, do inglês, *Multilayer Perceptron*). Elas são compostas por uma camada de entrada (*input layer*) para receber o sinal, uma camada de saída (*output layer*) que toma uma decisão ou previsão sobre a entrada, e entre esses dois, um número arbitrário de camadas ocultas (*hidden layers*) que está relacionado com a capacidade de aprendizagem.

Os neurônios na camada de entrada recebem os dados e os transferem para os neurônios na primeira camada oculta por meio das conexões ponderadas. Aqui, os dados são processados matematicamente e o resultado é transferido para os neurônios na próxima camada. Dessa forma, quanto mais próximo à camada de saída, mais complexas se tornam as funções implementadas. Por fim, os neurônios na última camada fornecem a saída da rede.

O j-ésimo neurônio em uma camada oculta processa os dados de entrada (x_i): (i) calculando a soma ponderada e adicionando um termo *bias* (θ_j) de acordo com a Equação 2.21 (AMATO *et al.*, 2013):

$$U_j = \sum_{i=1}^{m} x_i w_{ij} + \theta_j (j = 1, 2, ..., n) \tag{2.21}$$

(ii) transformando o potencial de ativação U_j por meio de uma função de ativação matemática adequada e (iii) transferindo o resultado para os neurônios da próxima camada. Várias funções de ativação estão disponíveis (GOODFELLOW *et al.*, 2016).

Para a realização do treinamento de uma rede MLP, vários algoritmos estão disponíveis e são em sua maioria do tipo supervisionado. O algoritmo de aprendizado mais conhecido para treinamento das redes MLP é o algoritmo *backpropagation*. Geralmente, os métodos de aprendizado aplicados para RNA do tipo MLP utilizam variações deste algoritmo. Apresenta-se na Tabela 2.5, as funções disponíveis no *Matlab* das possíveis variações citadas.

Tabela 2.5: Algoritmos de aprendizado e sua respectiva função no *Matlab*.

Algoritmo	**Descrição**
trainlm	*Backpropagation* Levenberg-Marquadt
traingd	*Backpropagation* de gradiente decrescente
traingdm	*Backpropagation* de gradiente decrescente com *momentum*
traingda	*Backpropagation* de gradiente decrescente com taxa adaptativa
traingdx	*Backpropagation* de gradiente decrescente com *momentum* e taxa adaptativa

Fonte: O autor.

De forma geral, o *backpropagation* é um algoritmo supervisionado que utiliza pares (entrada e saída desejada) para, por meio de um mecanismo de correção de erros, ajustar os pesos da rede. O treinamento ocorre em duas fases, conhecidas como fase *forward* e fase *backward*. A primeira delas é utilizada para definir a saída da rede para um dado padrão de entrada, enquanto a segunda utiliza a saída desejada e a saída fornecida pela rede para atualizar os pesos e limiares (*biases*) de suas conexões.

Para o sistema MLP desenvolvido, foi utilizado uma única camada oculta, mas variando o número de neurônios em cada iteração (20, 25, 30, 35 e 40) e 1 neurônio na camada de saída. Na primeira parte do treinamento foi utilizado o algoritmo *Backpropagation* de gradiente decrescente com *momentum* e taxa adaptativa (definido no *Matlab* pela função *traingdx*). Foi utilizada a tangente hiperbólica (definido no *Matlab* pela função *tansig*) como função de

ativação para a camada oculta e sigmóide (definido no *Matlab* pela função *logsig*) para a camada de saída. O treinamento é interrompido quando se alcança o número máximo de épocas (1000) ou o erro desejado ($1e^{-4}$) ou o gradiente mínimo de desempenho ($1e^{-5}$). A taxa de aprendizagem escolhida foi de 0,01. Após algumas iterações e redução de erros, o processo continua com uma função de treinamento de rede que atualiza os valores de peso e *bias* de acordo com a otimização de Levenberg-Marquardt (definido no *Matlab* pela função *trainlm*). Para esta segunda etapa, a única alteração na condição de parada foi no número máximo de épocas, alterado para 100. Apresenta-se na Figura 2.25, um exemplo de configuração utilizada em uma das iterações na geração dos modelos.

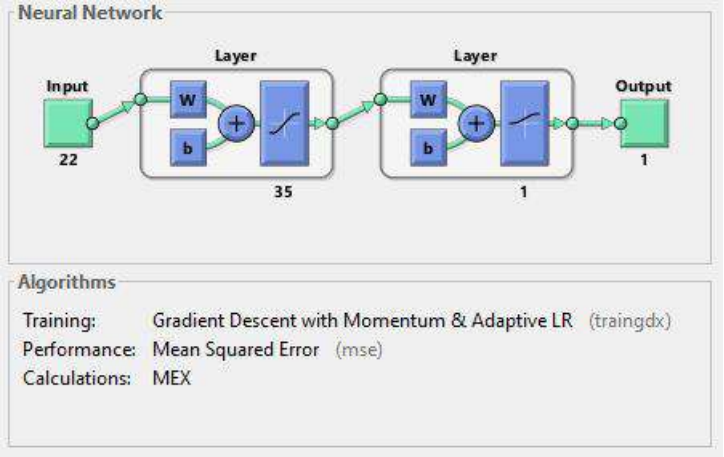


Figura 2.25: Exemplo de configuração da RNA.
Fonte: O autor.

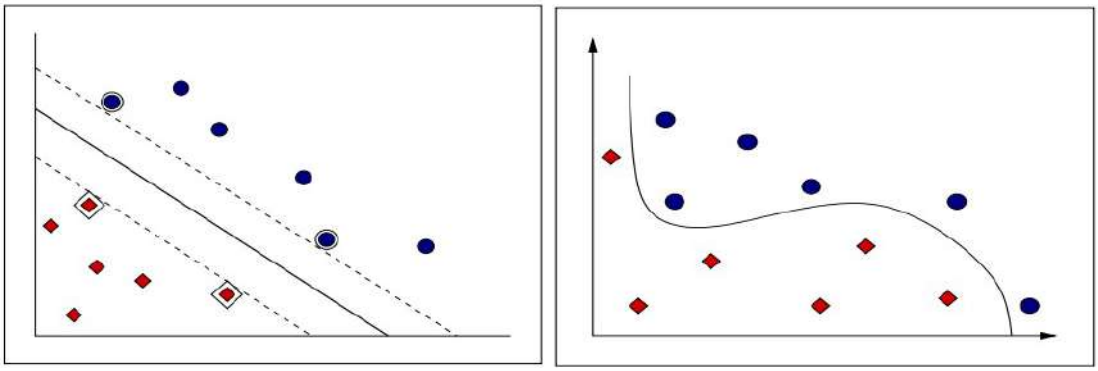
2.5.2. Máquinas de Vetores de Suporte

Segundo Haykin (2007), as máquinas de vetores de suporte (SVM, do inglês, *Support Vectors Machines*) podem ser consideradas como algoritmos de aprendizado supervisionado, baseado no princípio de minimização de risco estrutural, advindo das teorias de aprendizado estatístico. As SVM possuem a capacidade de interpretar o modelo matemático dos dados por meio dos vetores de suporte, apresentando um bom desempenho de generalização em problemas de classificação, apesar do fato de não incorporar conhecimento do domínio do problema.

O modelo mais simples do algoritmo, que também foi o primeiro a ser introduzido, é o chamado classificador de margem máxima. Este algoritmo supervisionado para classificação, proposto por Boser *et al.* (1992), desde então evoluiu para o que hoje é conhecido como as SVM. Este primeiro modelo foi utilizado apenas para os casos em que os dados eram linearmente separáveis, ficando restrito, portanto, à poucas aplicações práticas. Essas SVM vieram a ser definidas como sendo de margens rígidas (ou *hard margins*). Apesar dessa limitação para os casos práticos, o seu desenvolvimento foi importante para novos estudos e formulação de SVM mais sofisticadas, capazes de tratar, agora, os dados não linearmente separáveis, dando origem às SVM de margens flexíveis (ou *soft margins*). Apresenta-se nas figuras 2.26a e 2.26b,

uma ilustração de um conjunto de treinamento bidimensional linearmente separável e não linearmente separável, respectivamente. A linha contínua que separa os dados em duas classes distintas é chamada de superfície de decisão. No caso particular da Figura 2.26a, a linha é conhecida como hiperplano de separação, devido a linearidade da superfície de decisão.

O classificador de margem máxima citado anteriormente, otimiza limites no erro de generalização das máquinas lineares em termos da margem de separação entre as classes a qual é determinada pelo hiperplano de separação. Para isso, a estratégia é utilizar um hiperplano de separação ótima (ou de margem máxima) na separação dos dados. Segundo Vapnik (1999), um hiperplano é considerado de margem máxima se separa um conjunto de vetores sem erro e a distância entre os vetores (das classes opostas) mais próximos ao hiperplano é máxima.



(a) Linearmente separável. (b) Linearmente inseparável.

Figura 2.26: Espaço de características.
Fonte: O autor.

Para exemplificar o conceito de hiperplano ótimo, considera-se um conjunto linearmente separável de uma classificação binária. Uma classificação linear consiste em determinar uma função $f : X \subseteq \mathbb{R}^N \rightarrow \mathbb{R}^N$, que atribui um rótulo $(+1)$ se $f(x) > 0$ e (-1) caso contrário. Uma função linear pode ser representada pela Equação (2.22).

$$f(x) = \langle w.x \rangle + b = \sum_{i=1}^{n} w_i x_i + b \tag{2.22}$$

onde w denota o vetor de pesos, b o vetor de limiar ou *bias*. Os valores de w e b são obtidos pelo processo de aprendizagem a partir dos dados de entrada. Eles são responsáveis por controlar a função e a regra de decisão. O vetor peso define uma direção perpendicular ao hiperplano e com a variação do *bias* o hiperplano é movido paralelamente a ele mesmo. Dessa forma, pode-se afirmar que para a minimização do erro de generalização, deve-se maximizar a margem ρ, ilustrada na Figura 2.27, isto é, a distância mínima entre um hiperplano de separação das duas classes e os dados de entrada de cada classe que estejam mais próximos a esse hiperplano (HAYKIN, 2007).

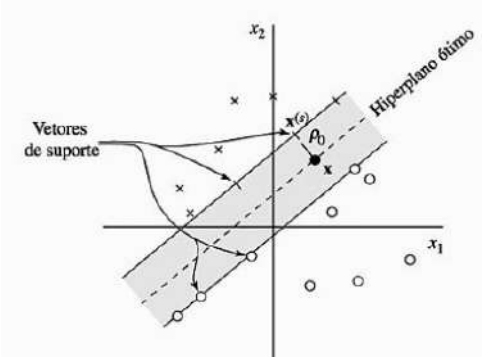


Figura 2.27: Representação de hiperplanos de separação de classes.
Fonte: Retirado de Haykin (2007).

Os vetores de suportes, indicados na Figura 2.27, desempenham um papel fundamental na operação desta classe de máquinas de aprendizagem. Em termos conceituais, devido à localização mais próxima da superfície de decisão, os vetores de suporte são os mais difíceis de classificar e, portanto, tem uma influência direta na localização ótima da superfície de decisão.

Para estender a SVM linear à resolução de problemas não lineares foram introduzidas funções que mapeiam o conjunto de treinamento em um espaço linearmente separável, o espaço de características de alta dimensionalidade. Para isso, a primeira operação é realizada de acordo com o teorema de Cover, descrito por Haykin (2007) da seguinte forma: "Um problema complexo de classificação de padrões tem mais probabilidade de ser separável linearmente em um espaço de alta dimensão do que em um espaço de baixa dimensão". Dessa forma, a SVM não linear realiza uma mudança de dimensionalidade, por meio das funções *Kernel*, caindo então em um problema de classificação linear, podendo fazer uso do hiperplano ótimo. Entretanto, o teorema de Cover não discute se o hiperplano de separação é ótimo, necessitando, portanto, de uma segunda operação. Especificamente, a segunda operação explora a ideia de construir um hiperplano ótimo, assim como no exemplo apresentado no início desta seção, mas com uma diferença fundamental; agora, ele é definido como uma função linear de vetores retirados do espaço de característica, em vez do espaço de entrada original. Apresenta-se na Figura 2.28 o processo de transformação de um domínio não linearmente separável, em um problema linearmente separável por meio do aumento da dimensão, em que é feito um mapeamento do espaço de entrada X em um novo espaço $Z = \phi(x)|x \in X$, em que ϕ_i são as funções *kernel*.

O treinamento realizado neste trabalho para a construção do modelo foi baseado em duas variações quanto a função *Kernel* utilizada. Foram considerados os *kernels* linear e polinomial. Apesar de serem apresentados no capítulo 2, o *kernel* sigmóide e o *Radial Basis Function* (RBM) não foram utilizados nos experimentos por dois motivos. O primeiro deles, se deve pelo fato de que no *kernel* RBF $\kappa(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$, segundo Mattera e Haykin (1999), existe uma dificuldade na definição do parâmetro σ, sendo necessário a utilização de alguma estratégia mais específica para determinar o valor mais indicado para o problema. A função disponível no *Matlab* (*fitcsvm*), apesar de utilizar um procedimento heurístico para estimar o valor de σ, após alguns testes realizados, não retornou um bom desempenho. Entretanto, essa

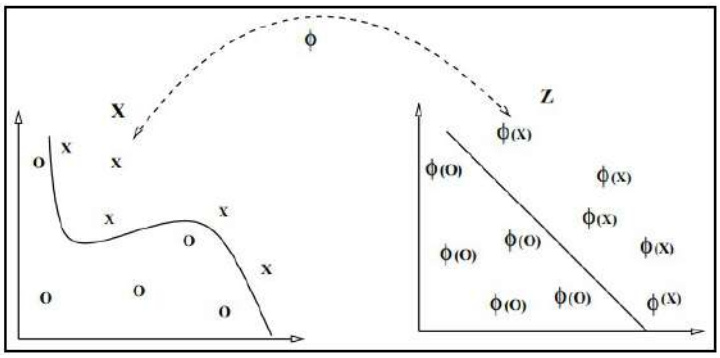


Figura 2.28: Mapeamento do espaço de entrada via função *Kernel*.
Fonte: O autor.

investigação ultrapassa o escopo desta dissertação. O segundo motivo é que, de acordo com Vapnik (1999), o *kernel* sigmóide $\kappa(x_i, x_j) = tanh(\beta_0 \langle x_i x_j \rangle) + \beta_1$ não satisfaz as condições de *Mercer* (Capítulo 2, Seção 2.5.4) para todos os valores de β_0 e β_1. Diante disso, também seria necessário realizar experimentos para definir que valores dessas variáveis (satisfazendo as condições de *Mercer*) seriam mais adequadas para o problema em foco, ultrapassando, também, o escopo desta dissertação. Para a realização dos testes, utilizou-se a função *fitcsvm* do *Matlab*, configurada para receber as três variações quanto aos *kernels* utilizados. O *kernel* polinomial é definido como $K(X_i, X_j) = (\gamma X_i \cdot X_j + C)^p, \gamma > 0$. Foi utilizado $p = [2,3]$, $C = 1$ e a função do *Matlab* seleciona um valor apropriado para γ usando um procedimento heurístico. Esse procedimento heurístico usa subamostragem, de modo que as estimativas podem variar de uma chamada para outra. Portanto, para reproduzir os resultados, define-se uma semente de número aleatório antes do treinamento.

2.5.3. Modelo de Mistura de Gaussianas

O modelo de mistura de gaussianas (GMM, do inglês, *Gaussian Mixture Models*) consiste de combinações ponderadas de funções densidade de probabilidade (PDF, do inglês, *Probability Density Functions*) gaussianas (PORTELA, 2015), cujas médias, matrizes de covariância e pesos são determinados durante o treinamento utilizando o algoritmo de maximização da esperança (EM, do inglês, *Expectation Maximization*).

Neste modelo, a mistura de gaussianas é ajustada de modo a representar os dados que apresentem maiores similaridades por meio das funções gaussianas, sendo o objetivo maximizar o critério de agrupamento em função da verossimilhança. Assim, as classes são descritas estatisticamente pelos parâmetros média e matriz de covariância que são inerentes à mistura e representam os dados que serão classificados com base na estatística.

O modelo de mistura gaussiana multidimensional é definido pelas componentes peso ϕ_i, as médias μ_i e matrizes de covariância Σ_i, representados pelas equações 2.23, 2.24 e 2.25. Para maiores detalhes, ver Bilmes *et al.* (1998).

$$p(\mathbf{x}) = \sum_{i=1}^{K} \phi_i N(\mathbf{x}|\mu_i, \Sigma_i) \tag{2.23}$$

$$N(\mathbf{x}|\mu_i, \Sigma_i) = \frac{1}{2\pi^{p/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)} \tag{2.24}$$

$$\sum_{i=1}^{K} \phi_i = 1 \tag{2.25}$$

O modelo baseado em GMM utilizado neste trabalho foi treinado a partir do algoritmo de EM e, ao longo dos experimentos, o número de componentes gaussianas da mistura foi variado entre 2 e 10 (2, 4, 6, 8 e 10). Além disso, foram obtidos dois modelos, sendo um para a representação dos dados associados aos dormentes danificados e outro para a representação dos dados associados aos trechos com dormentes saudáveis. Desta forma, embora o algoritmo clássico de maximização de esperança seja empregado para o treinamento não-supervisionado, neste trabalho utilizaram-se dados rotulados provenientes de trechos da EFVM como dormentes danificados e dados rotulados provenientes de trechos contendo apenas dormentes saudáveis, para a obtenção de modelos GMM independentes, cada qual para representar uma classe específica.

2.5.4. Métodos de *ensemble* e o *Adaboost*

Os métodos de *ensemble* para a classificação é parte ativa da pesquisa de reconhecimento de padrões, por geralmente apresentar uma melhor performance do que as técnicas individuais. O *ensemble* nada mais é do que um conjunto de vários classificadores treinados individualmente (classificadores de base), cujas decisões são de alguma forma combinada. Embora haja um número quase ilimitado de maneiras pelas quais isso pode ser alcançado, talvez existam três classes de técnicas de aprendizagem em conjunto que são mais comumente usadas na prática: *bagging* (BREIMAN, 1996), *stacking* (WOLPERT, 1992) e *boosting* (SCHAPIRE, 1990).

Bootstrap aggregation, ou apenas *bagging*, é um método de aprendizado de conjunto que busca um grupo diversificado de membros do conjunto variando os dados de treinamento. As previsões feitas pelos membros do conjunto são então combinadas usando estatísticas simples, como votação ou média. A chave para o método é a maneira pela qual cada amostra do conjunto de dados é preparada para treinar membros do *ensemble*. Cada modelo obtém sua própria amostra exclusiva do conjunto de dados. Apresenta-se na Figura 2.29 um exemplo da estrutura de *bagging*.

Stacking é um método de conjunto que busca um grupo diversificado de membros variando os tipos de modelo ajustados aos dados de treinamento e usando um modelo para combinar previsões. Os membros do conjunto são chamados de modelos de nível 0 e o modelo usado para combinar as previsões é chamado de modelo de nível 1. A hierarquia dos modelos de dois níveis

é a abordagem mais comum, embora mais camadas de modelos possam ser usadas. Por exemplo, em vez de um único modelo de nível 1, podemos ter 3 ou 5 modelos de nível 1 e um único modelo de nível 2 que combina as previsões dos modelos de nível 1 para fazer uma previsão. Apresenta-se na Figura 2.30 um exemplo da estrutura de *stacking*.

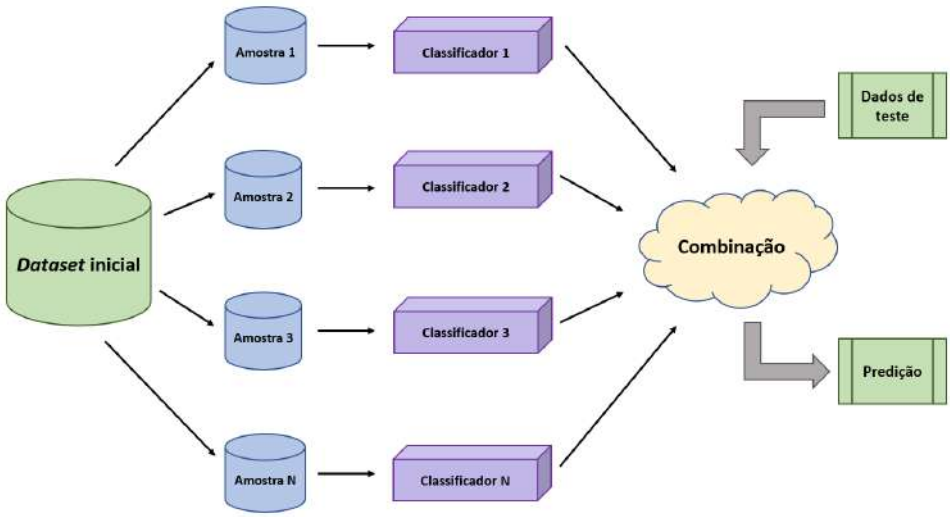


Figura 2.29: Bagging Ensemble.
Fonte: O Autor.

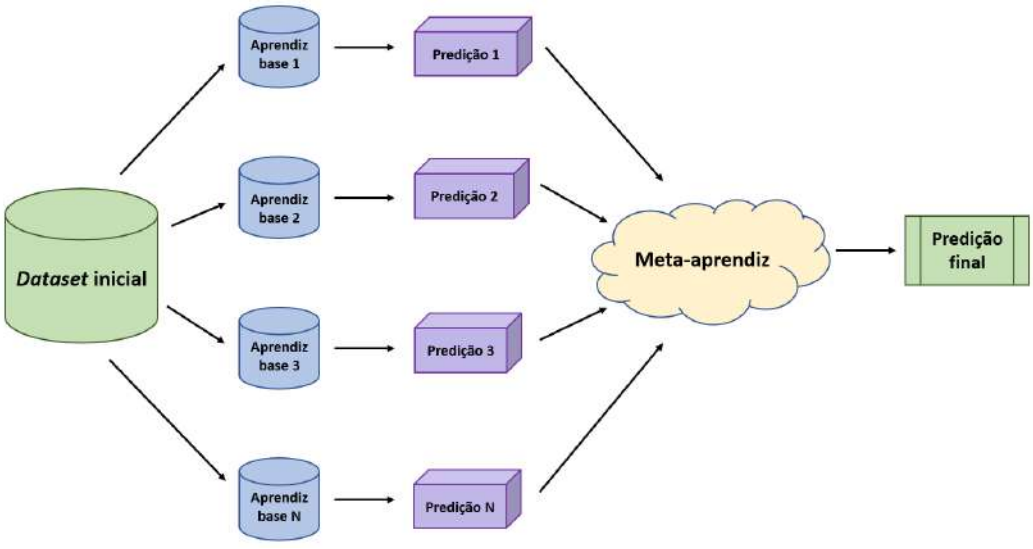


Figura 2.30: Stacking Ensemble.
Fonte: O Autor.

Boosting é um método de conjunto que procura alterar os dados de treinamento para focar a atenção em exemplos que modelos de ajuste anteriores no conjunto dos dados de treinamento erraram. Em outras palavras, o conjunto de dados de treinamento para cada classificador subsequente se concentra cada vez mais em instâncias classificadas incorretamente por classificadores gerados anteriormente. Os modelos são ajustados e adicionados ao conjunto sequencialmente de tal forma que o segundo modelo tenta corrigir as previsões do primeiro modelo,

o terceiro corrige o segundo modelo e assim por diante. Em geral, as técnicas de *ensemble* envolvem o uso de classificadores "fracos", a fim de combinar as suas previsões por meio de votação simples ou média, embora as contribuições sejam ponderadas proporcionalmente ao seu desempenho ou capacidade. O objetivo é gerar um classificador "forte" a partir de muitos classificadores "fracos". Normalmente, o conjunto de dados de treinamento é deixado inalterado e, em vez disso, o algoritmo de aprendizado é modificado para prestar mais ou menos atenção a exemplos específicos (linhas de dados) com base no fato de terem sido previstos corretamente ou incorretamente por membros do conjunto adicionados anteriormente. Por exemplo, as linhas de dados podem ser pesadas para indicar a quantidade de foco que um algoritmo de aprendizado deve dar ao aprender o modelo. Apresenta-se na Figura 2.31 um exemplo da estrutura de *boosting*.

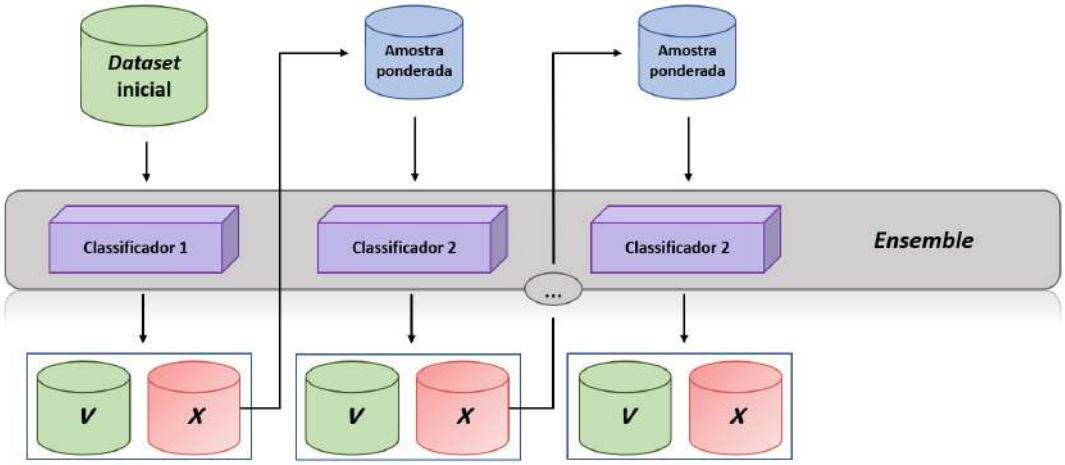


Figura 2.31: Boosting Ensemble.
Fonte: O Autor.

O algoritmo *AdaBoost*, introduzido por Freund e Schapire (1997), resolveu muitas das dificuldades práticas dos algoritmos de *boosting* anteriores. Desde o *AdaBoost*, muitos métodos de reforço foram desenvolvidos e alguns, como o reforço de gradiente estocástico, podem estar entre as técnicas mais eficazes para classificação e regressão em dados tabulares (estruturados). O *AdaBoost* chama um classificador fraco repetidamente em iterações $t = 1, 2, ..., T$. Para cada chamada a distribuição de pesos D_t é atualizada para indicar a importância do exemplo no conjunto de dados usado para classificação. A cada iteração os pesos de cada exemplo classificado incorretamente são aumentados (ou alternativamente, os pesos classificados corretamente são decrementados), para que então o novo classificador trabalhe em mais exemplos.

Neste trabalho, para a implementação do *Adaboost*, utilizou-se classificadores baseado em árvores de decisão como aprendiz fraco para fins de treinamento. Dessa forma, variou-se a complexidade do classificador (número de ciclos de aprendizagem do *ensemble* a serem realizados) de 20 a 100 (20, 40, 60, 80 e 100), definidos de forma empírica. Para a realização dos testes, utilizou-se a função *fitensemble* do *Matlab*, configurada com as características mencionadas.

2.5.5. Modelos Ocultos de Markov

Uma variável aleatória X pode ser definida como uma função que associa a cada possível observação ou saída ξ, um determinado valor $X(\xi)$. Assim, um processo estocástico é definido como uma sequência de variáveis aleatórias, cada qual associada a um determinado instante de tempo t, ou seja, uma função $X(\xi,t)$ (PAPOULIS, 1991). Quando o instante de tempo t pode assumir apenas valores inteiros, o processo estocástico é dito de tempo discreto, enquanto no caso em que assume quaisquer valores reais é considerado um processo de tempo contínuo.

Um processo estocástico pode apresentar variações de comportamento ao longo do tempo ou mantê-lo inalterado. Assim, um processo estocástico é definido como estacionário no sentido estrito quando as propriedades estatísticas não variam ao longo do tempo, ou de forma mais geral, é definido como estacionário no sentido amplo quando a média é constante e a função de autocorrelação não varia ao longo do tempo.

Adicionalmente, um processo markoviano é definido como um processo estocástico cuja saída ou estado futuro (no próximo instante de tempo) depende apenas da saída no presente (propriedade de Markov), independentemente do passado. Em particular, quando o número de estados do processo markoviano é finito, tem-se uma Cadeia de Markov, que é caracterizada pelos estados e pela matriz de transição de estados (P_{ij}) que contém as probabilidades associadas a permanência ou mudança de estado, conforme indicado na Figura 2.32.

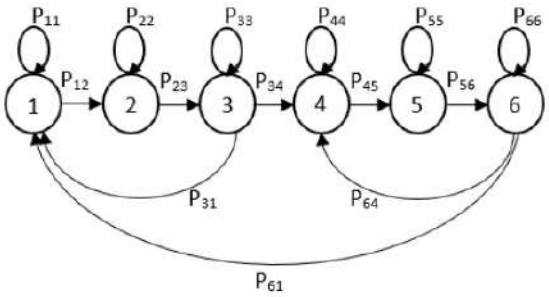


Figura 2.32: Modelo de Markov com 6 estados.
Fonte: O autor.

Uma cadeia de Markov pode ser utilizada para modelar o comportamento de um processo físico e a partir desta representação é possível realizar a estimativa da probabilidade de ocorrência de uma sequência específica de estados. Desta forma, a sequência de estados para a qual se deseja obter a probabilidade de ocorrência é conhecida à priori.

Por outro lado, existem problemas em que a sequência de estados não é conhecida a princípio, mas apenas os parâmetros que descrevem o comportamento do processo físico se encontram disponíveis. Neste caso, assumindo-se que o problema satisfaça a propriedade de Markov, a abordagem apropriada consiste na utilização dos modelos ocultos de Markov (HMM, do inglês, *Hidden Markov Models*).

O HMM é caracterizado por um conjunto de N estados, por uma matriz de transição de estados A dada pela Equação 2.26:

$$A = \begin{bmatrix} a_{11} & \dots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \dots & a_{NN} \end{bmatrix} \tag{2.26}$$

e pela sequência de T observações (parâmetros que descrevem o comportamento do processo) conforme indicado na Equação (2.27):

$$O = [o_1, o_2, ..., o_T] \tag{2.27}$$

Além disso, o HMM também é definido pela probabilidade de emissão de estado, ou sequência de verossimilhança das observações dada pela Equação (2.28):

$$B = b_i(o_t), \tag{2.28}$$

em que o_t são as observações e b_i é uma função densidade de probabilidade gaussiana do estado i ou uma combinação delas. Por fim, para a completa definição do HMM é necessária a definição da distribuição de probabilidade inicial dos estados, ou seja, a probabilidade π do estado i deve corresponder ao início do processo, de acordo com a Equação (2.29):

$$\pi = [\pi_1, \pi_1, ..., \pi_N,] \tag{2.29}$$

lembrando-se que $\sum_{i=1}^{N} \pi_i = 1$.

Em geral, a probabilidade de emissão de estado é dada por uma combinação de funções densidade de probabilidade gaussianas no espaço k-dimensional de características do problema de acordo com a Equação (2.30).

$$b_i(o_t) = \sum_{m=1}^{N_g} c_{im} G(o_t, \mu_{im}, U_{im}) \tag{2.30}$$

sendo $G(o_t, \mu_{im} U_{im})$ a função densidade de probabilidade gaussiana multidimensional da m-ésima componente do estado i, definida pela Equação(2.31):

$$G(o_t, \mu_{im} U_{im}) = \frac{1}{(2\pi)^{dim/2} |U_{im}|^{1/2}} e^{-[(o_t - \mu_{im}) U_{im}^{-1} (o_t - \mu_{im})']/2} \tag{2.31}$$

em que d_{im} é a dimensão do vetor de observações (características ou atributos), c_{im} é o peso da m-ésima componente gaussiana do estado i, N_g é o número de componentes gaussianas, μ_{im} e U_{im} são a média e a matriz de covariância, respectivamente.

Os parâmetros do HMM são ajustados por meio de treinamento não-supervisionado baseado em EM de acordo com o algoritmo de Baum-Welch (BAHL *et al.*, 1983). Após o treinamento e ajuste dos parâmetros HMM, utilizam-se as observações juntamente com o algoritmo de Viterbi (RABINER, 1989) para a determinação da sequência de estados mais prováveis.

Para o classificador HMM utilizado neste trabalho, variou-se o número de componentes

gaussianas em cada mistura entre 2 e 10 (2, 4, 6, 8 e 10), definidos de forma empírica. Além disso, deve-se destacar que o modelo HMM utilizado possui 6 estados, conforme ilustrado na Figura 2.32, sendo 3 estados para modelarem o comportamento dos trechos da via permanente contendo dormentes danificados e outros três estados para modelarem trechos com dormentes saudáveis. Deve-se destacar ainda que, após o treinamento do modelo não-supervisionado, é necessário estabelecer uma correspondência entre as classes obtidas e os diagnósticos desejados, o que pode ser realizado a partir da observação das saídas produzidas e a comparação com os dados de treinamento conhecidos. Neste sentido, foi implementado uma rotina específica de determinação das classes de defeito para o HMM.

2.6. Avaliação

Em linhas gerais, a avaliação de um modelo de classificação é feita a partir da comparação entre as classes preditas pelo modelo e as verdadeiras de cada amostra. Todas as métricas de classificação têm como objetivo comum medir quão distante o modelo está da classificação ideal, porém fazem isto de formas diferentes. Nesta pesquisa serão abordadas algumas métricas para a classificação binária desenvolvida, a fim de verificar o desempenho dos modelos. Nesse sentido, serão apresentadas nesta seção algumas dessas métricas.

2.6.1. Matriz de Confusão

Uma forma bastante simples de visualizar a performance de um modelo de classificação é através de uma Matriz de Confusão (MC). Em análise preditiva, a MC é, normalmente, uma tabela com duas linhas e duas colunas que relata o Número de Falsos Positivos (NFP), Número de Falsos Negativos (NFN), Número de Verdadeiros Positivos (NVP) e Número de Verdadeiros Negativos (NVN). No caso desta aplicação, o NFP pode ser computado pela quantidade de dormentes saudáveis que foram diagnosticados como defeituosos. O NFN está relacionado com a quantidade de dormentes defeituosos que foram diagnosticados como saudáveis. O NVP é feito pela quantidade de dormentes defeituosos que foram diagnosticados corretamente. Já o NVN diz respeito à quantidade de dormentes saudáveis que foram diagnosticados corretamente. Entretanto, para esta pesquisa, outras métricas serão apresentadas na matriz de confusão. A MC apresentada na avaliação dos classificadores segue o modelo apresentado pela Figura 2.33.

Considerando a parte colorida do modelo utilizado, as linhas da matriz correspondem à classe predita (*Output Class*) e as colunas representam a classe verdadeira (*Target Class*). A diagonal principal correspondem às observações que estão corretamente classificadas (cor verde), enquanto as células da diagonal secundária correspondem às observações classificadas incorretamente (cor vermelha). A contagem das métricas são apresentadas tanto em números absolutos quanto em porcentagens da classe real, já que o número de exemplos em cada classe pode variar. A coluna na extrema direita da matriz mostra as porcentagens de todos os exem-

plos previstos para pertencer a cada classe que são classificados correta e incorretamente. Essas métricas são frequentemente chamadas de precisão (ou Valor Preditivo Positivo, VPP) e taxa de descoberta falsa (ou Valor Preditivo Negativo, VPN). A primeira mede a probabilidade da presença do defeito quando o diagnóstico for positivo. A segunda é a probabilidade da ausência do defeito quando o diagnóstico for negativo. Da mesma forma, as métricas presentes na linha da parte inferior do gráfico são frequentemente chamadas de sensibilidade (ou Taxa de Verdadeiro Positivo, TVP) e especificidade (ou Taxa de Verdadeiro Negativo, TVN). A primeira é a probabilidade do resultado positivo nos dormentes defeituosos (verdadeiro positivo). A segunda é a probabilidade de resultado negativo nos dormentes saudáveis (verdadeiro negativo). A célula no canto inferior direito do gráfico mostra a acurácia, que nos diz quantos exemplos foram de fato classificados corretamente, independente da classe.

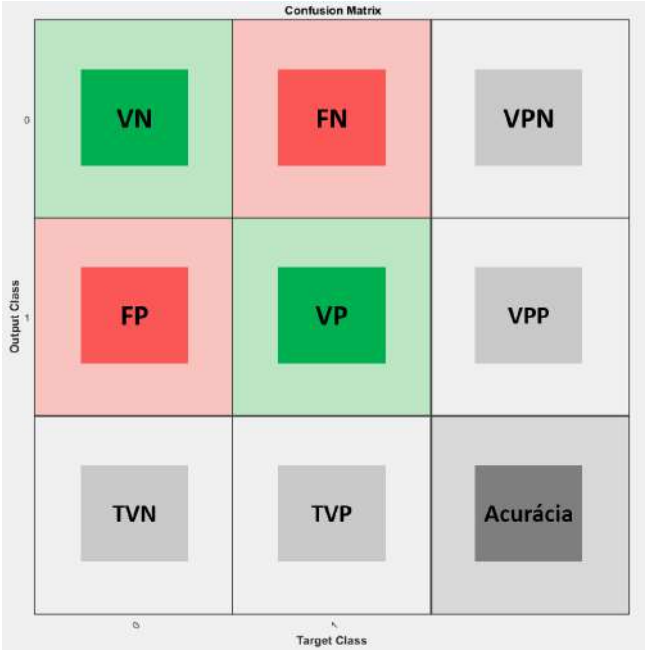


Figura 2.33: Modelo da MC utilizada na pesquisa.
Fonte: O autor.

O cálculo da sensibilidade, especificidade, precisão, valor preditivo negativo e acurácia pode ser feito a partir das equações 2.32, 2.33, 2.34, 2.35 e 2.36, respectivamente.

$$sensibilidade = \frac{NVP}{NVP + NFN} \tag{2.32}$$

$$especificidade = \frac{NVN}{NVN + NFP} \tag{2.33}$$

$$precisão = \frac{NVP}{NVP + NFP} \tag{2.34}$$

$$VPN = \frac{NVP}{NVP + NFP} \tag{2.35}$$

$$acurácia = \frac{NVP + NVN}{NVP + NVN + NFP + NFN} \tag{2.36}$$

De modo geral, apesar do cálculo e da importância de todas essas métricas, para este trabalho especificamente, algumas delas não são tão relevantes. As duas que trazem maior significado para o problema é a taxa de verdadeiros positivos (*sensibilidade*) e a taxa de falsos positivos ($1 - especificidade$), ambas representadas na última linha da MC.

2.6.2. Métricas Principais

As métricas apresentadas anteriormente são calculadas nesta pesquisa a partir da quantidade de amostras presentes em um determinado conjunto de dados. Vale lembrar que o período de amostragem espacial dos dados é de 25 cm e, portanto, menor do que a distância média entre dormentes adjacentes, que varia tipicamente entre 55 e 60 cm. Nesse sentido, para que houvesse uma interpretação coerente com relação ao diagnóstico dos dormentes, duas outras métricas foram avaliadas. Neste trabalho, as métricas foram definidas como: Taxa de Acerto (TA) e Taxa de Esforço Desnecessário (TED). Portanto, o que diferencia essas duas novas métricas das anteriores, é o fato de serem calculadas a partir do diagnóstico dos dormentes e não sobre a quantidade de amostras.

A TA é definida pela razão entre o número de dormentes identificados corretamente como defeituosos e o número total de dormentes danificados existentes nos elementos da EFVM que foram considerados no experimento. Vale ressaltar que a TA não é a métrica sensibilidade apresentada na Seção 2.6.1, apesar de parecer. Enquanto a sensibilidade mede a taxa das amostras diagnosticadas corretamente como defeituosas, a TA meda a taxa de localização dos dormentes danificados (isso acontece pelo motivo explicado anteriormente de que a amostragem dos dados é feita em intervalos diferentes do espaçamento entre os dormentes).

A TED, é definida pela razão entre o número de dormentes identificados incorretamente como danificados e a quantidade total de dormentes saudáveis. Aqui, vale reforçar que a TED não é a mesma medida da taxa de falsos positivos, apesar de parecer. Isso acontece pelo mesmo motivo supracitado; enquanto uma medida avalia a quantidade de amostras, a outra está preocupada com a quantidade de dormentes avaliados nos experimentos. Além disso, a utilização dessas métricas traz uma interpretação mais coerente para a compreensão e aplicação por parte da empresa, uma vez que medem o percentual do esforço que precisa ser feito para se obter uma certa taxa de localização dos dormentes defeituosos.

Neste ponto, deve-se destacar que a estratégia adotada no protocolo experimental consiste em realizar o diagnóstico a partir do CC no sentido de destacar os trechos da EFVM nos quais o modelo classificador indica a existência de dormentes de aço danificados. Assim, para

um determinado valor aceitável da máxima TED, haverá um desempenho correspondente em termos da TA. Neste trabalho, admitiu-se uma TED máxima de 40%, com o intuito de encontrar um desempenho satisfatório na detecção de trechos contendo dormentes defeituosos e, ao mesmo tempo, procurando reduzir o esforço durante o processo de inspeção. Em outras palavras, isso significa que para obter uma certa TA na localização dos dormentes, a equipe de inspeção deverá exercer um esforço desnecessário de no máximo 40%. De todo modo, este valor pode ser alterado e, consequentemente, afetará o desempenho com relação a TA, como será apresentado na seção 3.1.2.

Essa possível flexibilidade com relação à TED se deve principalmente pela quantidade de variações que foram realizadas nos experimentos. Dessa forma, os resultados apresentados na Seção 3.1 (primeira fase da pesquisa) evidenciam essa possibilidade. A partir do momento em que define-se uma configuração única para o experimento, é possível que nem todos os desempenhos quanto à TED se mantenham abaixo de 40%, como pode ser visto na seção 3.2 (segunda fase da pesquisa). De todo modo, durante a seleção dos melhores parâmetros em cada uma das configurações, buscou-se encontrar aqueles que entregavam uma TED menor ou próximo a 40%.

3. Análise dos Resultados

Neste capítulo serão discutidos os resultados obtidos em cada uma das duas fases implementadas. Na Seção 3.1 será apresentado a primeira abordagem, indicando a melhor configuração obtida dentre todas as técnicas escolhidas. A segunda abordagem, apresentada na Seção 3.2, aponta os resultados do sistema *ensemble* desenvolvido, com o intuito de potencializar os desempenhos obtidos na etapa anterior.

3.1. Primeira Fase do Experimento

Inicialmente, os experimentos planejados visavam a modelagem, de forma conjunta, de todos os elementos das diversas supervisões e das duas linhas. Os primeiros resultados apontaram para a dificuldade de avanço nesta direção, uma vez que os desempenhos obtidos não ultrapassavam os 70% de taxa de acerto na localização dos dormentes danificados. Assim, considerando-se as variações significativas que podem existir em características do solo e do clima ao longo da ferrovia, dentre outras, optou-se por modificar a estratégia de modelagem do problema, que passou a consistir na determinação de um modelo para cada configuração. Dessa forma, utilizaram-se os elementos das supervisões de Conselheiro Pena (CP), Governador Valadares (GV) e Mário Carvalho (MR), conforme apresentado na Tabela 3.1, totalizando 12 diferentes configurações.

Tabela 3.1: Configurações analisadas.

Supervisão	Tipo	Linha
CP	Curvas	1
		2
	Tangentes	1
		2
GV	Curvas	1
		2
	Tangentes	1
		2
MR	Curvas	1
		2
	Tangentes	1
		2

Fonte: O autor.

Apresenta-se nas tabelas 3.2 e 3.3 a quantidade de elementos e instâncias em cada uma das configurações. Como foi apresentado na Seção 2.3.3, a relação entre dados saudáveis e defeituosos encontra-se desequilibrada. Desequilíbrio de dados refere-se à distribuição desigual de classes dentro de um conjunto de dados, ou seja, há muito menos eventos de uma classe em comparação com a(s) outra(s). Para contornar o problema de desequilíbrio de classe, as instâncias de treinamento são reamostradas. O conceito básico é alterar as proporções das classes (distribuição a priori) dos dados de treinamento para obter um classificador que possa prever efetivamente a classe minoritária (os dormentes de aço com defeito). Para resolver esse

problema, foram ajustadas as proporções de classe dos conjuntos de dados sub-amostrando aleatoriamente a classe maior. Portanto, do total indicado nas tabelas, 50% refere-se à classe dos saudáveis e os outros 50% à classe dos defeituosos.

Tabela 3.2: Estrutura da base de dados do primeiro experimento (curvas).

Supervisão/Linha	Número de Elementos	Número de instâncias
CP/1	13	6824
CP/2	20	11728
GV/1	28	5298
GV/2	20	5536
MR/1	11	6792
MR/2	21	3960

Fonte: O autor.

Tabela 3.3: Estrutura da base de dados do primeiro experimento (tangentes).

Supervisão/Linha	Número de Elementos	Número de instâncias
CP/1	7	2706
CP/2	8	300
GV/1	25	5312
GV/2	15	6530
MR/1	11	378
MR/2	22	1396

Fonte: O autor.

Novamente, foram utilizados os dados referentes às inspeções 3, 4 e 5 para efeito de treinamento e seleção de modelos/parâmetros, enquanto os dados da inspeção 6 foram utilizados para teste. Foi desenvolvido um sistema que produz um modelo para cada variação aplicada, considerando as técnicas utilizadas na pesquisa:

- **Tamanho das janelas:** 32, 64 e 128;
- **Extração de características:** análise espacial, espectral e wavelet;
- **Seleção de características ou redução de dimensionalidade:** analise de componentes principais e a razão discriminante de Fisher;
- **Número de parâmetros/componentes:** 12 variações no número de parâmetros da FDR e no número de componentes principais da PCA;
- **Classificadores:** redes neurais artificiais, modelo de mistura de gaussianas e modelos ocultos de Markov;

- **Complexidade dos classificadores:** 5 variações na complexidade de cada classificador, de acordo com a característica individual;
- **Combinações de sinais:** os sinais de geometria da via permanente foram analisados individualmente e de forma combinadas entre eles, totalizando trinta e uma possíveis combinações.

Os experimentos iniciais foram executados a partir dos dados de curvas e os resultados obtidos procuraram identificar a combinação de técnicas que forneceram a maior taxa de acerto na identificação dos dormentes danificados, considerando-se apenas aqueles desempenhos que apresentaram uma TED menor do que 40%. Foram avaliados os desempenhos dos 100.440 modelos gerados a partir das variações aplicadas. Dessa forma, buscou-se encontrar os intervalos de confiança para a média (com nível de significância de 5%), utilizando-se os 10 melhores modelos de cada configuração. Apresentam-se na Tabela 3.4 os resultados para as curvas de cada supervisão/linha considerada neste trabalho.

Tabela 3.4: Melhor desempenho geral (considerando todas as técnicas utilizadas) para as curvas de cada supervisão/linha, com base nas duas métricas apresentadas.

Supervisão/Linha	**Configuração**	**TA (%)**	**TED (%)**
CP/1	Janela 128 - Parametrização Espacial - PCA - HMM	96 ± 2	35 ± 2
CP/2	Janela 128 - Parametrização Espacial - FDR - GMM	86 ± 2	23 ± 3
GV/1	Janela 128 - Parametrização Espacial - FDR - HMM	89 ± 1	26 ± 2
GV/2	Janela 128 - Parametrização Espacial - FDR - RNA	84 ± 2	38 ± 1
MR/1	Janela 32 - Parametrização Espacial - FDR - HMM	99 ± 1	34 ± 3
MR/2	Janela 32 - Parametrização Wavelet - PCA - HMM	90 ± 3	36 ± 2

Fonte: O autor.

Na sequência, com o intuito de se obter a melhor configuração experimental comum a todas as supervisões/linhas, realizaram-se análises estatísticas comparativas entre os modelos. A partir dos resultados foi possível fazer comparações entre os desempenhos obtidos por cada técnica utilizada em cada etapa. Nesse sentido, foi utilizado um método de pontuação durante as comparações entre eles. Para descobrir qual tamanho da janela de dados foi responsável pelos melhores resultados, primeiramente fixava-se uma configuração (por exemplo, GV/Linha1/Espacial/PCA/RNA) e os 200 melhores resultados gerados por cada janela eram comparados para a mesma configuração fixada. As comparações foram realizadas entre duas técnicas separadamente (tamanho 32 x 64, tamanho 32 x 128 e tamanho 64 x 128). Dessa forma, o mesmo procedimento foi realizado para cada variação da configuração que era fixada e, ao final, tinha-se a pontuação total de cada tamanho de janela. Além do tamanho das janelas de dados, as técnicas de extração e seleção de características também foram comparadas, seguindo a mesma ideia apresentada.

Nesse sentido, foram realizadas duas análises estatísticas para efeito de comparação entre os resultados obtidos pelas diferentes técnicas. A primeira é a análise de variância unidirecional (ANOVA, do inglês, *One-way analysis of variance*) onde a proposta é determinar se os dados de vários grupos (níveis) de um fator possuem uma média comum (MARTINS, 2001). A ANOVA testa a hipótese de que todas as médias dos grupos são iguais contra a hipótese alternativa de que pelo menos um grupo é diferente dos demais. O outro é o teste de Wilcoxon (KERBY, 2014), que testa a hipótese nula de que os dados em x e y são amostras de distribuições contínuas com medianas iguais, contra a alternativa de que não são. Para ambos os métodos, foi utilizado um nível de significância de 5%.

Seguindo essa lógica para todos os casos possíveis, a análise forneceu o resultado indicado na Figura 3.1, na qual pode-se destacar a janela de tamanho 128 como a vencedora.

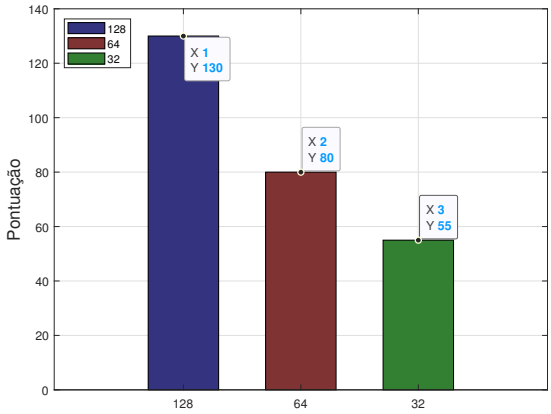


Figura 3.1: Determinação da melhor janela de dados.
Fonte: O autor.

O mesmo procedimento de comparação apresentado acima também foi utilizado para a obtenção dos próximos resultados que serão apresentados, ou seja, todas as condições do experimento são fixadas, com exceção do aspecto que se deseja analisar. Dessa forma, o método de parametrização que se mostrou mais adequado, conforme é possível verificar na Figura 3.2, foi o espacial.

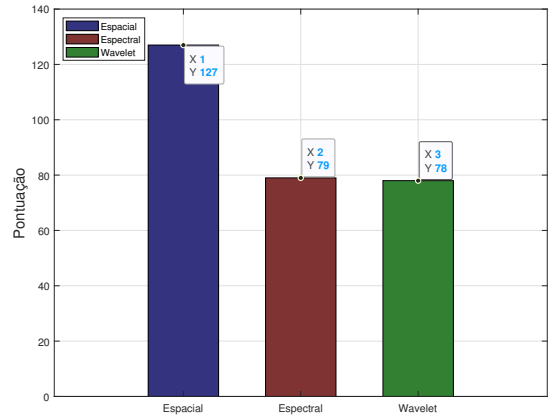


Figura 3.2: Determinação do melhor método de extração de características.
Fonte: O autor.

Na sequência, investigou-se também o melhor método de seleção de características ou redução de dimensionalidade, e os resultados comparativos mostram um melhor desempenho do FDR, conforme indicado na Figura 3.3.

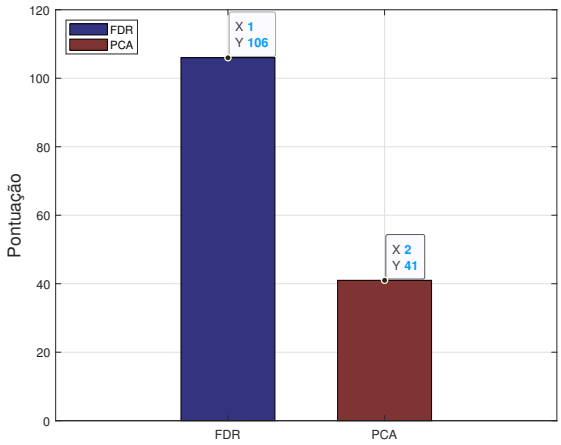


Figura 3.3: Determinação do melhor método de seleção de características.
Fonte: O autor.

Posteriormente, uma vez determinada a melhor configuração experimental comum a todas as supervisões/linhas investigadas, foram selecionados, agora, os resultados para os elementos curvas, a partir desta melhor configuração comum, variando-se apenas os classificadores. Dessa forma, foram obtidos os resultados presentes na Tabela 3.5. Um ponto importante a se destacar foi que nos melhores resultados gerais apresentados na Tabela 3.4, apareceram 3 configurações contendo as técnicas vencedoras (CP/Linha2, GV/Linha1 e GV/Linha2). Na comparação com as outras 3 configurações, apesar da mínima redução na TA, o resultado ainda permaneceu satisfatório para a condição, validando a configuração comum para todas as supervisões/linha. Além disso, pode-se notar que os três classificadores utilizados apareceram na mesma quantidade de vezes nos melhores resultados, mostrando que a princípio não existe uma superioridade de nenhum deles.

Tabela 3.5: Melhor desempenho para as curvas de cada supervisão/linha, fixando os melhores parâmetros (variando apenas os classificadores), com base nas duas métricas apresentadas.

Supervisão/Linha	**Configuração**	**TA (%)**	**TED (%)**
CP/1	Janela 128 - Parametrização Espacial - FDR - HMM	95 ± 1	34 ± 3
CP/2	Janela 128 - Parametrização Espacial - FDR - GMM	86 ± 2	23 ± 3
GV/1	Janela 128 - Parametrização Espacial - FDR - GMM	89 ± 1	26 ± 2
GV/2	Janela 128 - Parametrização Espacial - FDR - RNA	84 ± 2	38 ± 1
MR/1	Janela 128 - Parametrização Espacial - FDR - RNA	98 ± 1	14 ± 5
MR/2	Janela 128 - Parametrização Espacial - FDR - HMM	84 ± 3	38 ± 2

Fonte: O autor.

Por fim, realizou-se o mesmo procedimento para a modelagem das tangentes, considerando apenas a melhor configuração experimental comum a todas as supervisões/linhas. Foram calculados também os intervalos de confiança para a média (com nível de significância de 5%), utilizando-se os 10 melhores resultados de cada configuração. Apresentam-se os desempenhos na Tabela 3.6.

Tabela 3.6: Melhor desempenho para as tangentes de cada supervisão/linha, fixando os melhores parâmetros (variando apenas os classificadores), com base nas duas métricas apresentadas.

Supervisão/Linha	**Configuração**	**TA (%)**	**TED (%)**
CP/1	Janela 128 - Parametrização Espacial - FDR - RNA	95 ± 1	24 ± 3
CP/2	Janela 128 - Parametrização Espacial - FDR - RNA	99 ± 1	15 ± 3
GV/1	Janela 128 - Parametrização Espacial - FDR - RNA	86 ± 1	36 ± 2
GV/2	Janela 128 - Parametrização Espacial - FDR - RNA	81 ± 2	34 ± 2
MR/1	Janela 128 - Parametrização Espacial - FDR - RNA	99 ± 1	9 ± 1
MR/2	Janela 128 - Parametrização Espacial - FDR - GMM	97 ± 3	30 ± 2

Fonte: O autor.

Assim como para as curvas, os resultados apresentados para as tangentes indicaram um desempenho superior a 80% para todas as configurações. Ao se comparar os resultados das tabelas 3.5 e 3.6 para uma mesma configuração, nota-se que em alguns casos os resultados para as curvas são superiores, em outros são os da tangentes. Nesse sentido, a princípio, não é possível afirmar se o problema das tangentes é mais fácil de ser tratado do que o das curvas, ou vice-versa. Outro ponto importante a se destacar sobre os resultados das tangentes é com relação ao classificador. Diferente do que aconteceu para as curvas, a RNA apareceu como o melhor classificador em 5 das 6 configurações analisadas, mostrando uma certa superioridade quando trata-se das tangentes.

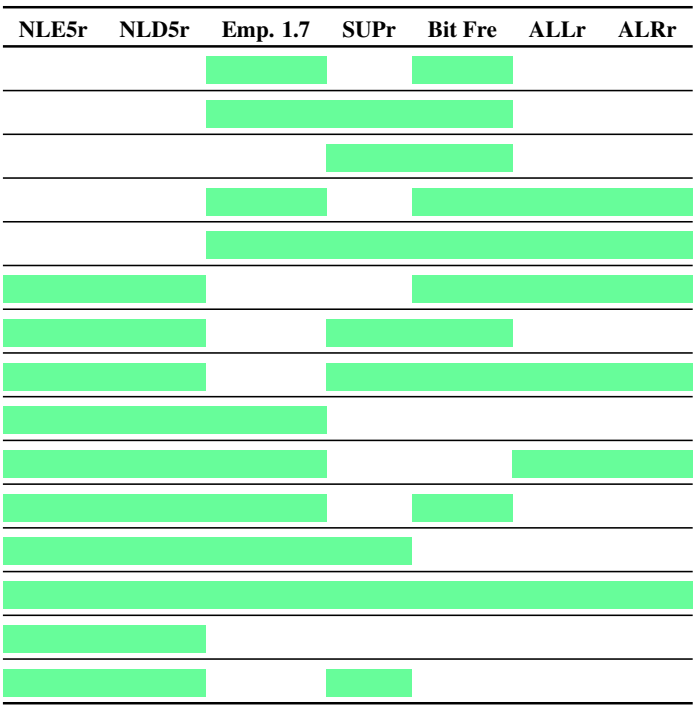
3.1.1. Conjunto dos Sinais

Os primeiros experimentos foram realizados a partir da variação das 31 combinações mencionadas na seção anterior. Dessa forma, foi possível analisar quais das combinações estavam mais relacionadas com as curvas ou com as tangentes. Em outras palavras, foram verificadas quais combinações eram responsáveis pelos melhores desempenhos das curvas e quais eram das tangentes. Apresentam-se nas tabelas 3.7 e 3.8, as combinações de sinais que forneceram os melhores resultados para cada tipo dos elementos. Da esquerda para a direita, representam os sinais de nivelamento esquerdo (NLE5r), nivelamento direito (NLD5r), empeno (Emp. 1.7), superelevação (SUPr), bitola (Bit Fre), alinhamento esquerdo (ALLr) e alinhamento direito (ALRr). Vale ressaltar que a ordem combinações apresentadas em cada tabela, não diz respeito ao desempenho de cada uma delas.

Além disso, outras análises permitiram determinar a frequência de ocorrência de cada sinal geométrico para os melhores modelos gerados no experimento, a qual se encontra descrita

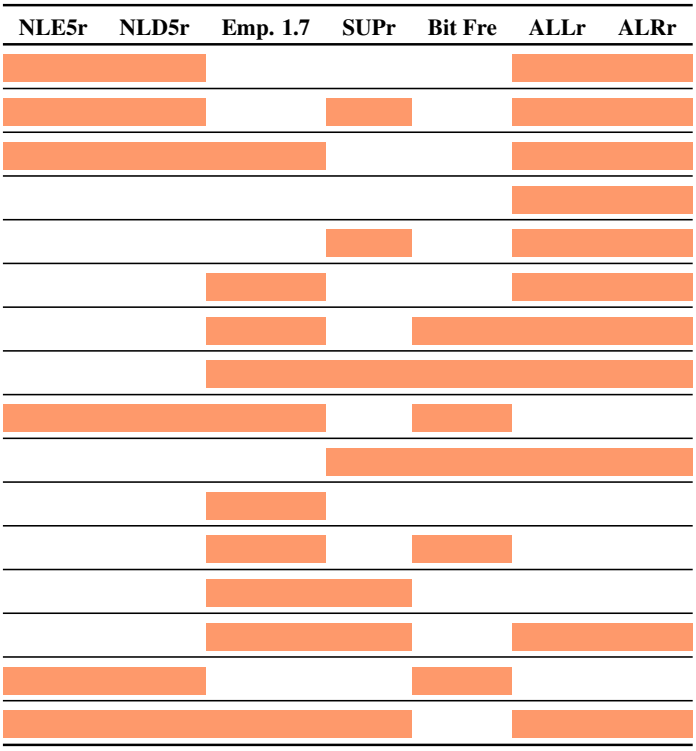
na Figura 3.4. Este resultado destacou a importância dos sinais de nivelamento, bitola e empeno para caracterização dos danos em dormentes de aço localizados nas curvas, enquanto o alinhamento, empeno e superelevação se mostraram mais relevantes para o caso das tangentes.

Tabela 3.7: Combinações de sinais com os melhores resultados para as curvas.

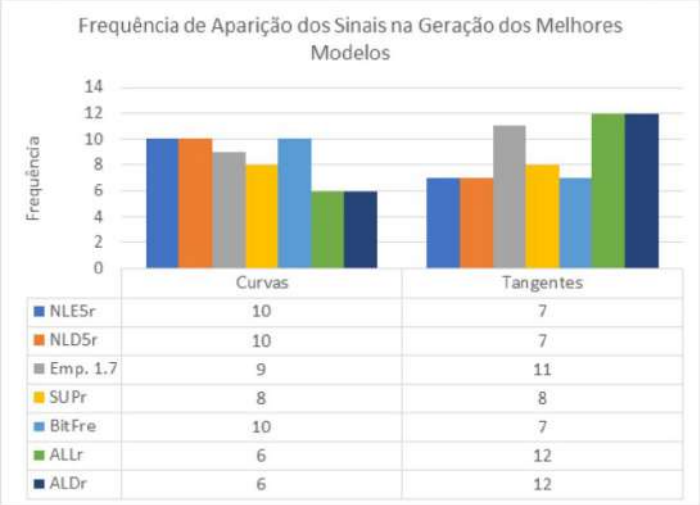


Fonte: O autor.

Tabela 3.8: Combinações de sinais com os melhores resultados para as tangentes.



Fonte: O autor.



	Curvas	Tangentes
NLE5r	10	7
NLD5r	10	7
Emp. 1.7	9	11
SUPr	8	8
BitFre	10	7
ALLr	6	12
ALDr	6	12

Figura 3.4: Frequência de ocorrência dos sinais geométricos.
Fonte: O autor.

3.1.2. Efeito da Máxima Taxa de Esforço Desnecessário Aceitável

Um aspecto importante que deve ser ressaltado é a definição da máxima taxa de esforço desnecessário aceitável, o que corresponde, em termos práticos, ao percentual de esforço em vão que deverá ser empregado pelas equipes de inspeção para alcançar uma certa taxa de localização dos dormentes defeituosos. Em síntese, os resultados apontam que quanto maior for esta máxima taxa aceitável, mais próximo de 100% será a taxa de acerto na identificação dos problemas, conforme se pode observar nas figuras 3.5 e 3.6.

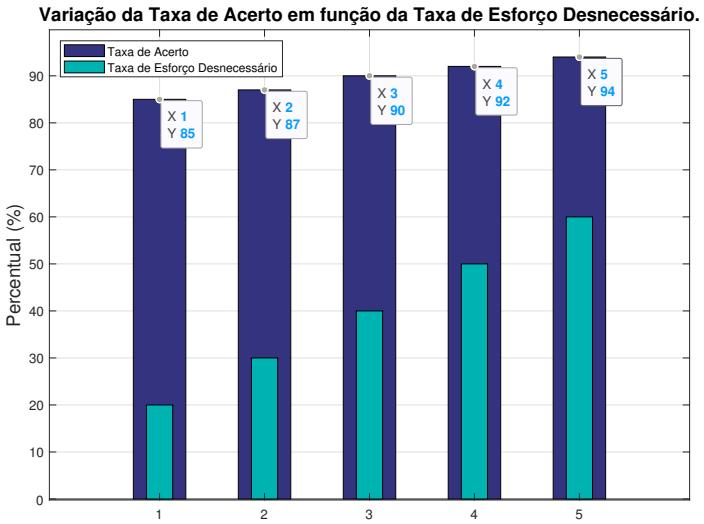


Figura 3.5: Desempenho médio obtido para a melhor configuração (curvas).
Fonte: O autor.

Elas apresentam a taxa de acerto média na detecção dos dormentes de aço danificados em função da máxima taxa de esforço desnecessário aceitável, utilizando-se a melhor configuração experimental (janela com 128 amostras, parâmetros espaciais, utilizando o método

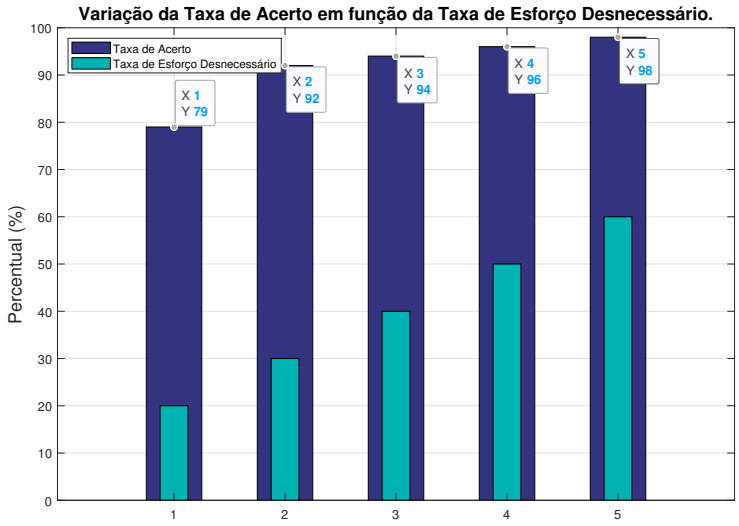


Figura 3.6: Desempenho médio obtido para a melhor configuração (tangentes).
Fonte: O autor.

FDR e considerando apenas o classificador RNA). Para isso, de forma a exemplificar, foi alterado o valor da TED aceitável entre 20% e 60%. Dessa forma, nota-se que é possível atingir o equilíbrio desejado entre o custo operacional da inspeção e a eliminação dos dormentes danificados da via permanente para assegurar a confiabilidade da ferrovia.

3.1.3. Geração do Diagnóstico

Após a execução da etapa de treinamento e validação dos modelos, realizou-se a etapa de inferência com os dados da 3ª inspeção do ano de 2020. Os resultados serviram como diagnóstico para os elementos das supervisões de CP, GV e MR com a finalidade de verificação em campo. Para isso, foram gerados arquivos indicando as posições de defeito para cada uma das supervisões/linha. Apresentam-se na Tabela 3.9 as informações presentes nos arquivos gerados. Destacam-se as colunas 6 e 7, as quais indicam os intervalos de defeito do diagnóstico em metros, baseado na posição de início deste mesmo elemento.

Tabela 3.9: Segmento de um arquivo de diagnóstico gerado para inspeção em campo.

ID	EH	SUP	Linha	Início elem (m)	Extensão (m)	Início defeito (m)	Fim defeito (m)
CURVA 2	30/31	CP	Linha 1	186999	233	14	32
CURVA 2	30/31	CP	Linha 1	186999	233	64	74
CURVA 3	30/31	CP	Linha 1	187545	311	28	34
CURVA 10	30/31	CP	Linha 1	191402	267	74	75
CURVA 10	30/31	CP	Linha 1	191402	267	146	156
CURVA 10	30/31	CP	Linha 1	191402	267	158	165
CURVA 10	30/31	CP	Linha 1	191402	267	235	238

Fonte: O autor.

3.2. Segunda Fase do Experimento

Os resultados e conhecimentos obtidos na etapa anterior serviram como base para implementar o sistema desenvolvido nesta fase. Dessa forma, manteve-se fixo o tamanho da janela de dados (128), a técnica de extração de características (espacial) e a técnica de seleção de característica (FDR). Além disso, para cada configuração e para cada classificador, utilizou-se uma única combinação de sinais. Ao reduzir o espaço de busca, optou-se por inserir novos classificadores (SVM e ADB), a fim de criar um sistema *ensemble* mais robusto.

Assim como na etapa anterior, foram estudadas as supervisões de CP, GV e MR com os dados referentes às inspeções 3, 4 e 5 para efeito de treinamento e seleção de modelos e parâmetros, enquanto os dados da inspeção 6 foram utilizados para teste dos modelos. Entretanto, vale relembrar que os dados utilizados nesse experimento são referentes, agora, ao ano de 2020. Apresenta-se nas tabelas 3.10 e 3.11 a quantidade de elementos e de instâncias de treinamento em cada configuração escolhida. Vale ressaltar que a quantidade de instâncias utilizadas em cada configuração seguiu o mesmo critério apresentado na seção 3.1.

Tabela 3.10: Quantidade de elementos e instâncias por configuração (curvas).

Supervisão/Linha	Quantidade de elementos	Quantidade de instâncias
CP/1	11	1772
CP/2	24	9078
GV/1	17	1654
GV/2	20	2630
MR/1	32	24104
MR/2	28	3786

Fonte: O autor.

Tabela 3.11: Quantidade de elementos e instâncias por configuração (tangentes).

Supervisão/Linha	Quantidade de elementos	Quantidade de instâncias
CP/1	12	2562
CP/2	8	3700
GV/1	19	1238
GV/2	18	500
MR/1	24	600
MR/2	12	5548

Fonte: O autor.

3.2.1. Resultados Individuais

A estrutura do sistema desenvolvido na na primeira fase do experimento basicamente se manteve a mesma, com exceção das técnicas que foram fixadas. Dessa forma, os experimentos iniciais consistiram na execução do sistema para cada um dos classificadores individualmente, de forma que fosse possível verificar o potencial individual e, ao mesmo tempo, gerar

os modelos que seriam utilizados no *ensemble*. Portanto, para cada variação das configurações, realizou-se a repetição do experimento 10 vezes e o desempenho médio obtido de TA e TED, juntamente com o intervalo de confiança para a média (com nível de significância de 5%) estão apresentados nas tabelas 3.12, 3.13 e 3.14.

Tabela 3.12: Resultados para a supervisão de CP.

Conselheiro Pena (CP)								
	Curvas				**Tangentes**			
	Linha 1		**Linha 2**		**Linha 1**		**Linha 2**	
	TA (%)	**TED (%)**	**TA (%)**	**TED (%)**	**TA (%)**	**TED (%)**	**TA (%)**	**TED (%)**
RNA	99 ± 1	6 ± 1	74 ± 6	33 ± 5	98 ± 2	35 ± 5	86 ± 4	32 ± 5
GMM	96 ± 1	7 ± 1	68 ± 7	26 ± 4	97 ± 4	23 ± 2	84 ± 7	27 ± 4
HMM	87 ± 8	53 ± 6	69 ± 8	45 ± 2	85 ± 8	43 ± 6	71 ± 9	44 ± 5
SVM	99 ± 1	2 ± 1	82 ± 1	37 ± 1	99 ± 1	24 ± 1	79 ± 1	26 ± 1
ADB	99 ± 1	4 ± 1	79 ± 4	35 ± 3	95 ± 2	20 ± 1	91 ± 1	38 ± 1

Fonte: O autor.

Tabela 3.13: Resultados para a supervisão de GV.

Governador Valadares (GV)								
	Curvas				**Tangentes**			
	Linha 1		**Linha 2**		**Linha 1**		**Linha 2**	
	TA (%)	**TED (%)**	**TA (%)**	**TED (%)**	**TA (%)**	**TED (%)**	**TA (%)**	**TED (%)**
RNA	94 ± 8	28 ± 3	81 ± 5	21 ± 4	72 ± 1	17 ± 2	95 ± 8	40 ± 6
GMM	85 ± 8	27 ± 2	80 ± 3	14 ± 1	70 ± 8	28 ± 1	99 ± 1	36 ± 2
HMM	81 ± 11	51 ± 3	85 ± 8	50 ± 3	75 ± 10	48 ± 5	90 ± 8	53 ± 7
SVM	99 ± 1	10 ± 1	79 ± 1	19 ± 1	86 ± 1	39 ± 1	99 ± 1	12 ± 1
ADB	93 ± 5	13 ± 2	89 ± 2	39 ± 1	70 ± 5	45 ± 4	99 ± 1	14 ± 2

Fonte: O autor.

Tabela 3.14: Resultados para a supervisão de MR.

Mário Carvalho (MR)								
	Curvas				**Tangentes**			
	Linha 1		**Linha 2**		**Linha 1**		**Linha 2**	
	TA (%)	**TED (%)**	**TA (%)**	**TED (%)**	**TA (%)**	**TED (%)**	**TA (%)**	**TED (%)**
RNA	81 ± 1	14 ± 2	78 ± 10	34 ± 7	97 ± 8	25 ± 2	70 ± 3	32 ± 5
GMM	81 ± 2	20 ± 2	81 ± 4	37 ± 3	80 ± 12	29 ± 1	62 ± 3	26 ± 2
HMM	86 ± 3	49 ± 4	77 ± 13	48 ± 4	93 ± 9	54 ± 3	65 ± 9	50 ± 4
SVM	81 ± 1	12 ± 1	94 ± 1	54 ± 1	99 ± 1	10 ± 1	87 ± 1	60 ± 1
ADB	86 ± 1	31 ± 1	72 ± 4	24 ± 2	99 ± 1	39 ± 6	78 ± 2	29 ± 2

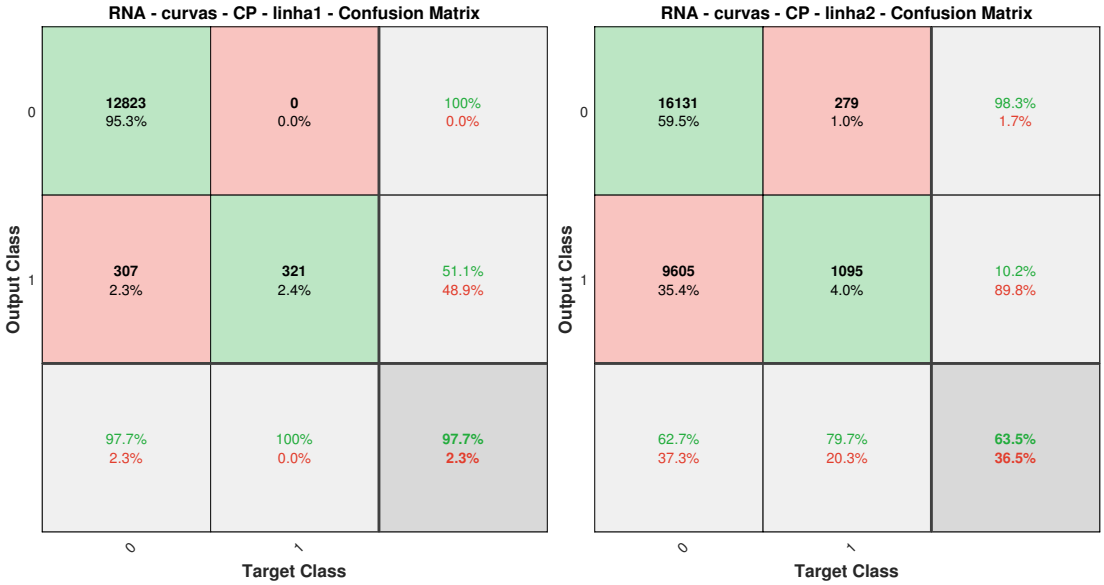
Fonte: O autor.

O motivo de não utilizar a validação cruzada no processo de verificação do desempenho dos modelos foi pelo fato de que a sequência temporal dos dados (a ordem das inspeções) foi levado em consideração para o processo de treinamento e teste. A aleatoriedade na seleção dos conjuntos de dados pode não ser a melhor escolha em alguns casos práticos. Nesse sentido, para tornar a avaliação robusta contra o acaso, optou-se apenas por repetir cada experimento 10 vezes para cada configuração em divisões aleatórias de treinamento/teste (mas mantendo sempre a 6ª inspeção para teste) e relatou-se a média e o intervalo de confiança com nível de significância de 5%.

Ao analisar as tabelas de resultados, algumas informações relevantes podem ser retiradas. O classificador HMM, para todas as configurações, apresentou uma TED maior que 40%, ou seja, ao tentar fixar apenas uma combinação dos sinais, nenhum modelo gerado pelo HMM conseguiu superar este desempenho. Além disso, em muitos casos, mesmo com a TED elevada o HMM retornou uma TA inferior a outros classificadores. Por outro lado, no geral, em se tratando da TA, os 5 classificadores entregaram desempenhos comparáveis (em alguns poucos casos os resultados de um ou outro classificador acabaram se destacando mais que outros) em cada configuração. Esta última análise foi importante para garantir que nenhum classificador pudesse influenciar negativamente o sistema *ensemble* que posteriormente será desenvolvido. Por fim, pode-se dizer que os resultados encontrados nesta etapa do experimento serviram para validar todo o sistema desenvolvido, por evidenciar a sua capacidade de localização dos defeitos em dormente de aço a partir do sinais de geometria da via permanente.

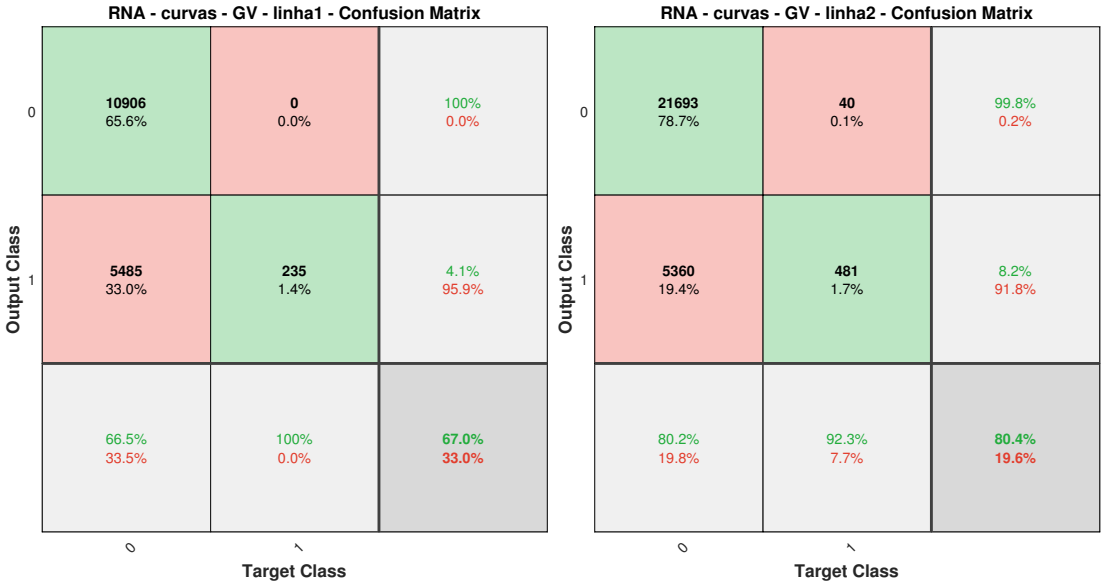
Além dos resultados encontrados para a TA e TED, foram geradas as matrizes de confusão para o melhor modelo de cada uma das configurações. Considerando a quantidade de classificadores utilizados e de modelos gerados (um para cada supervisão/linha), optou-se pela apresentação das matrizes de confusão somente para um classificador (RNA). Para os outros classificadores, os resultados encontram-se em anexo no final deste trabalho. Portanto, apresentam-se nas figuras 3.7a, 3.7b, 3.8a, 3.8b, 3.9a, 3.9b, 3.10a, 3.10b, 3.11a, 3.11b, 3.12a e 3.12b as matrizes de confusão obtidas para cada configuração, a fim de verificar o percentual de algumas das métricas utilizadas e também a quantidade de amostras nos dados de teste.

Dentre as métricas apresentadas, a taxa de falsos positivos (1-especificidade) e a taxa de verdadeiros positivos (sensibilidade) são as mais representativas para o problema em questão. Considerando que os dados de teste estão em sua maioria compostos de dados saudáveis, é natural que a precisão sempre seja muito baixa. Da mesma forma, a acurácia não é uma métrica muito representativa para o problema. Um exemplo disso é quando o modelo acerta grande parte da classe dos saudáveis, mas não é capaz de identificar a classe dos defeituosos. Em um caso como esse, ainda sim é possível que a acurácia encontrada seja alta, mesmo sem identificar nenhum defeito. Isso acontece pelo mesmo motivo apresentado para o caso da precisão.



(a) CP/Linha1. (b) CP/Linha2.

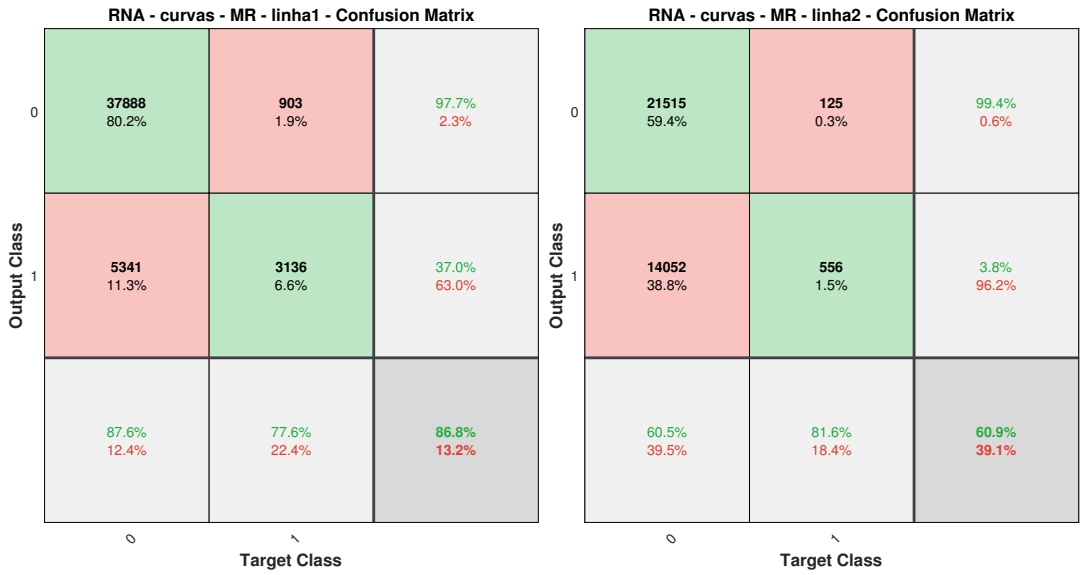
Figura 3.7: Matriz de Confusão para a RNA (Curvas - Conselheiro Pena).
Fonte: O autor.



(a) GV/Linha1. (b) GV/Linha2.

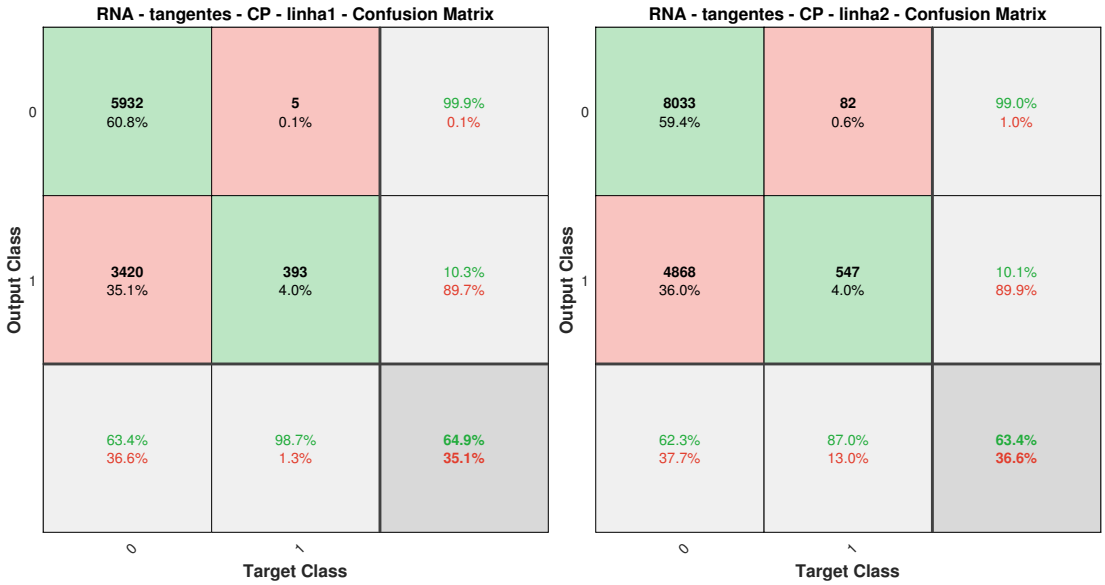
Figura 3.8: Matriz de Confusão para a RNA (Curvas - Governador Valadares).
Fonte: O autor.

Para as curvas, utilizando-se o classificador RNA, o menor percentual de sensibilidade obtido foi de 77,6% para a supervisão de MR (linha 1) e o melhor foi de 100% para CP (linha 1) e GV (linha 1). Com relação à taxa de falsos positivo, o pior caso foi para supervisão de MR (linha 2) com o percentual de 39,5%.



(a) MR/Linha1. (b) MR/Linha2.

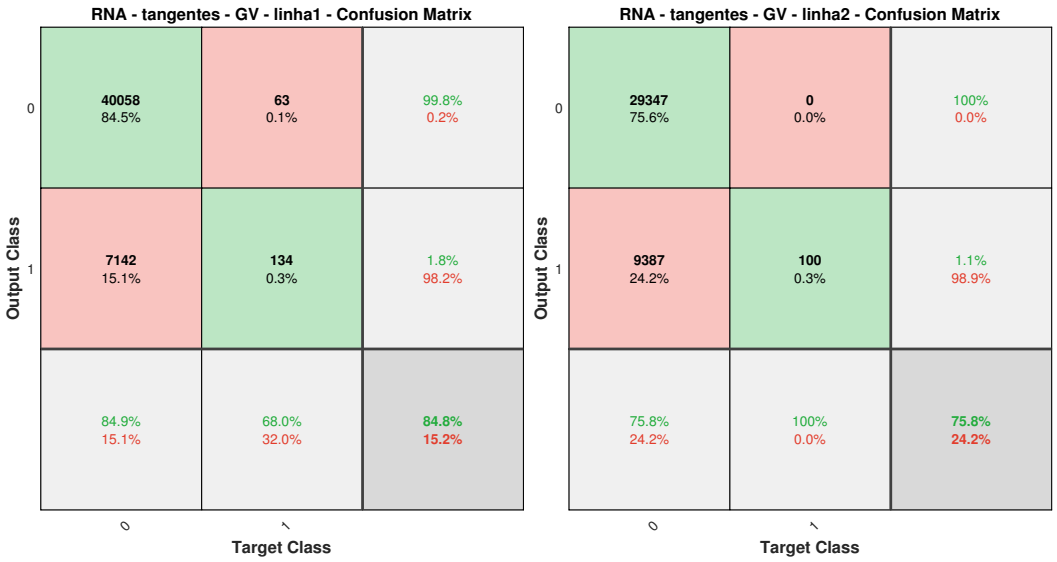
Figura 3.9: Matriz de Confusão para a RNA (Curvas - Mário Carvalho).
Fonte: O autor.



(a) CP/Linha1. (b) CP/Linha2.

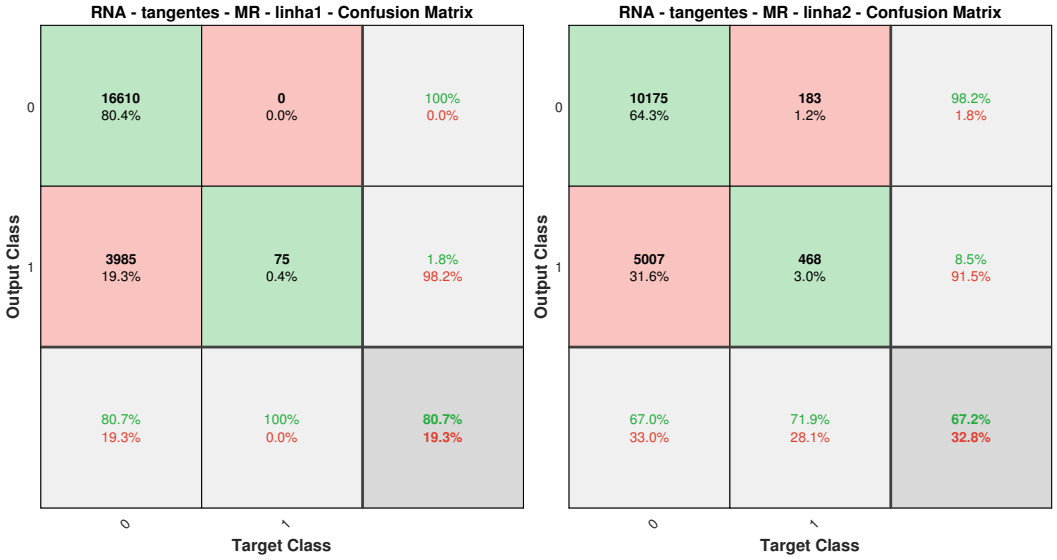
Figura 3.10: Matriz de Confusão para a RNA (Tangentes - Conselheiro Pena).
Fonte: O autor.

Para as tangentes, utilizando-se o classificador RNA, o menor percentual de sensibilidade obtido foi de 68% para a supervisão de GV (linha 1) e os melhores foram de 100% para CP (linha 2) e MR (linha 1). Com relação à taxa de falsos positivo, o pior caso foi para supervisão de GV (linha 2) com o percentual de 37,7%.



(a) GV/Linha1. (b) GV/Linha2.

Figura 3.11: Matriz de Confusão para a RNA (Tangentes - Governador Valadares).
Fonte: O autor.



(a) MR/Linha1. (b) MR/Linha2.

Figura 3.12: Matriz de Confusão para a RNA (Tangentes - Mário Carvalho).
Fonte: O autor.

3.2.2. *Ensembles*

Após a execução dos classificadores de forma individual, foram executados os experimentos para o *ensemble* com os 5 classificadores de forma que fosse possível comparar com os resultados anteriormente obtidos, apresentados nas tabelas 3.12, 3.13 e 3.14. Dentre as técnicas de *ensemble* disponíveis, optou-se pela escolha de duas abordagens mais simples, de forma que fosse possível aproveitar os melhores modelos obtidos no passo anterior. São elas a saber: votação por maioria e grupos por votação.

3.2.2.1. Votação por Maioria

A primeira técnica implementada foi o método de votação por maioria. A ideia por trás dessa estratégia é combinar classificadores de aprendizado de máquina conceitualmente diferentes e usar um voto majoritário (*hard voting*) para prever os rótulos de classe. Tal classificador pode ser útil para um conjunto de modelos de desempenho igualmente bom, a fim de equilibrar suas fraquezas individuais. Neste método, o rótulo de classe previsto para uma determinada amostra é o rótulo de classe que representa a maioria (moda) dos rótulos de classe previstos por cada classificador individual. Por exemplo, se em uma determinada predição um *ensemble* composto por 5 classificadores retornar os seguintes resultados para uma certa amostra: classificador 1 (classe 1), classificador 2 (classe 1), classificador 3 (classe 2), classificador 4 (classe 2) e classificador 5 (classe 1). Pelo método da votação por maioria, a amostra seria classificada como sendo da "classe 1", pelo fato de que 3 dos 5 classificadores predisseram esta classe.

Seguindo a mesma ideia apresentada no exemplo, foi desenvolvido um sistema *ensemble* com os 5 classificadores utilizados nesta pesquisa. Como resultado, foram encontrados os desempenhos (TA e TED) para cada supervisão/linha, considerando curvas e tangentes. Apresenta-se na Tabela 3.15, o desempenho obtido para essa primeira abordagem, com a utilização do voto majoritário.

Tabela 3.15: Resultado da técnica de votação por maioria.

Tipo	Supervisão/Linha	TA (%)	TED (%)
Curvas	CP1	100,00	1,07
	CP2	81,48	45,82
	GV1	100,00	14,35
	GV2	95,83	11,26
	MR1	83,64	10,32
	MR2	82,67	24,87
Tangentes	CP1	98,14	19,84
	CP2	88,89	41,97
	GV1	86,37	20,47
	GV2	100,00	12,12
	MR1	100,00	19,46
	MR2	89,19	41,85

Fonte: O autor.

Sobre os resultados obtidos, o menor percentual de TA encontrado foi de 81,48% enquanto a maior TED foi de 45,82%, ambos para a configuração de CP2/curvas. De modo geral, os resultados encontrados com a estratégia de votação por maioria não superaram em todos os casos o desempenho individual dos classificadores apresentados nas tabelas 3.12, 3.13 e 3.14.

Em alguns casos eles foram bem semelhantes (CP1/curvas e GV1/curvas), em outros o *ensemble* mantém a TA mas consegue uma boa redução da TED (MR1/curvas, GV1/tangentes e MR1/tangentes). No melhor dos casos, o *ensemble* apresenta resultado superior na TA e, ao mesmo tempo, uma queda na TED (GV2/curvas).

3.2.2.2. Grupos por Votação

A segunda técnica aplicada para o desenvolvimento do sistema *ensemble* foi a divisão de grupos por votação. Agora, a saída prevista não receberá mais uma votação majoritária, mas a soma da previsão de cada um dos classificadores. Apresenta-se na Tabela 3.16 um exemplo da aplicação desenvolvida, em que cada coluna representa a predição de cada um dos classifidadores, com exceção da última que representa a soma da predição. Vale lembrar que o "0" indica um diagnóstico saudável e o "1" de defeituoso.

Tabela 3.16: Exemplo da técnica de grupos por votação.

C1	C2	C3	C4	C5	Resultado
1	1	1	1	1	5
1	1	1	1	1	5
1	1	1	0	1	4
1	1	0	1	1	4
1	0	0	1	1	3
0	1	0	1	1	3
0	1	0	1	0	2
0	0	1	0	0	1
0	0	0	0	0	0

Fonte: O autor.

Serão chamadas de pertencentes ao Grupo 5, aquelas amostras em que todos os classificadores apontaram como sendo da classe dos defeituosos ou, em outras palavras, aquelas amostras em que a última coluna da Tabela 3.16 tiver o número 5. Analogamente, serão chamadas de pertencentes ao Grupo 4, aquelas amostras em que 4 dos 5 classificadores apontaram como sendo da classe dos defeituosos. Portanto, serão criados 6 grupos, sendo que cada um deles estará relacionado com o número de classificadores que retornaram um diagnóstico de defeito para uma determinada amostra. Estes grupos foram analisadas individualmente quanto a TA e TED, para que fosse possível entender o potencial de cada um na detecção dos defeitos nos dormentes de aço. Os resultados obtidos para os 6 grupos das curvas e tangentes de cada uma das supervisões/linha estão apresentados nas tabelas 3.17 e 3.18.

Esses resultados evidenciam as diferentes contribuições de TA e TED para cada um dos grupos. Nesse sentido, é possível notar que o Grupo 5, em todos os casos, apresentou a maior TA entre todos os grupos e, além disso, uma baixa TED. Os outros grupos não apresentaram

um comportamento comum em todos os casos, mas seguiram uma lógica, na maioria dos casos, de que o desempenho diminui à medida que parte-se do Grupo 5 para o Grupo 1.

Tabela 3.17: Resultados dos grupos de cada uma das supervisões/linha para as curvas.

CURVAS							
		Grupo 5	**Grupo 4**	**Grupo 3**	**Grupo 2**	**Grupo 1**	**Grupo 0**
CP1	**TA (%)**	100	0,00	0,00	0,00	0,00	0,00
	TED (%)	0,28	0,55	1,07	2,69	35,77	57,26
CP2	**TA (%)**	54,32	24,69	2,47	10,49	3,70	4,32
	TED (%)	6,81	12,58	13,61	13,62	14,60	33,73
GV1	**TA (%)**	92,86	7,14	0,00	0,00	0,00	0,00
	TED (%)	2,48	4,28	9,42	15,75	27,91	38,75
GV2	**TA (%)**	70,83	8,33	16,67	4,17	0,00	0,00
	TED (%)	7,39	5,18	5,46	16,49	34,76	28,82
MR1	**TA (%)**	55,15	19,70	8,79	6,67	6,97	2,73
	TED (%)	1,62	4,65	7,11	18,71	37,97	21,38
MR2	**TA (%)**	57,33	24,00	1,33	14,67	2,67	0,00
	TED (%)	2,49	16,79	9,78	20,74	28,88	19,43

Fonte: O autor.

Tabela 3.18: Resultados dos grupos de cada uma das supervisões/linha para as tangentes.

TANGENTES							
		Grupo 5	**Grupo 4**	**Grupo 3**	**Grupo 2**	**Grupo 1**	**Grupo 0**
CP1	**TA (%)**	74,07	22,22	1,85	1,85	0,00	0,00
	TED (%)	1,43	4,48	11,37	15,34	22,57	41,27
CP2	**TA (%)**	68,25	20,63	0,00	1,59	1,59	7,94
	TED (%)	7,01	14,46	12,82	13,61	16,22	31,23
GV1	**TA (%)**	50,00	22,73	13,64	9,09	4,55	0,00
	TED (%)	2,82	6,81	10,77	23,98	32,38	22,85
GV2	**TA (%)**	100	0,00	0,00	0,00	0,00	0,00
	TED (%)	1,91	2,85	6,16	21,57	42,06	25,19
MR1	**TA (%)**	66,67	33,33	0,00	0,00	0,00	0,00
	TED (%)	2,54	6,44	10,09	20,91	33,45	21,21
MR2	**TA (%)**	48,65	27,03	13,51	10,81	0,00	0,00
	TED (%)	6,69	16,55	12,83	29,81	19,34	10,66

Fonte: O autor.

É importante estar claro que, individualmente, o Grupo 5 não irá superar o resultado obtido pela técnica de votação por maioria ou mesmo para os classificadores individuais. Entretanto, pode ser que, junto a outro grupo, esse desempenho se aproxime dos outros resultados. Neste último caso, mesmo não alcançando o melhor resultado com relação à TA, a TED cai

significantemente nos Grupos 5 e 4, trazendo uma grande contribuição para o desempenho. Além disso, a aplicação dessa metodologia permitiu que os resultados fossem estratificados e, consequentemente, que fosse possível a criação de um sistema de prioridades de acordo com o compromisso entre o esforço empregado na inspeção e a taxa de acerto na localização dos dormentes danificados. Em outras palavras, agora é possível procurar por trechos da via permanente contendo dormentes defeituosos com uma maior probabilidade de localizá-los e, ao mesmo tempo, com uma menor chance de realizar um esforço desnecessário.

Para exemplificar, alguns dos resultados atuais serão comparados aos obtidos pela técnica de votação por maioria. Na configuração de CP/linha2/curvas, o resultado por voto majoritário foi de 81,48% de TA e 45,82% de TED. Isso significa que ao percorrer por todos os elementos dessa configuração, existe a possibilidade de localização dos defeitos de 81,48%, mas com taxa de 45,82% de esforço desnecessário que deverá ser empregado. Para a outra metodologia, o Grupo 5 foi responsável por um desempenho de 54,32% de TA e 6,81% de TED, enquanto que para o Grupo 4 foi de 24,69% de TA e 12,58% de TED. Juntos, alcançam um percentual de 79,51% de TA e 19,39% de TED que, se comparados aos resultados anteriores tem-se uma redução na TA de 1,97% mas, por outro lado, uma redução na TED de 25,92%.

Diferentes combinações de grupos também poderiam ser feitas nas outras configurações analisadas com o objetivo de encontrar a relação desejada entre TA e TED. Um outro exemplo de melhoria no desempenho foi na configuração de MR/linha1/tangentes. Para o caso do voto majoritário o desempenho foi de 100,00% na TA e 19,46% na TED, enquanto que para os grupos 4 e 5 juntos, o desempenho foi de 100,00% de TA e 8,98% de TED, igualando a taxa de localização dos defeitos, mas reduzindo o esforço empregado em 10,48%. Em alguns casos, os resultados obtidos tanto pelo voto majoritário quanto pelos diagnósticos individuais poderão ser melhores ou bem parecidos, mesmo com a combinação entre os grupos. De todo modo, isso não afeta o resultado da metodologia, que tem como principal funcionalidade a apresentação do diagnóstico de maneira estratificada.

Como esperado, é possível notar que a soma da TA dos 3 grupos mais significativos (5, 4 e 3) da técnica de grupos por votação é exatamente igual à TA do resultado pela votação por maioria. Por outro lado, a taxa de falso positivo é diluída entre os grupos de uma forma diferente, fazendo com que em vários casos ela fosse reduzida. Mesmo assim, o potencial dessa segunda estratégia vai além do que apenas tentar melhorar o desempenho, ela entrega a possibilidade da criação de um procedimento diferente, que de outra forma não seria possível.

3.2.3. Geração do Diagnóstico (*ensemble*)

Assim como apresentado na Seção 3.1.3, após a etapa de treinamento e validação, se faz possível a realização da etapa de inferência com os novos dados apresentados aos modelos. Dessa forma, a saída do diagnóstico será representada pelos intervalos de defeitos encontrados. Levando em consideração a metodologia do sistema *ensemble* utilizada nesta última etapa,

espera-se que os diagnósticos também sejam estratificados de acordo com os 6 grupos apresentados. Encontra-se na Tabela 3.19, um exemplo de apresentação do arquivo de saída gerado para indicar os trechos de defeito em cada grupo. A primeira coluna contém a indicação do elemento (tipo e número) e, na sequência, tem-se a EH, supervisão, linha, posição de início do elemento a partir do início da ferrovia, extensão do elemento, posição de início do defeito, posição de fim do defeito e o grupo indicado pelo *ensemble*.

Este segmento do arquivo gerado é apenas um exemplo de como poderia ser estruturado para a indicação dos trechos defeituosos em cada um dos grupos. O fato é que ele poderia ser configurado de outras formas, de acordo com a necessidade da empresa. Com o processo automatizado, a customização se torna ainda mais simples.

Tabela 3.19: Arquivo de diagnóstico gerado para inspeção em campo (*ensemble*).

ID	EH	SUP	Linha	Início elem. (m)	Extensão (m)	Início defeito (m)	Fim defeito (m)	Grupo
'TANGENTE 9'	'33/34'	'CP'	'Linha 1'	213419	419	228	229	Grupo 5
'TANGENTE 8'	'31/32'	'CP'	'Linha 1'	197715	285	37	48	Grupo 5
'TANGENTE 8'	'31/32'	'CP'	'Linha 1'	197715	285	50	55	Grupo 5
'TANGENTE 8'	'31/32'	'CP'	'Linha 1'	197715	285	109	109	Grupo 5
'TANGENTE 8'	'31/32'	'CP'	'Linha 1'	197715	285	115	124	Grupo 5
'TANGENTE 8'	'31/32'	'CP'	'Linha 1'	197715	285	130	141	Grupo 5
'TANGENTE 8'	'31/32'	'CP'	'Linha 1'	197715	285	143	152	Grupo 5
'TANGENTE 9'	'33/34'	'CP'	'Linha 1'	213419	419	50	50	Grupo 4
'TANGENTE 9'	'33/34'	'CP'	'Linha 1'	213419	419	123	124	Grupo 4
'TANGENTE 9'	'33/34'	'CP'	'Linha 1'	213419	419	175	175	Grupo 4
'TANGENTE 9'	'33/34'	'CP'	'Linha 1'	213419	419	179	179	Grupo 4
'TANGENTE 9'	'33/34'	'CP'	'Linha 1'	213419	419	184	186	Grupo 4
'TANGENTE 8'	'31/32'	'CP'	'Linha 1'	197715	285	49	50	Grupo 4
'TANGENTE 8'	'31/32'	'CP'	'Linha 1'	197715	285	57	60	Grupo 4
'TANGENTE 8'	'31/32'	'CP'	'Linha 1'	197715	285	71	73	Grupo 4
				...				
'TANGENTE 9'	'33/34'	'CP'	'Linha 1'	213419	419	1	2	Grupo 1
'TANGENTE 9'	'33/34'	'CP'	'Linha 1'	213419	419	4	7	Grupo 1
'TANGENTE 9'	'33/34'	'CP'	'Linha 1'	213419	419	20	26	Grupo 1

Fonte: O autor.

3.3. Discussão dos Resultados

A divisão deste trabalho em duas fases foi importante para a validação da metodologia utilizada, além de possibilitar a exploração de diferentes técnicas durante o desenvolvimento do projeto. A partir da execução das diversas ferramentas na primeira fase (Seção 3.1), foi possível aplicar análises comparativas com o objetivo de encontrar aquelas que mais se adequaram à resolução do problema. Nesse sentido, considerando o escopo da pesquisa, é possível afirmar

qual tamanho da janela de dados (128), qual técnica de extração de características (espacial) e qual técnica de seleção de característica (FDR) possui um maior potencial para desenvolver uma ferramenta capaz de detectar a presença de defeitos em dormentes de aço.

Após as primeiras conclusões alcançadas durante a primeira fase, buscou-se compreender o potencial de cada classificador utilizado durante a modelagem do problema (Seção 3.2.1). A partir dos resultados individuais foi possível identificar que, apesar do desempenho semelhante entre eles, alguns classificadores como a RNA, a SVM e o ADB se sobressaíram com relação ao GMM e ao HMM. Além disso, notou-se também que, de forma geral, o HMM apresentou os piores resultados, tanto para a TA quanto para a TED.

Com relação às estratégias de implementação do projeto na prática, algumas opções poderiam ser levadas em conta, considerando todas as abordagens apresentadas. Uma delas, e provavelmente a mais interessante, seria a utilização do *ensemble* de classificadores para a obtenção dos resultados estratificados e assim ter a possibilidade de um inspeção realizada por níveis de prioridades. Por outro lado, não sendo este o interesse da empresa, seria claramente possível a utilização de apenas um dos classificadores para um diagnóstico completo de todo trecho analisado, mas levando em consideração a TED inerente a cada modelo. Ao optar pela segunda opção, a RNA, a SVM ou o ADB seriam as melhores escolhas dentre as técnicas aqui estudadas.

Para a base de dados utilizada na pesquisa, ou seja, os sinais geométricos da via permanente, diversas análises também foram realizadas, principalmente no que diz respeito à combinação entre os 7 sinais escolhidos. Como resultado, algumas observações indicaram uma relação maior entre as curvas com os sinais de nivelamento, bitola e empeno, enquanto que para as tangentes, uma maior relação com os sinais de alinhamento, empeno e superelevação. De todo modo, a partir dos resultados encontrados, não foi possível inferir precisamente a respeito de quais sinais são melhores na identificação do problema como um todo. Ao tentar fixar uma única combinação de sinais tanto para as curvas quanto para as tangentes, os desempenhos oscilavam muito de uma supervisão/linha para outra. Portanto, para encontrar os melhores modelos, fez-se necessário utilizar combinações diferentes para cada configuração.

A partir da discussão dos resultados destacados acima, é possível configurar diferentes estratégias em busca do melhor modelo, porém, considerando a divisão do problema por configuração (supervisão, linha e tipo do elemento). Nesse sentido, é importante estar claro que a escolha pela metodologia utilizada na pesquisa exige a utilização de um modelo para cada configuração. Vale lembrar que essa foi uma escolha definida no início do projeto ao se observar que as variações significativas nas características do solo e do clima ao longo da ferrovia, impactam diretamente no comportamento dos sinais da geometria.

4. Conclusão

A logística do transporte de cargas se baseia fortemente nos sistemas ferroviários em muitos países. Nesse sentido, empregou-se neste trabalho métodos de aprendizado de máquina para a detecção de defeitos em dormentes de aço. Mais especificadamente, a hipótese inicial de que os sinais de geometria da via permanente (nivelamento longitudinal e transversal, alinhamento, empeno e bitola) poderiam ser utilizados para a localização dos trechos da ferrovia em que se encontram dormentes de aço danificados foi comprovada, e os resultados obtidos nos experimentos se mostraram promissores. No decorrer da pesquisa, verificou-se que ao tentar gerar um único modelo para diferentes supervisões, os desempenhos não ultrapassavam os 70% na taxa de localização dos dormentes danificados. O motivo do baixo desempenho deve-se principalmente às variações significativas nas características do solo e do clima ao longo da ferrovia, que impactam diretamente no comportamento dos sinais de geometria da via permanente. Dessa forma, verificou-se que uma melhor solução para este problema seria por meio da geração de um modelo para cada configuração definida por supervisão, linha e tipo do elemento (curvas ou tangentes).

Com a implantação da primeira fase da pesquisa, foi possível determinar a melhor configuração experimental comum para todas as configurações investigadas, que consistiu na utilização da janela de dados com 128 amostras, método de extração de características espacial e seleção de características com a razão discriminante de Fisher (FDR, do inglês, *Fisher's Discriminant Ratio*). Nestes casos, a Taxa de Acerto (TA) na localização dos dormentes danificados média variou entre 84% e 98% para as curvas e entre 81% e 99% para as tangentes, considerando uma Taxa de Esforço Desnecessário (TED) máxima limitada em 40%. Ao final, encontrou-se uma maior relação entre as curvas com os sinais de nivelamento, bitola e empeno, enquanto que para as tangentes, uma maior relação com os sinais de alinhamento, empeno e superelevação.

Para a segunda fase do experimento, valeram-se dos resultados obtidos na fase anterior, mas com uma nova base de dados. Dessa forma, foi possível validar o sistema desenvolvido desde a extração dos dados gerados pelo Carro Controle (CC) até a geração dos diagnósticos e, além disso, verificar que as técnicas escolhidas foram eficazes, por manterem desempenhos semelhantes à aqueles obtidos na primeira fase. A execução do treinamento individual para cada um dos 5 classificadores serviu tanto para a verificação dos desempenhos individuais quanto para a criação dos modelos que seriam combinados posteriormente. A nova metodologia proposta neste trabalho demonstrou eficácia na detecção de defeitos em dormentes de aço com Taxa de Acerto acima de 80% e Taxa de Falso Positivo abaixo de 40%, na maioria dos casos.

Com relação ao sistema *ensemble*, pode-se dizer que os resultados encontrados foram muito promissores devido à sua característica de apresentação. No método de grupos por votação ficou evidente a contribuição positiva causada pela estratificação dos resultados. A maneira com que esse método foi implementado possibilitou a criação de um sistema de pri-

oridades de acordo com o compromisso desejado entre o esforço empregado na inspeção e a taxa de acerto na localização dos dormentes danificados. Adicionalmente, a combinação entre os grupos foi capaz de superar o desempenho dos resultados individuais quanto a TA em alguns casos, além de apresentarem um pequeno percentual de TED nos grupos mais significativos.

4.1. Contribuições

As tecnologias para análise das condições estruturais de componentes ferroviários têm atraído muita atenção da academia nos últimos anos, proporcionando diversos estudos nesse seguimento. Entretanto, pouco se encontra na literatura quando se trata de análise das condições estruturais dos dormentes de aço. Portanto, a principal contribuição técnico científica desta pesquisa foi a implementação dos métodos de aprendizado de máquina para detecção de dormentes de aço defeituosos presentes nas ferrovias.

As estratégias utilizadas na pesquisa evidenciaram o potencial do sistema desenvolvido para a localização de trechos da estrada de ferro contendo dormentes de aço defeituosos. A partir dos resultados apresentados, fica evidente a contribuição industrial do trabalho para o aumento da confiabilidade da infraestrutura ferroviária, reduzindo custos de manutenção. Além disso, existe uma contribuição importante de saúde e segurança, uma vez que possibilita a redução significativa das horas de exposição dos inspetores aos riscos físicos e condições climáticas intensas. Em um paralelo com a condição atual, em que as inspeções são realizadas visualmente por toda a ferrovia, e com base nos resultados obtidos, evidenciou-se a possibilidade da localização de grande parte dos defeitos sem a necessidade de percorrer todo o percurso. Mais do que isso, os últimos resultados apontaram para um caminho em que é possível realizar a estratificação dos resultados com o objetivo de aumentar a acurácia do sistema e, ao mesmo tempo, levando em consideração o compromisso desejado entre o esforço empregado na inspeção e a taxa de acerto na localização dos dormentes de aço danificados.

Com relação às contribuições para a academia, foi desenvolvido um artigo com os resultados iniciais da pesquisa para o VI Encontro ANTF de Ferrovias (SILVA *et al.*, 2021) e um outro artigo foi submetido para o Congresso Brasileiro de Automática (CBA) com parte dos resultados obtidos nesta pesquisa. Além disso, encontra-se em desenvolvimento um artigo que será enviado para o IEEE *Transactions on Intelligent Transportation Systems* (T-ITS) contemplando todos os resultados encontrados ao longo do desenvolvimento do trabalho.

5. Recomendações para Trabalhos Futuros

A metodologia aqui utilizada, em que foram aplicadas diferentes técnicas em cada etapa do sistema, permitiu encontrar os métodos responsáveis pelos melhores resultados até o momento. Entretanto, várias outras técnicas existentes na literatura poderiam ser utilizadas para melhorar ainda mais o desempenho do sistema, como por exemplo, outras técnicas de seleção de características, investigação mais ampla das diversas famílias Wavelet que ainda não foram testadas, implementação de novos classificadores como o *Random Forest*, *XGBoost*, ou mesmo uma aplicação de *deep learning*. Além disso, ainda é possível aprofundar a análise sobre as técnicas utilizadas, buscando investigar, por exemplo, outras variações na complexidade dos classificadores e na quantidade de parâmetros do método de seleção de características ou mesmo novas técnicas de extração de características. Adicionalmente, novas estratégias poderiam ser empregadas na concepção dos *ensembles*. Uma outra possibilidade seria uma abordagem ao problema baseado em técnicas de detecção de anomalias, em que se modela apenas a classe dos dados normais, visto que o conjunto de dados é bastante desbalanceada.

Como mencionado ao longo do texto, a estratégia de modelagem utilizada na pesquisa baseou-se na divisão dos modelos por supervisão, linha e tipo do elemento, justificado pelo fato de que as variações significativas nas características do solo e do clima ao longo da ferrovia, impactam diretamente no comportamento dos sinais da geometria. Entretanto, considerando que os modelos ainda fossem gerados para cada supervisão, seria possível testar a possibilidade de se juntar os elementos de linhas diferentes (1 e 2) e ainda de tipos diferentes (curvas e tangentes), reduzindo significantemente o número de modelos. Com relação à seleção dos dados de treinamento para a geração dos modelos, é possível utilizar uma estratégia de sobreamostragem dos dados de defeito, considerando-se que neste trabalho foi utilizado um processo de subamostragem.

Com relação ao sistema *ensemble* desenvolvido, espera-se verificar em campo alguma relação existente entre as classes estratificadas e o nível de criticidade do defeito diagnosticado. Por exemplo, verificar se os trechos diagnosticados pelo resultado do Grupo 5 apresentam uma maior criticidade em relação aos grupos menos significativos. Se for comprovada essa relação, seria possível atuar de maneira mais eficiente, alocando as horas de trabalho por parte da equipe de manutenção para aquelas regiões mais críticas.

Referências Bibliográficas

ABNT, A. B. D. N. T. "NBR 16387: Via férrea - Classificação de vias." 2020.

AMATO, F., LÓPEZ, A., PEÑA-MÉNDEZ, E. M., et al.. "Artificial neural networks in medical diagnosis". 2013.

ANTT. *Fiscalização do Transporte Ferroviário de Cargas*, 2a edição ed., abril 2018.

BAHL, L. R., JELINEK, F., MERCER, R. L. "A Maximum Likelihood Approach to Continuous Speech Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. PAMI-5, n. 2, pp. 179–190, mar 1983.

BILMES, J. A., OTHERS. "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models", *International Computer Science Institute*, v. 4, n. 510, pp. 126, 1998.

BISHOP, C. M., OTHERS. *Neural networks for pattern recognition*. Oxford university press, 1995.

BOSER, B. E., GUYON, I. M., VAPNIK, V. N. "A training algorithm for optimal margin classifiers". ACM Press, 1992.

BRASIL, G. D. "EFVM - Estrada de Ferro Vitória a Minas Gerais". 2020. Disponível em: <https://www.ppi.gov.br/efvm-estrada-de-ferro-vitoria-a-minas#:~:text=A%20Estrada%20de%20Ferro%20Vit%C3%B3ria,o%20vale%20do%20Rio%20Doce.>.

BREIMAN, L. "Bagging predictors", *Machine learning*, v. 24, n. 2, pp. 123–140, 1996.

BRINA, H. L. *Estradas de Ferro - Via Permanente*. Editora UFMG, 1988.

CHAPMAN, P., CLINTON, J., KERBER, R., et al.. *CRISP-DM 1.0*, 2000.

CLARK, A., KAEWUNRUEN, S., JANELIUKSTIS, R., et al.. "Damage Detection in Railway Prestressed Concrete Sleepers using Acoustic Emission", *IOP Conference Series: Materials Science and Engineering*, 2017.

COIMBRA, M. *Modos de Falha dos Componentes da Via Permanente Ferroviária e Seus Efeitos no Meio Ambiente*. Tese de Mestrado, Instituto Militar de Engenharia, 2008.

COX, P. H. *Análise e síntese de um processador digital wavelet*. Tese de Doutorado, 2004.

DELFOROUZI, A., TABATABAEI, A. H., KHAN, M. H., et al.. "A vision-based method for automatic crack detection in railway sleepers". Em: *International Conference on Computer Recognition Systems*, pp. 130–139. Springer, 2017.

DINIZ, P. S. R., DA SILVA, E. A. B., NETTO, S. L. *Processamento Digital de Sinais*. 2014.

FERDOUS, W., MANALO, A. "Failures of mainline railway sleepers and suggested remedies–review of current practice", *Engineering Failure Analysis*, v. 44, pp. 17–35, 2014.

FRANCA, A. S., VASSALLO, R. F. "A method of classifying railway sleepers and surface defects in real environment", *IEEE Sensors Journal*, v. 21, n. 10, pp. 11301–11309, 2020.

FREUND, Y., SCHAPIRE, R. E. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", *Journal of Computer and System Sciences*, v. 55, n. 1, pp. 119–139, aug 1997.

GOODFELLOW, I., BENGIO, Y., COURVILLE, A. *Deep Learning*. MIT Press, 2016.

HAYKIN, S. *Redes Neurais: Princípios e Prática*. Artmed, 2007.

JAIN, A., DUIN, R., MAO, J. "Statistical Pattern Recognition: A Review", *IEEE Trans. Pattern Anal. Mach. Intell.*, v. 22, pp. 4–37, 01 2000.

JAIN, A. K., CHANDRASEKARAN, B. "39 Dimensionality and sample size considerations in pattern recognition practice", *Handbook of statistics*, v. 2, pp. 835–855, 1982.

KAEWUNRUEN, S., REMENNIKOV, A. M. "Dynamic effect on vibration signatures of cracks in railway prestressed concrete sleepers". Em: *Advanced Materials Research*, v. 41, pp. 233–239. Trans Tech Publ, 2008.

KAEWUNRUEN, S., YOU, R., ISHIDA, M. "Composites for timber-replacement bearers in railway switches and crossings", *Infrastructures*, v. 2, n. 4, pp. 13, 2017.

KERBY, D. "The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation", *Comprehensive Psychology*, v. 3, jan 2014.

KHAN, H. U. R., SIDDIQUE, M., ZAMAN, K., et al.. "The impact of air transportation, railways transportation, and port container traffic on energy demand, customs duty,

and economic growth: Evidence from a panel of low-, middle-, and high-income countries", *Journal of Air Transport Management*, v. 70, pp. 18–35, 2018.

KOUROUSSIS, G., CAUCHETEUR, C., KINET, D., et al.. "Review of Trackside Monitoring Solutions: From Strain Gages to Optical Fibre Sensors", *Sensors*, v. 15, pp. 20115–20139, 08 2015.

KRUMMENACHER, G., ONG, C. S., KOLLER, S., et al.. "Wheel defect detection with machine learning", *IEEE Transactions on Intelligent Transportation Systems*, v. 19, n. 4, pp. 1176–1187, 2017.

LACERDA, A. L. M., FILHO, P., BRITO, J., et al.. "Detection of faults in three phase induction motors using wavelet packet analysis". 01 2011.

MALLAT, S. *A Wavelet Tour of Signal Processing*. Cambridge: Academic Press, 1998.

MARTINS, G. D. A. *Estatística geral e aplicada*. Atlas, 2001.

MATTERA, D., HAYKIN, S. "Support Vector Machines for Dynamic Reconstruction of a Chaotic System". Em: *Advances in Kernel Methods: Support Vector Learning*, p. 211–241, Cambridge, MA, USA, MIT Press, 1999. ISBN: 0262194163.

MESQUITA, A. L., SANTIAGO, D. F., BEZERRA, R. A., et al.. "Detecção De Falhas Em Rolamentos Usando Transformadas Tempo-Frequência–Comparação Com Análise De Envelope." *Mecánica Computacional*, , n. 1, pp. 1938–1954, 2002.

MORAIS, E. C. *Reconhecimento de Padrões e Redes Neurais Artificiais em Predição de Estruturas Secundárias de Proteínas*. Tese de Doutorado, Universidade Federaldo Rio de Janeiro - UFRJ, 2010.

NG, A. K., MARTUA, L., SUN, G. "Influence of Sleeper Distance on Rail Corrugation Growth". Em: *2018 International Conference on Intelligent Rail Transportation (ICIRT)*, pp. 1–5. IEEE, 2018.

NG, A. K., MARTUA, L., SUN, G. "Dynamic modelling and acceleration signal analysis of rail surface defects for enhanced rail condition monitoring and diagnosis". Em: *2019 4th International Conference on Intelligent Transportation Engineering (ICITE)*, pp. 69–73. IEEE, 2019.

OLIVEIRA, H. *Análise de Sinais para Engenheiros: Uma abordagem via WAVELETS*. 01 2007.

OPPENHEIM, A., WILLSKY, A., NAWAB, S. *Sinais e sistemas*. Prentice-Hall, 2010.

PAPOULIS, A. *Probability, Random Variables and Stochastic Processes*. 1991.

PLASSER. “Carro Controle eletrônico auto propelido EM80H”. 2022. Disponível em: <https://www.plasser.com.br/pt/maquinas-sistemas/em80h.html>.

PORTELA, N. M. *Modelo de Mistura de Gaussianas Fuzzy Contextual*. Tese de Mestrado, Universidade Federal de Pernambuco (UFPE), 2015.

RABINER, L. “A tutorial on hidden Markov models and selected applications in speech recognition”, *Proceedings of the IEEE*, v. 77, n. 2, pp. 257–286, 1989.

SANTIAGO, D. F. D. A., PEDERIVA, R. “Influência Da Resolução Tempo-Frequência Da Wavelet De Morlet No Diagnóstico De Falhas De Máquinas Rotativas.” *Mecánica Computacional*, pp. 2538–2550, 2003.

SCHAPIRE, R. E. “The strength of weak learnability”, *Machine learning*, v. 5, n. 2, pp. 197–227, 1990.

SHARMA, L., YADAV, D. K., SINGH, A. “Fisher’s linear discriminant ratio based threshold for moving human detection in thermal video”, *Infrared Physics & Technology*, v. 78, pp. 118–128, 2016.

SILVA, L. P. F., LIMA, G. H. S., LAURETT, N. S., et al.. “Identificação Automática de Trechos da Estrada de Ferro Vitória-Minas com Dormentes Danificados”, *VI Encontro ANTF de Ferrovias*, 2021.

SMITH, L. I. “A Tutorial on Principal Components Analysis”, p. 27, 2002.

STEWART, J. *Cálculo (Tradução da 7ª edição norte-americana)*, v. 1. São Paulo: Cengage Learning, 2013.

THEODORIDIS, S., KOUTROUMBAS, K. *Pattern Recognition & Matlab Intro*. 4th ed. USA, Academic Press, Inc., 2010.

VALE. “Dormentes de aço substituem os de madeira e ajudam a preservar o meio ambiente”. março 2013. Disponível em: <http://www.vale.com/brasil/pt/aboutvale/news/paginas/dormentes-de-aco-ajudam-a-preservar-o-meio-ambiente.aspx>.

VAPNIK, V. N. *The Nature of Statistical Learning Theory*. Springer, November 1999.

VLI. “Estrada de Ferro Vitória Minas - EFVM (concessão Vale)”. 2017. Disponível em: <https://www.vli-logistica.com.br/conheca-a-vli/ferrovias/efvm-concessao-vale/>.

WEBB, A. R. *Statistical pattern recognition*. John Wiley & Sons, 2003.

WOLPERT, D. H. “Stacked generalization”, *Neural networks*, v. 5, n. 2, pp. 241–259, 1992.

YARED, G. F. G., BARBOSA, C. H. N. R., LEITE, S. N. C., et al.. "Vibration Analysis for Crack Detection in Railroad Steel Sleepers", *International Heavy Haul STS Conference (IHHA 2019)*, 2019.

YELLA, S., DOUGHERTY, M., K., G. N. "Condition monitoring of wooden railway sleepers", *Transportation Research*, 2009.

ZHAO, J., CHAN, A., BURROW, M. "Reliability analysis and maintenance decision for railway sleepers using track condition information", *Journal of the Operational Research Society*, v. 58, n. 8, pp. 1047–1055, 2007.

Apêndice A: Ambiente IHM

A portabilidade e a capacidade de replicação do diagnóstico, com a possível inclusão de novas supervisões a partir da obtenção de novos modelos, motivaram o desenvolvimento de rotinas no ambiente do *Matlab* com o intuito específico de servir como uma interface para usuários do sistema. Essa interface foi exportada para uma aplicação *standalone* que independe do ambiente no qual os códigos foram desenvolvidos, necessitando apenas da instalação de uma plataforma de execução (*Matlab Runtime*) gratuita, disponibilizada pela *Mathworks*, para executar o programa.

Uma vez iniciada a execução, tal aplicativo abre um terminal do sistema operacional no qual está operando, onde serão emitidos avisos com relação à execução e informações complementares. A primeira tela apresentada ao usuário permite-o escolher estre as opções: Extrair Dados, Extrair Rótulos, Parametrizar Dados, Treinar um Classificador ou Realizar Diagnóstico, conforme mostrado na Figura 1. Após a seleção de uma delas, outras opções serão apresentadas ao usuário de acordo com cada uma escolha.

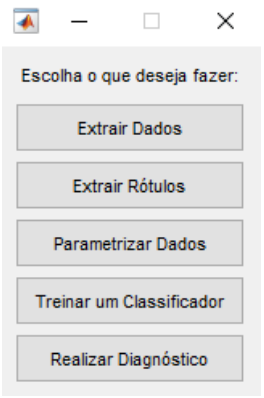


Figura 1: Opções de escolha para o usuário.
Fonte: Próprio autor.

Extrair Dados

Caso a opção seja para extrair os dados, aparecerá uma nova janela solicitando ao usuário que selecione o caminho para arquivo contendo a relação de todos os elementos listados da EFVM. Feito isso, o usuário precisará escolher o ano da inspeção e o caminho para o arquivo contendo a base de dados do CC para uma determinada inspeção realizada.

Extrair Rótulos

Se a opção for para extrair os rótulos, aparecerá uma nova janela solicitando ao usuário que selecione o caminho para arquivo contendo todos os elementos da EFVM rotulados. Feito isso, o usuário precisará escolher apenas o ano da inspeção analisada.

Parametrizar dados

Se o usuário desejar parametrizar os dados, a única informação necessária será a do ano da inspeção analisada.

Treinar um Classificador

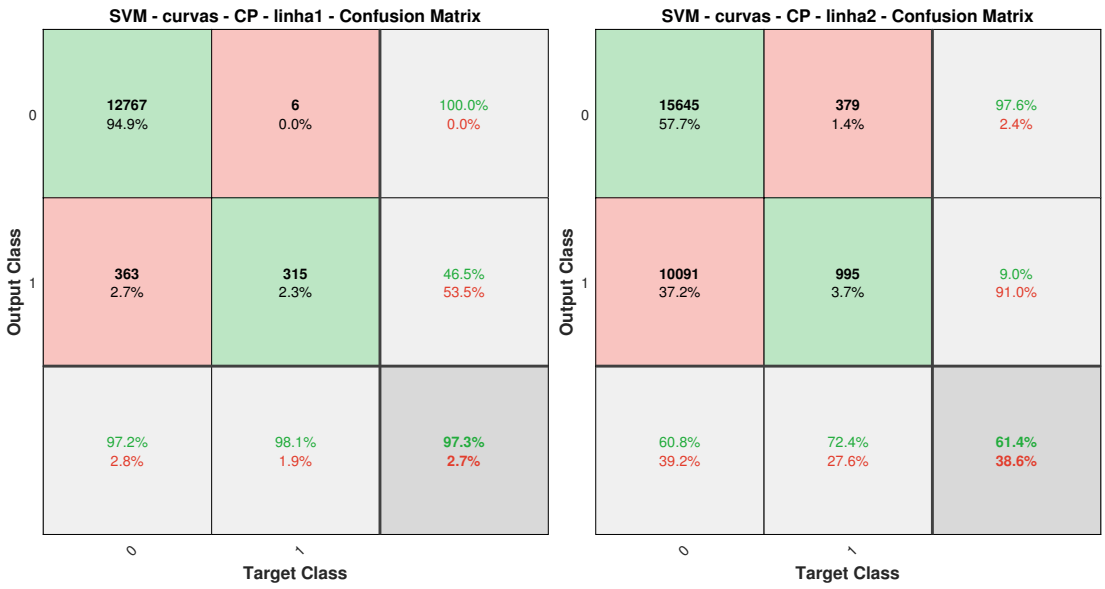
Se a opção for para treinar um classificador, aparecerá uma nova janela solicitando ao usuário que defina as três opções: o ano da inspeção, a supervisão e a linha. Feito isso, o usuário precisará escolher entre o tipo de elemento que deseja analisar (curvas ou tangentes). Em seguida, deve-se escolher qual classificador será utilizado para o treinamento entre as 5 opções indicadas.

Realizar Diagnóstico

Caso a opção seja para realizar um diagnóstico, aparecerá uma nova janela solicitando ao usuário que escolha o método individual (neste caso aparecerão as opções dos 5 classificadores utilizados) ou utilizando o *ensemble* (neste caso aparecerá a opção de votação por maioria e grupos por votação). Feito isso, o usuário precisará escolher entre o tipo de diagnóstico que deseja emitir (Figuras ou planilhas). Em seguida, deve-se escolher a supervisão e a linha para qual deseja-se realizar o diagnóstico. Depois, o usuário precisará escolher entre o tipo de elemento (curvas ou tangentes). Por fim, deve-se escolher o ano dos modelos que serão carregados, o ano da inspeção que se pretende diagnosticar (realizar a inferência) e o número da inspeção.

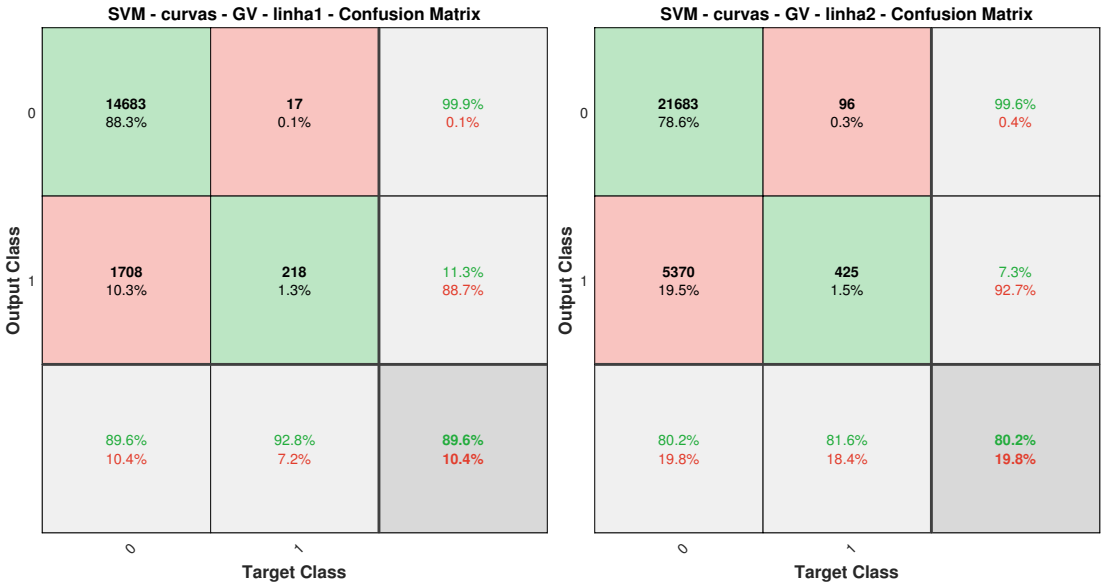
Apêndice B: Matriz de Confusão

Na seção 3.2, foram apresentados os resultados da segunda fase do experimento, contemplando o desempenho individual de cada classificador. Uma das formas de análise de desempenho foi por meio da MC que, naquele caso, foram apresentadas somente para o classificador RNA. Entretanto, também foram geradas as outras matrizes de confusão para os outros classificadores analisados. Portanto, seguem-se as matrizes de confusão para cada configuração analisada, considerando os outros quatro classificadores respectivamente: SVM, ADB, GMM e HMM.



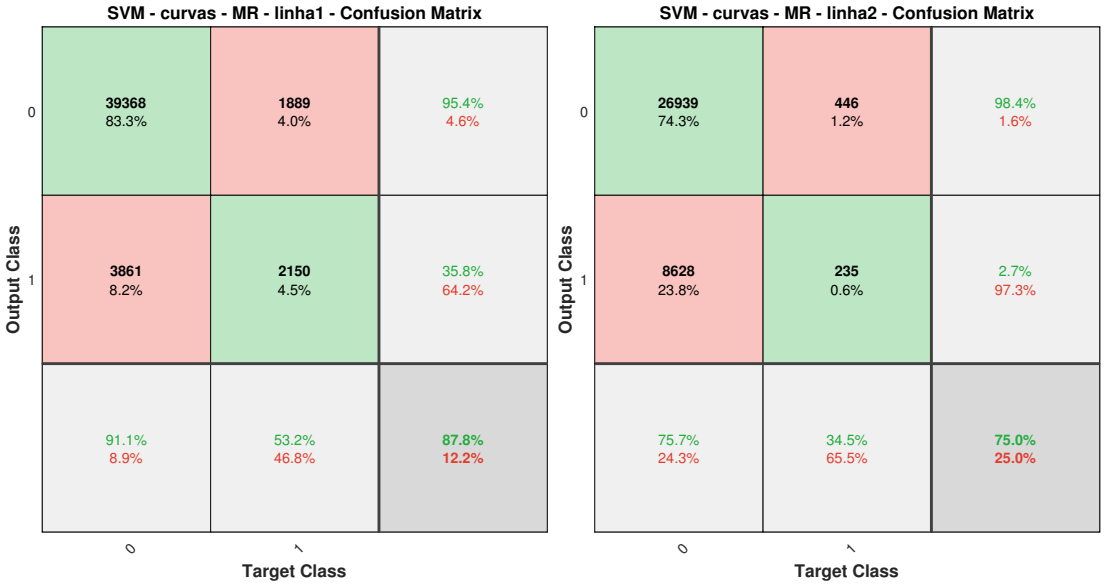
(a) CP/Linha1. (b) CP/Linha2.

Figura 2: Matriz de Confusão para a SVM (Curvas - Conselheiro Pena).
Fonte: O autor.



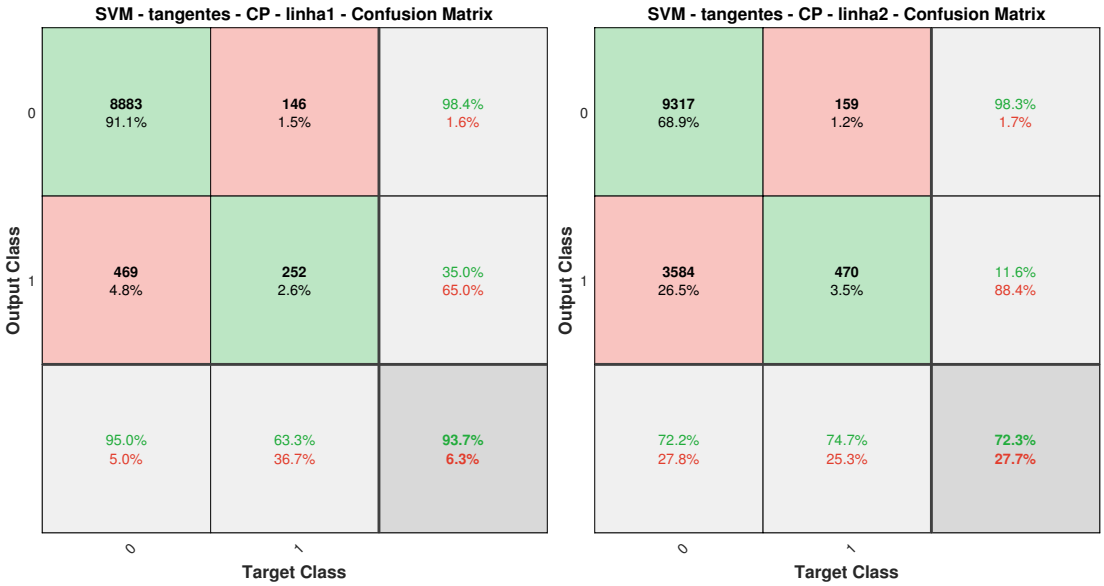
(a) GV/Linha1. (b) GV/Linha2.

Figura 3: Matriz de Confusão para a SVM (Curvas - Governador Valadares).
Fonte: O autor.



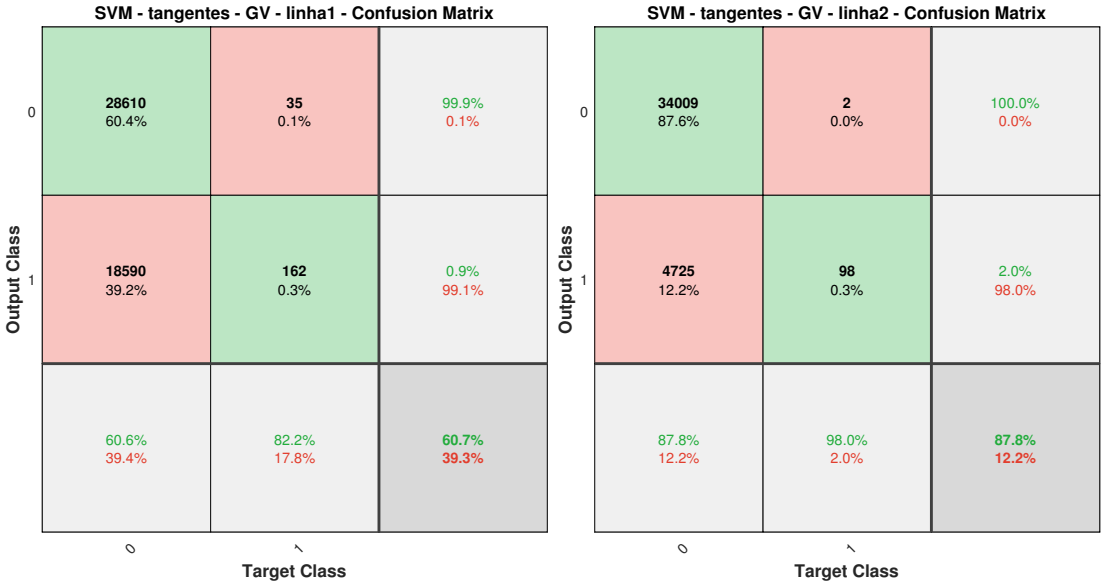
(a) MR/Linha1. (b) MR/Linha2.

Figura 4: Matriz de Confusão para a SVM (Curvas - Mário Carvalho).
Fonte: O autor.



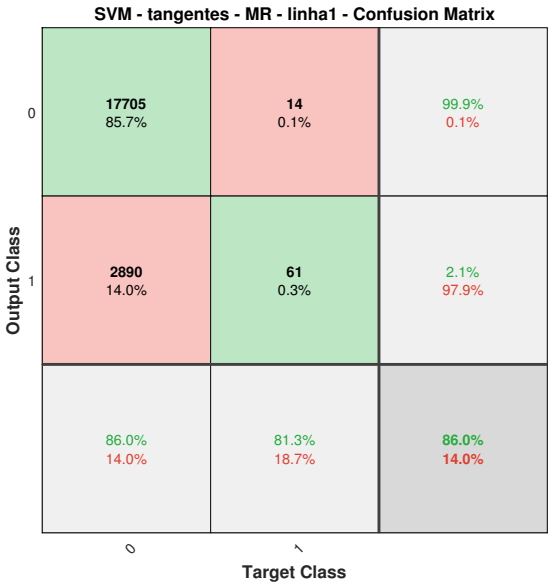
(a) CP/Linha1. (b) CP/Linha2.

Figura 5: Matriz de Confusão para a SVM (Tangentes - Conselheiro Pena).
Fonte: O autor.

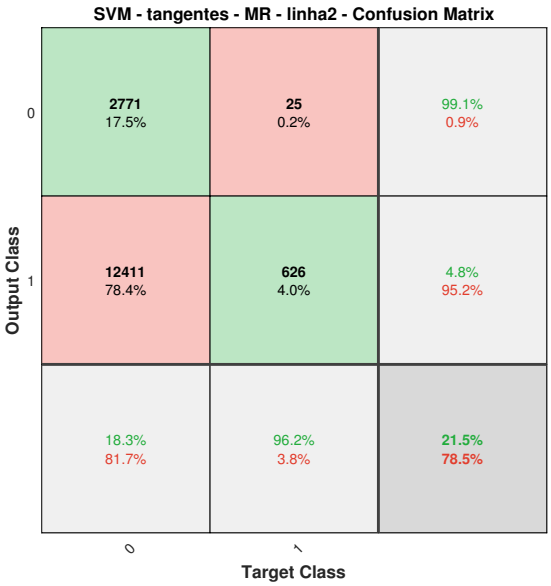


(a) GV/Linha1. (b) GV/Linha2.

Figura 6: Matriz de Confusão para a SVM (Tangentes - Governador Valadares).
Fonte: O autor.

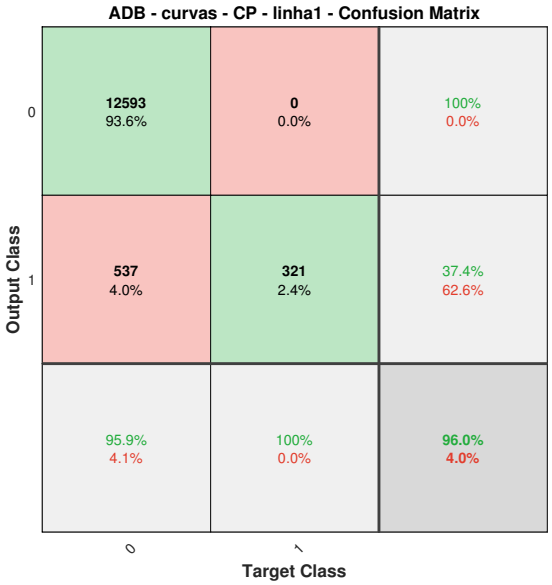


(a) MR/Linha1.

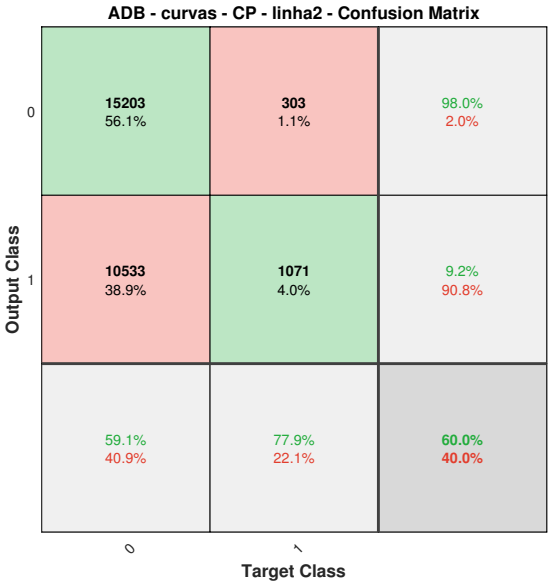


(b) MR/Linha2.

Figura 7: Matriz de Confusão para a SVM (Tangentes - Mário Carvalho).
Fonte: O autor.

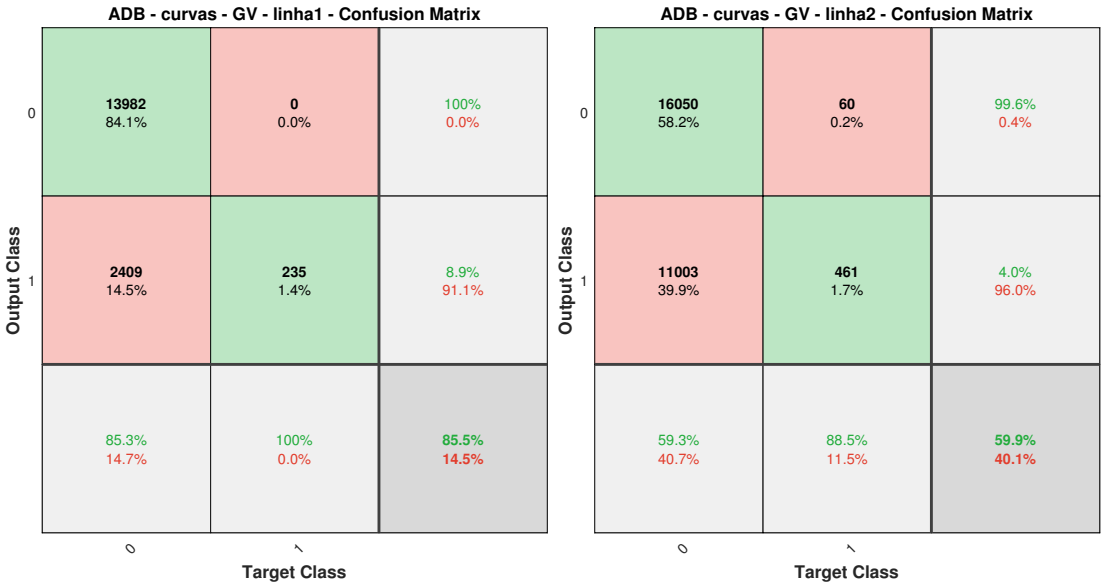


(a) CP/Linha1.



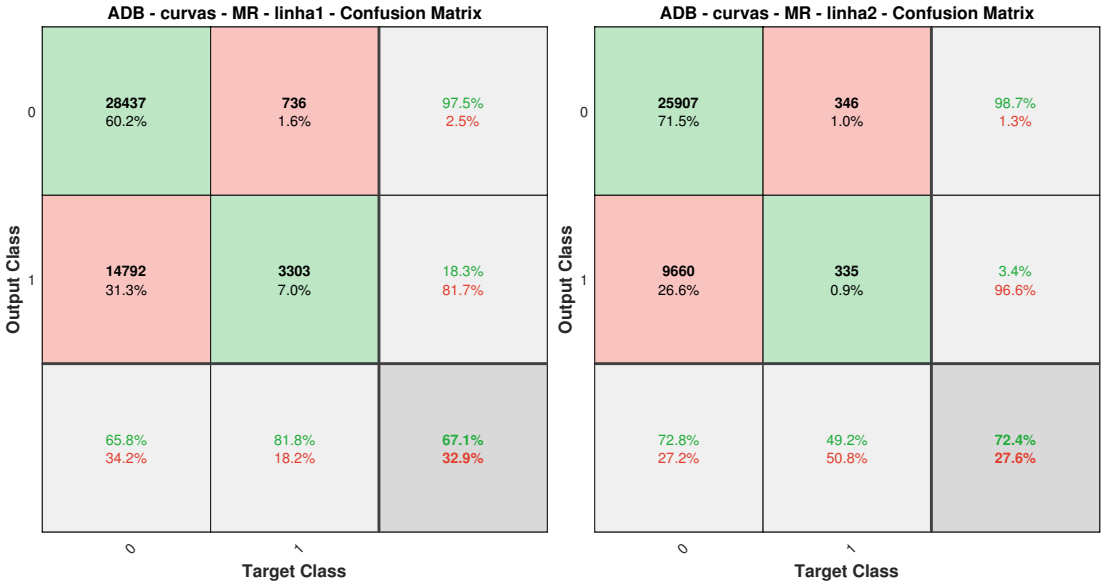
(b) CP/Linha2.

Figura 8: Matriz de Confusão para a ADB (Curvas - Conselheiro Pena).
Fonte: O autor.



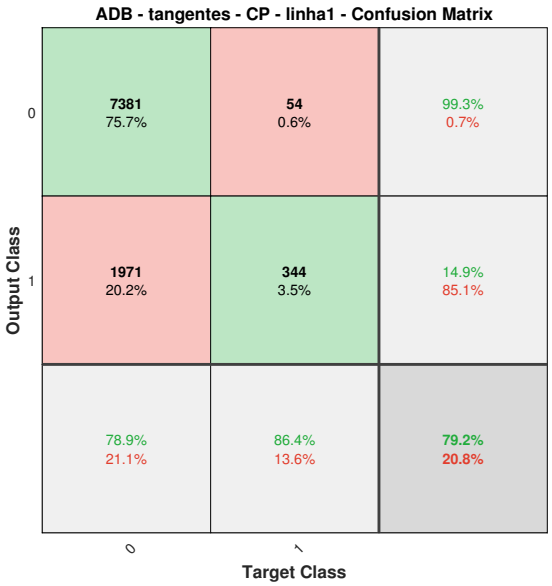
(a) GV/Linha1. (b) GV/Linha2.

Figura 9: Matriz de Confusão para a ADB (Curvas - Governador Valadares).
Fonte: O autor.

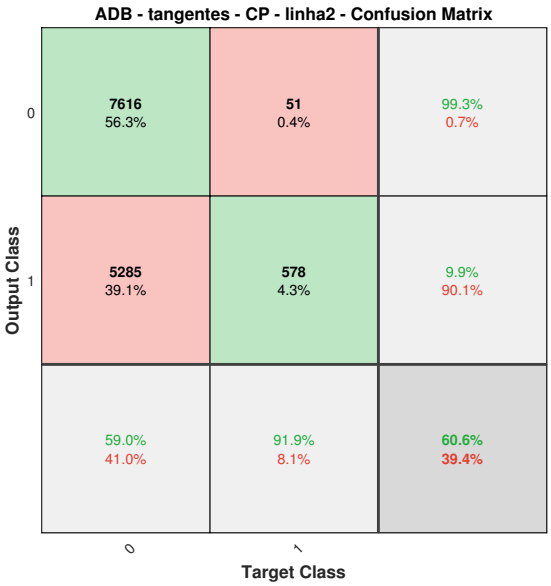


(a) MR/Linha1. (b) MR/Linha2.

Figura 10: Matriz de Confusão para a ADB (Curvas - Mário Carvalho).
Fonte: O autor.

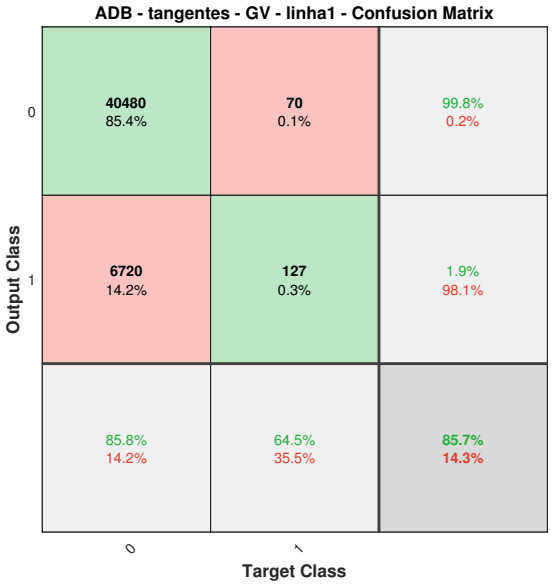


(a) CP/Linha1.

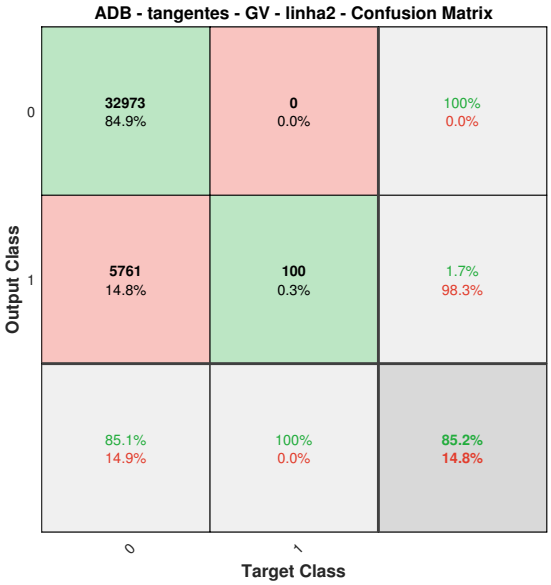


(b) CP/Linha2.

Figura 11: Matriz de Confusão para a ADB (Tangentes - Conselheiro Pena).
Fonte: O autor.

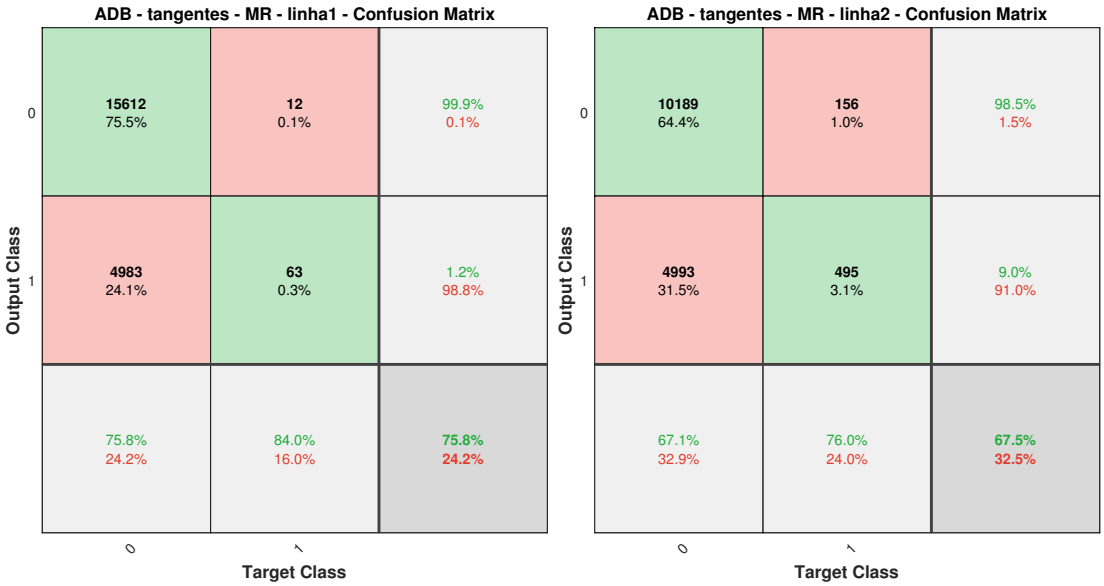


(a) GV/Linha1.



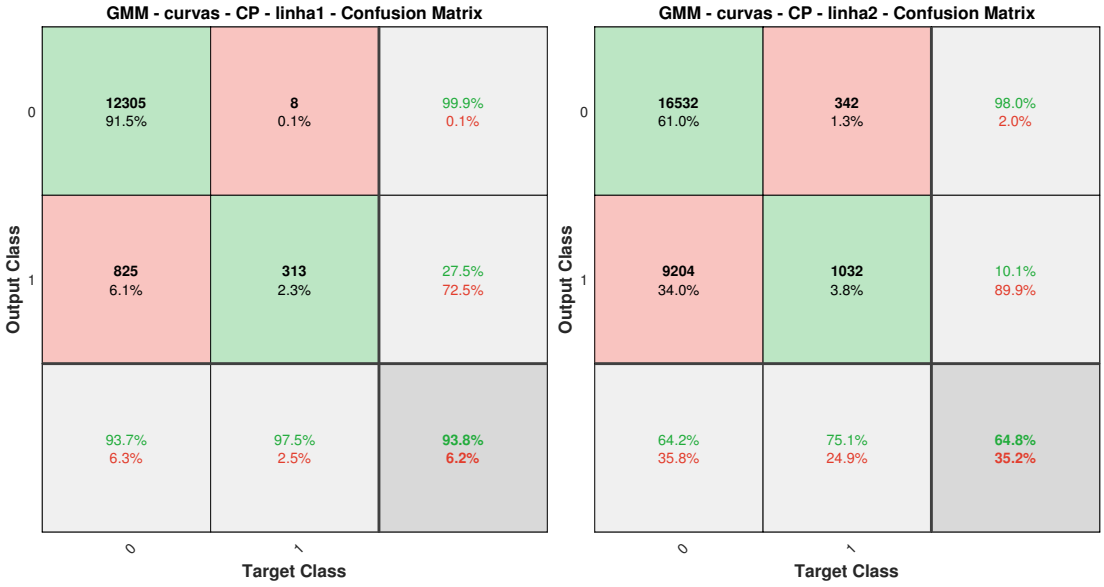
(b) GV/Linha2.

Figura 12: Matriz de Confusão para a ADB (Tangentes - Governador Valadares).
Fonte: O autor.



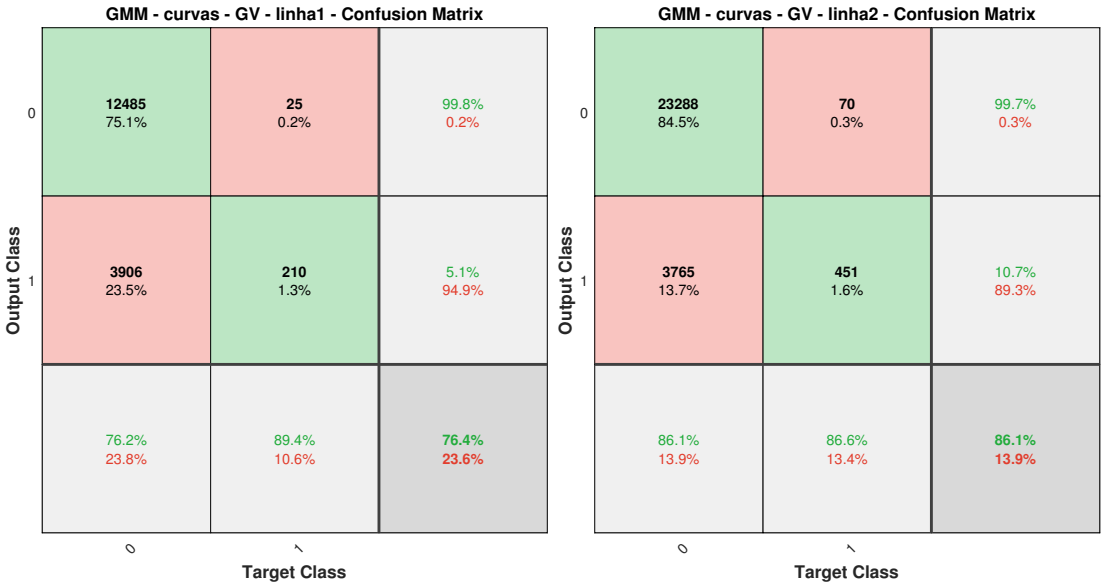
(a) MR/Linha1. (b) MR/Linha2.

Figura 13: Matriz de Confusão para a ADB (Tangentes - Mário Carvalho).
Fonte: O autor.



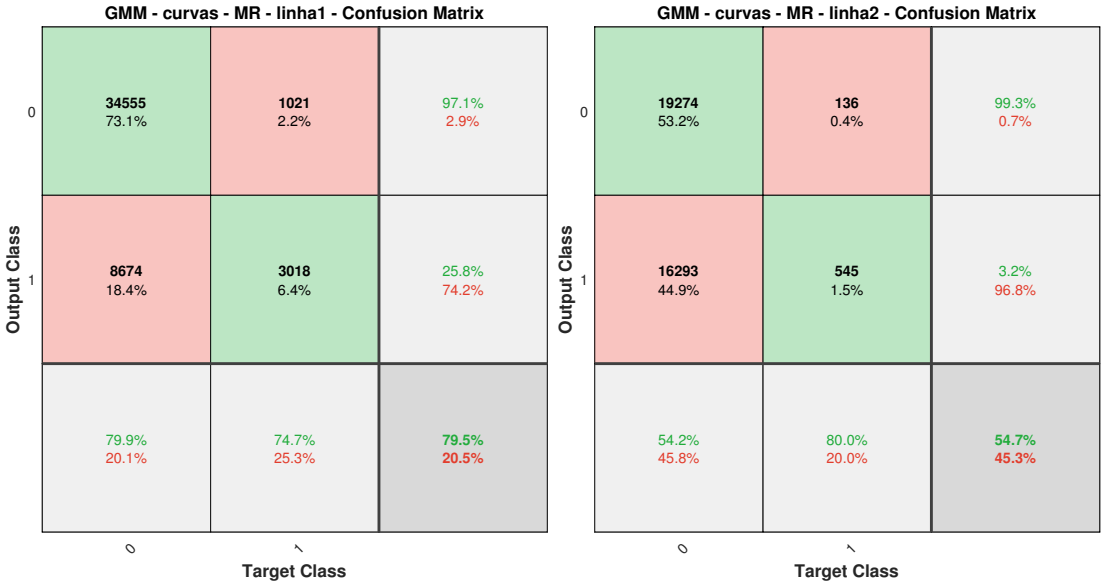
(a) CP/Linha1. (b) CP/Linha2.

Figura 14: Matriz de Confusão para a GMM (Curvas - Conselheiro Pena).
Fonte: O autor.



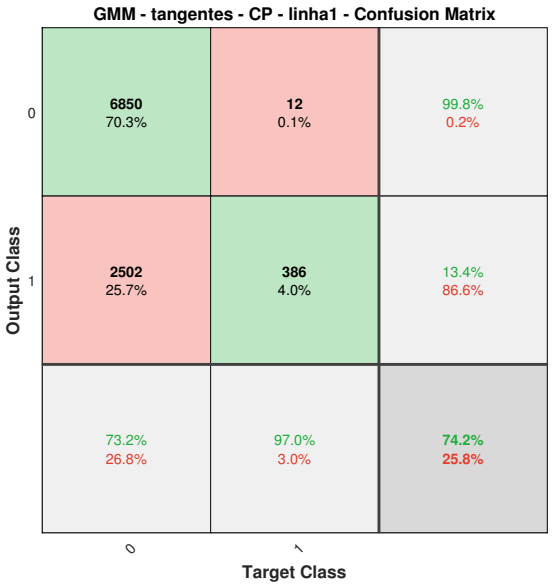
(a) GV/Linha1. (b) GV/Linha2.

Figura 15: Matriz de Confusão para a GMM (Curvas - Governador Valadares).
Fonte: O autor.

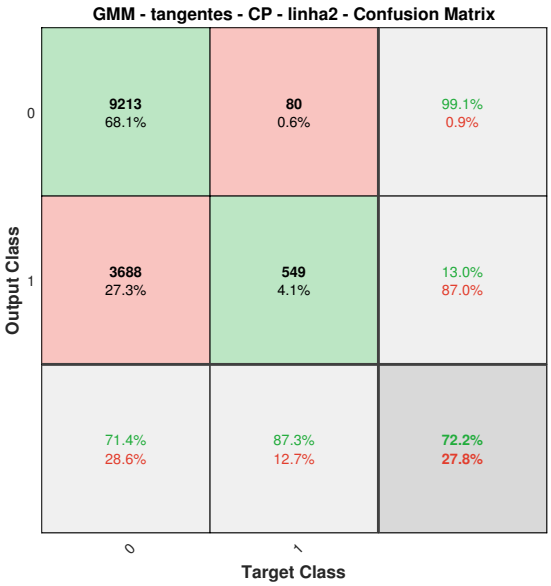


(a) MR/Linha1. (b) MR/Linha2.

Figura 16: Matriz de Confusão para a GMM (Curvas - Mário Carvalho).
Fonte: O autor.

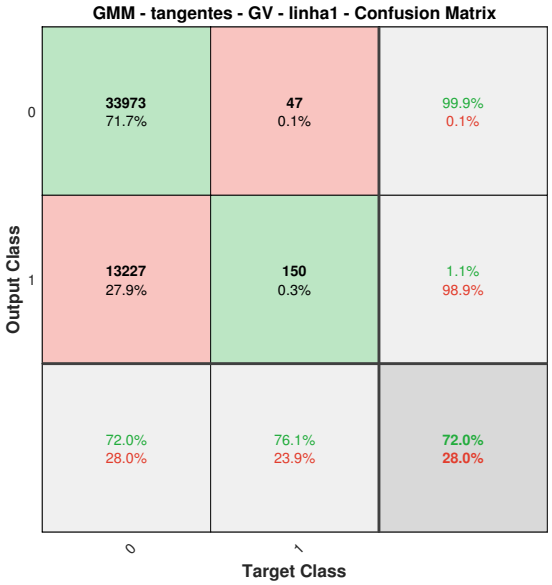


(a) CP/Linha1.

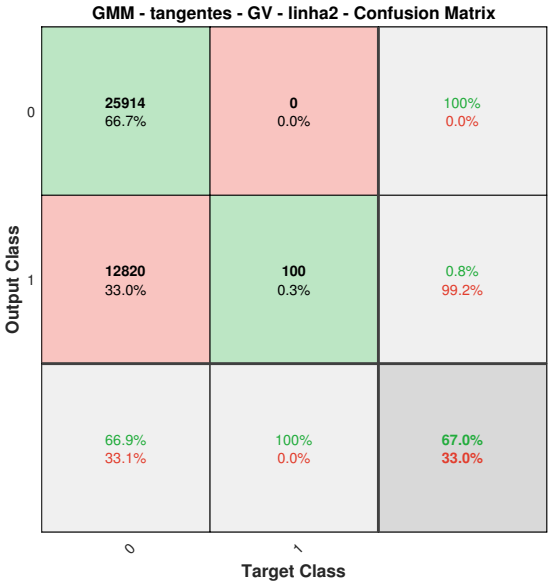


(b) CP/Linha2.

Figura 17: Matriz de Confusão para a GMM (Tangentes - Conselheiro Pena).
Fonte: O autor.

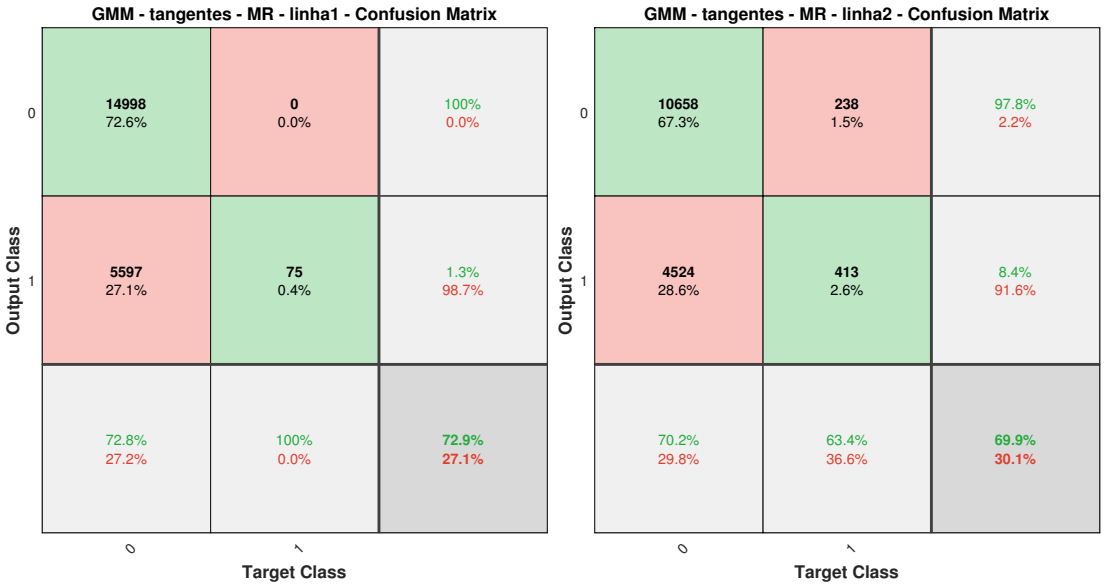


(a) GV/Linha1.



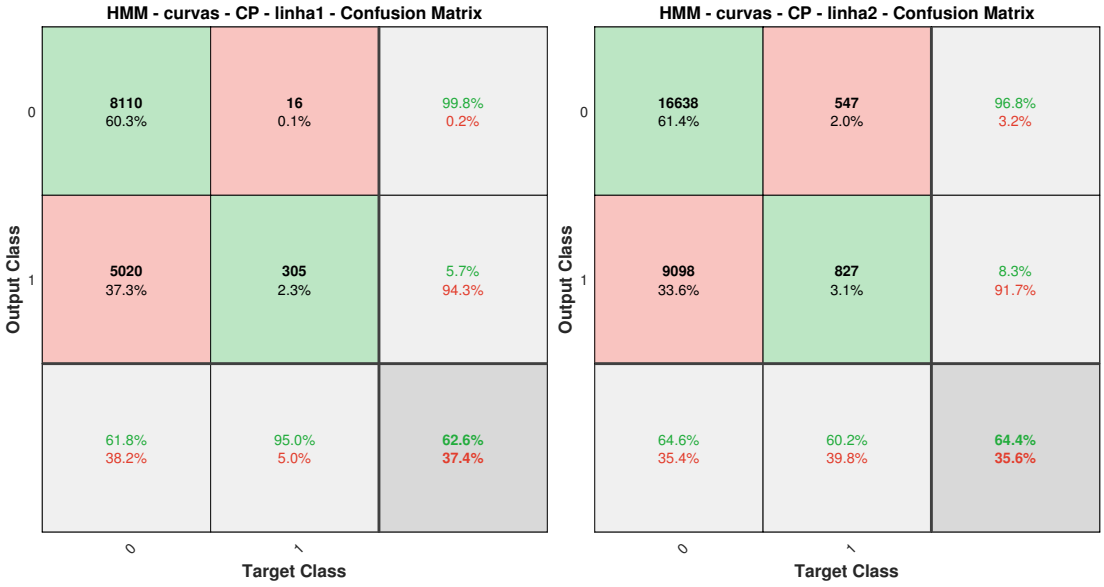
(b) GV/Linha2.

Figura 18: Matriz de Confusão para a GMM (Tangentes - Governador Valadares).
Fonte: O autor.



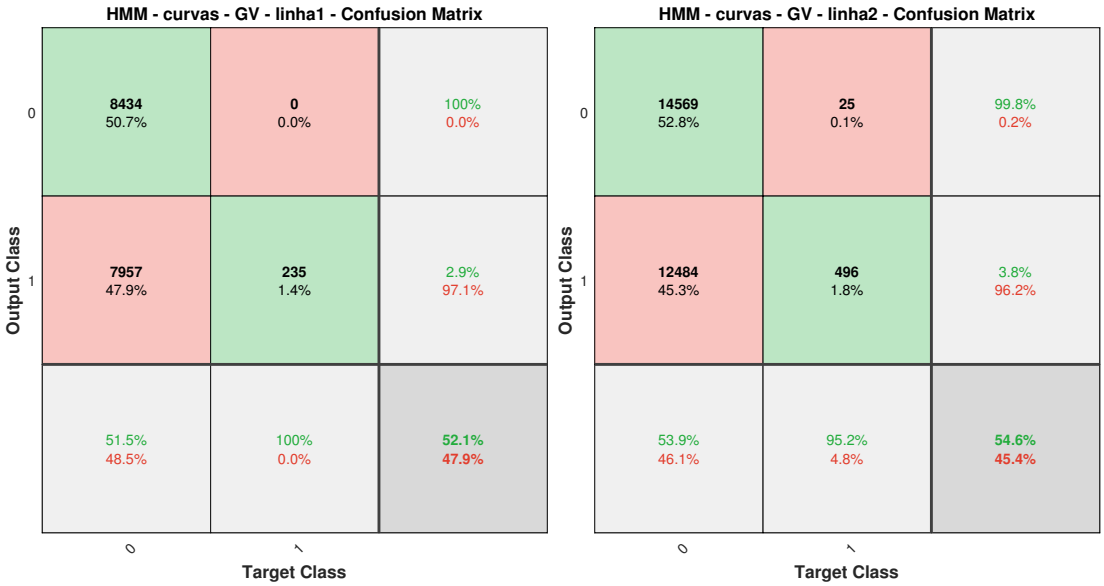
(a) MR/Linha1. (b) MR/Linha2.

Figura 19: Matriz de Confusão para a GMM (Tangentes - Mário Carvalho).
Fonte: O autor.



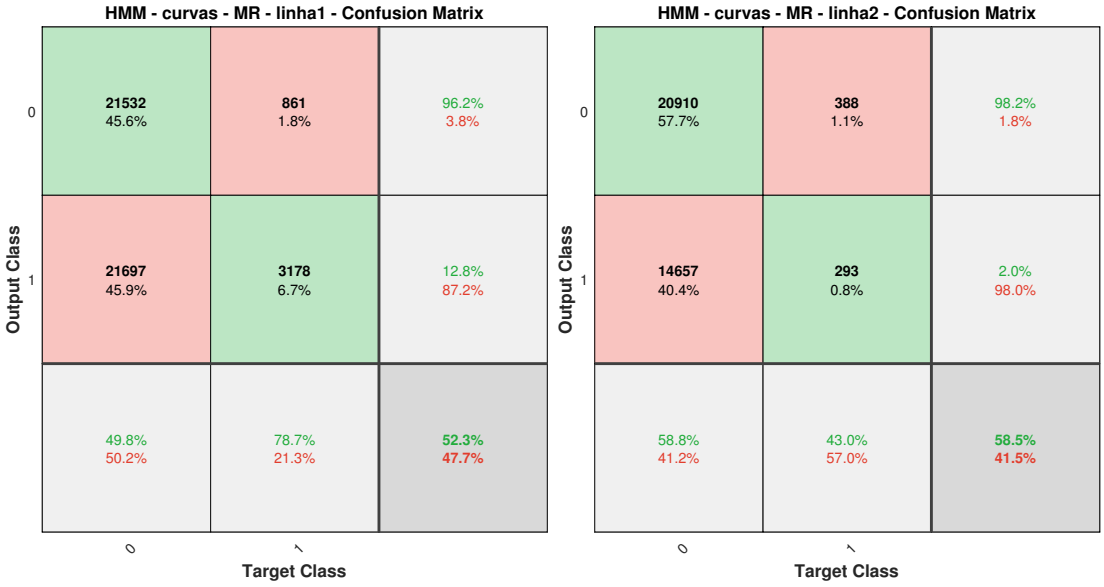
(a) CP/Linha1. (b) CP/Linha2.

Figura 20: Matriz de Confusão para a HMM (Curvas - Conselheiro Pena).
Fonte: O autor.



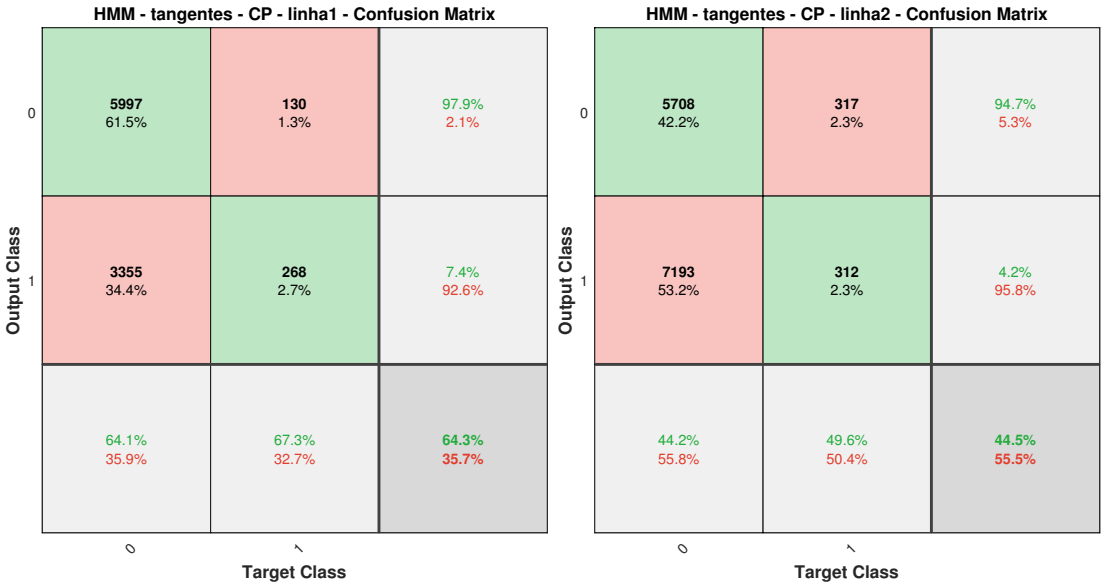
(a) GV/Linha1. (b) GV/Linha2.

Figura 21: Matriz de Confusão para a HMM (Curvas - Governador Valadares).
Fonte: O autor.



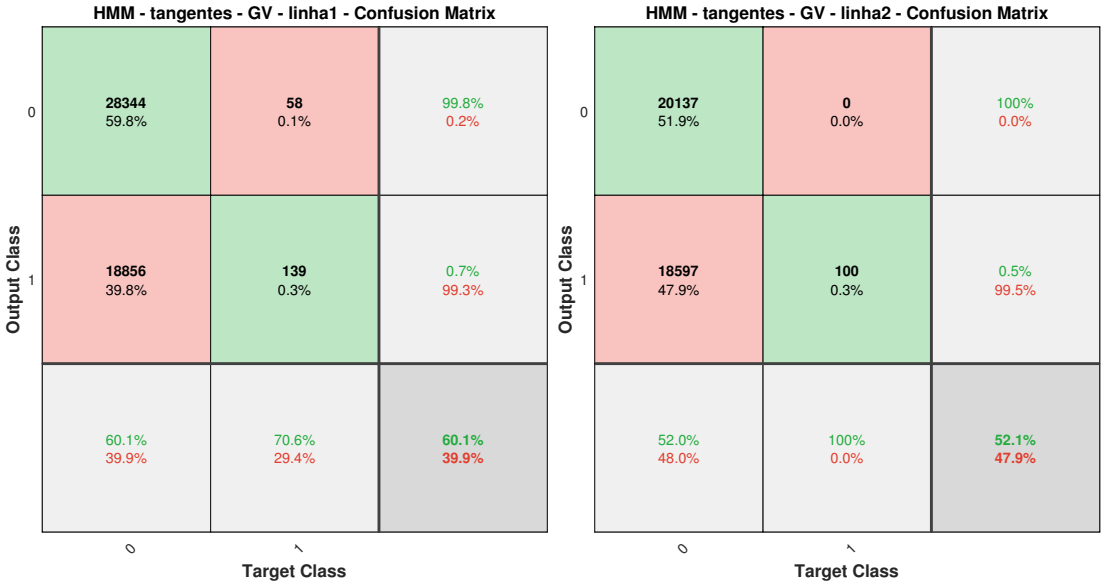
(a) MR/Linha1. (b) MR/Linha2.

Figura 22: Matriz de Confusão para a HMM (Curvas - Mário Carvalho).
Fonte: O autor.



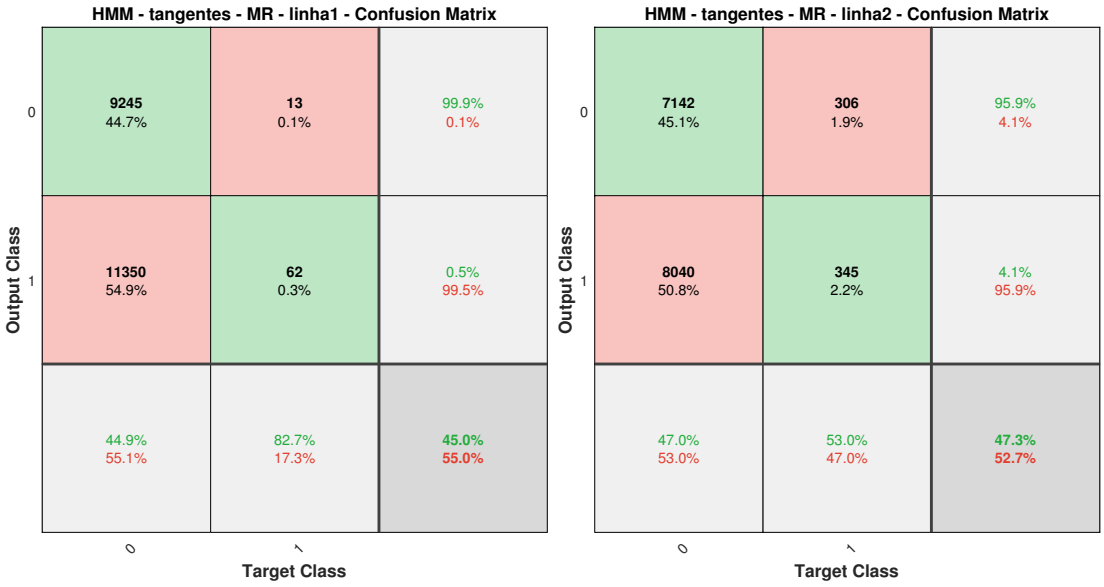
(a) CP/Linha1. (b) CP/Linha2.

Figura 23: Matriz de Confusão para a HMM (Tangentes - Conselheiro Pena).
Fonte: O autor.



(a) GV/Linha1. (b) GV/Linha2.

Figura 24: Matriz de Confusão para a HMM (Tangentes - Governador Valadares).
Fonte: O autor.



(a) MR/Linha1.

(b) MR/Linha2.

Figura 25: Matriz de Confusão para a HMM (Tangentes - Mário Carvalho).
Fonte: O autor.