REVISTA MILITAR de
CIÊNCIA e TECNOLOGIA

Print version: ISSN 0102-3543

Volume XL - 4° Quarter 2023

ENGENHARIA MILITAR

Mobilidade, Contramobilidade e Proteção

Revista Militar de Ciência e Tecnologia
Comandante do Exército:
Gen Ex Tomás Miguel Ribeiro Paiva

Departamento de Ciência e Tecnologia:
Gen Ex Achilles Furlan Neto

Departamento de Educação e Cultura do Exército:
Gen Ex Francisco Carlos Machado Silva

Comandante do Instituto Militar de Engenharia:
Gen Div Juraci Ferreira Galdino

Diretor da BIBLIEx:
Cel Art Marcos Walfrido Ricarte Figueiredo

Corpo Redatorial:

Editor-chefe:
Prof. Dr. Fernando Manuel Araújo Moreira

Editor-chefe adjunto:
Prof. Dr. Álvaro José Boareto Mendes

Editores adjuntos:
Prof. Dr. Marcelo de Miranda Reis
Prof. Dr. Paulo Henrique Coelho Maranhão

Editores de área:
Engenharia Civil, Transportes e Ciências Ambientais: Prof. Dr. Filipe Almeida Corrêa do Nascimento • Engenharia Elétrica e áreas afins: Prof. Dr. Antônio Eduardo Carrilho da Cunha • Engenharia Mecânica e áreas afins: Prof. Dr. André Luiz Tenório Rezende • Engenharia Química e áreas afins: Prof. Dr. Julio Zukerman Schpector • Engenharia Cartográfica: Prof. Dr. Vagner Braga Nunes Coelho • Engenharia Nuclear: Profa. Dra. Inaya Correa Barbosa Lima • Ciência e Engenharia de Materiais: Prof. Dr. André Ben-Hur da Silva Figueiredo • Ciência e Engenharia da Computação: Prof. Dr. Paulo Cesar Salgado Vidal • Engenharia de Defesa: Prof. Dr. Giuseppe Miceli Junior • Outras áreas: Prof. Dr. Marcelo de Miranda Reis

Editores associados externos:
• Dr. André Fenili – Universidade Federal do ABC, Santo André, SP • Dr. Fernando Fachini Filho – Instituto Nacional de Pesquisas Espaciais, SP • Dr. José Carlos Costa da Silva Pinto – Universidade Federal do Rio de Janeiro, RJ • Dr. José Carlos Maldonado – Universidade de São Paulo, São Carlos, SP • Drª. Júlia Célia Mercedes Strauch – Escola Nacional de Ciências Estatísticas, RJ • Dr. Luiz Pereira Calôba – Universidade Federal do Rio de Janeiro, RJ • Dr. Otto Corrêa Rotunno Filho – COPPE/Universidade Federal do Rio de Janeiro, RJ • Dr. Richard Magdalena Stephan – COPPE/Universidade Federal do Rio de Janeiro, RJ • Dr. Webe João Mansur – COPPE/Universidade Federal do Rio de Janeiro, RJ • Dr. Carlos Alberto Nunes Consenza – COPPE/Universidade Federal do Rio de Janeiro, RJ

Direção, Impressão e Distribuição:

Diretor:
Cel Art Marcos Walfrido Ricarte Figueiredo

Editor executivo:
Cap R1 Antônio Carlos Manhães de Souza

Administração, revisão, programação e diagramação:
Coordenação: INSTITUTO MILITAR DE ENGENHARIA – IME

Redação, Diagramação e OJS:
Rubenildo Pithon de Barros
Marcela Fagundes Casotti
Luiz Tadeu Carqueija Mota

Revisão Tikinet:
Revisores: Cristina Saez; Guilherme Oliveira; Piero Kanaan • Diagramador: Ilário Junior

Impressão e distribuição:
BIBLIEx

Desde 1949
"A Gráfica do Exército" - Compromisso com a Qualidade

Impresso na Gráfica do Exército
Al. Mal. Rondon - Setor de Garagens - QGEx - SMU - CEP:70630-901 - Brasília - DF
Tel: 3415-4367 - Site: http://www.graficadoexercito.eb.mil.br
E-mail: divcmcl@graficadoexercito.eb.mil.br

PUBLICAÇÃO TRIMESTRAL

REVISTA MILITAR de CIÊNCIA e TECNOLOGIA

REVISTA MILITAR DE CIÊNCIA E TECNOLOGIA - Volume XL
4º Trimestre de 2023

INSTITUTO MILITAR DE ENGENHARIA - IME
Praça General Tibúrcio, 80 - Praia Vermelha -
Rio de Janeiro-RJ - CEP 22.290-270 -
Tel.: (21) 2546-7115
Website: https://ebrevistas.eb.mil.br/CT - Email: rmct@ime.eb.br

BIBLIOTECA DO EXÉRCITO
EDITORA (BIBLIEx)
Palácio Duque de Caxias - Praça D. de Caxias, 25
3º andar - Ala Marcílio Dias
Centro - Rio de Janeiro-RJ
CEP 20.221-260
Tel.: (21) 2519-5707

ACESSE NOSSA REVISTA DIGITAL

Nossa capa:
Homenagem ao Quadro de Engenheiros Militares

EDITORIAL

REVISTA MILITAR de CIÊNCIA e TECNOLOGIA

The convergence of advanced materials, manufacturing processes, and artificial intelligence techniques is reshaping modern engineering and infrastructure. The six lines of research presented here — from microalloyed alloys and heat treatments to malware detection and road pathology analysis — demonstrate not only isolated advances but an integrated panorama where materials science, additive manufacturing, and statistical methods powered by neural networks complement each other to provide greater performance, safety, and sustainability.

The combination of alloy development and surface characterization is central to structural and biomedical applications. Studies on low-carbon microalloyed steels subjected to the Quenching and Partitioning process reveal opportunities to increase toughness and mechanical strength without sacrificing ductility, optimizing microstructures through thermal control. In parallel, the analysis of titanium surfaces during silanization demonstrates how targeted surface treatments can modulate adhesion, biocompatibility, and corrosion resistance, which are essential for implants and aerospace components. The interaction between the core microstructure (thermally treated alloys) and surface chemistry (silanization) defines the final functional behavior of parts, suggesting that integrated material and surface design can bring exponential performance gains.

The introduction of additive manufacturing by directed energy deposition (DED-AM) using atomized H13 tool steel powder in Brazil represents a strategic leap: reduced lead time, greater geometric complexity, and the possibility of localized repairs. However, a successful transition requires attention to powder quality, densification control, anisotropic properties, and integration with subsequent treatments such as quenching and partitioning. Developing a national supply chain for atomized powders strengthens technological autonomy and reduces logistics costs, but it must proceed alongside rigorous characterization protocols (microstructure, hardness, mechanical properties) and evaluations of compatibility between additive processes and subsequent heat treatments.

The use of convolutional neural networks for identifying and classifying road pathologies and detecting novelties in malware families illustrates two areas where machine learning is transforming prevention and response. On highways, convolutional models integrated with embedded systems and high-resolution images allow prioritization of interventions based on risk, cost, and social impact, while statistical methodologies for reducing wear on railway tracks, when compared to actual wear, provide a feedback loop that optimizes predictive maintenance. In cybersecurity, detecting emerging variations in malware families requires pipelines that combine behavioral analysis, dynamic signatures, and anomaly detection algorithms, promoting faster and more resilient responses.

The intersection of materials science, additive manufacturing, and artificial intelligence provides a powerful platform to rethink how we design, produce, and maintain structures and systems. Prioritizing integration between experimental processes, local supply chains, and intelligent models is not just an efficiency strategy: it is an imperative to ensure safety, longevity, and technological autonomy in the face of the industrial and digital challenges of the 21st century.

In conclusion, the intersection of materials science, additive manufacturing, and artificial intelligence provides a powerful platform to rethink how we design, produce, and maintain structures and systems. Prioritizing the integration of experimental processes, local supply chains, and intelligent models is not just an efficiency strategy: it is an imperative to ensure safety, longevity, and technological autonomy in the face of the industrial and digital challenges of the 21st century. It is in this context that the Military Engineering Institute is situated, a corporate university founded in 1792 whose primary objective is to meet the needs of the Brazilian Army, mainly by enhancing its combat power. This is pursued without neglecting dual-use applications that can contribute as much as possible to civil society.

In this context, the role to be played by the Military Magazine of Science and Technology (RMCT, an Open Access publication), one of the regular publications of the Brazilian Army, emerges as paramount in the field of military science and technology.

Excellent reading to everyone!

Dr. Eng. Fernando M. Araujo-Moreira
Editor-in-chief of RMCT

SUMMARY

VOL.40 Nº4 2023
https://doi.org/10.22491/rmct.v40i4.11950.en

Evaluation of micro-alloyed steels with low carbon and high mechanical strength, subjected to the Hardening and Partitioning process

Flávia Tolomelli[1,]*, Andersan dos Santos Paula[1], Fernando Cosme Rizzo Assunção[2]
[1]Instituto Militar de Engenharia, Rio de Janeiro, RJ, Brasil
[2] Pontifícia Universidade Católica do Rio de Janeiro - PUC-RJ, Rio de Janeiro, RJ, Brasil
***flavia.tolomelli@ime.eb.br**

ABSTRACT: Quenching and partitioning can produce AHSS (advanced high strength steel) with good toughness and ductility. This study evaluated the performance of two alloys of these steel forms (with Ti or Cr addition) that were subjected to one-stage quenching and partitioning. This study used the ThermoCalc program (TCFE9) to determine austenite composition and empirical equations to determine the main transformation temperatures of interest in this process, which were used as a basis for quenching heat treatments in oil. The alloys were characterized by optical and scanning electron microscopy and phase quantification on Image J. The latter was used to indirectly evaluate the retained austenite content as a function of the MA constituent fraction (martensite-austenite). Results found that the alloy with Ti-addition showed higher fractions of MA and higher yield strength, tensile strength limit, and total elongation than than with Cr, indicating its greater aptitude for quenching and partitioning.

KEYWORDS: *Microalloyed steels; Q&P; ThermoCalc; retained austenite.*

RESUMO: Por meio do processo de têmpera e partição podem ser obtidos aços AHSS (Aços Avançados de Alta Resistência) com boa tenacidade e ductilidade. Neste trabalho é avaliado o desempenho de duas ligas desses aços (com adição de Ti ou Cr) submetidas ao processo de Têmpera e Partição (T&P) de um estágio. O estudo envolveu o uso do programa ThermoCalc (TCFE9) para determinar a composição da austenita, bem como equações empíricas para determinar as principais temperaturas de transformação de interesse nesse processo, usadas como base para os tratamentos térmicos de têmpera com resfriamento em óleo. A caracterização das ligas foi feita usando microscopia ótica e eletrônica de varredura, como também quantificação de fase com auxílio do software Image J. Esse último para avaliar de forma indireta o percentual de austenita retida em função da fração de constituinte MA (martensita-austenita). Como resultado foi observado que a liga com adição de Ti comparada à liga com adição de Cr além de apresentar maiores frações de MA, apresentou maiores valores de limite de escoamento, limite de resistência e alongamento total, indicando que ela apresenta maior aptidão ao processo T&P.

PALAVRAS-CHAVE: Aços microligados; T&P; ThermoCalc; Austenita retida.

1. Introduction[1]

The third generation of advanced high-strength steels (AHSS) are aimed at adding greater toughness and ductility to first-generation AHSS. This is possible because these steels contain a high volumetric fraction of fine-grained ferrite, carbide-free bainite, martensite, and retained austenite in their microstructure. The ultrafine-grained ferrite, bainite, and martensite constituents provide the level of mechanical strength, while the ductility and high strain-hardening exponent are highly dependent on the fraction and mechanical stability of the austenite (GRAJCAR, KUZIAK, ZALECKI, 2012; MATLOCK, SPEER, 2009). The concept behind the microstructure design of third-generation AHSS is to achieve a considerable amount of retained austenite (> 20% by volume) within a martensitic/ferritic matrix. To obtain this sophisticated multiphase structure, complex thermal processing routes are employed (BLECK et al., 2019). Thus, steels produced through the quenching and partitioning process (Q&P steels) stand out.

The quenching and partitioning process (Q&P process), initially proposed by Speer, has proven to be a viable route for obtaining steels with mixed microconstituents, making them attractive due

---

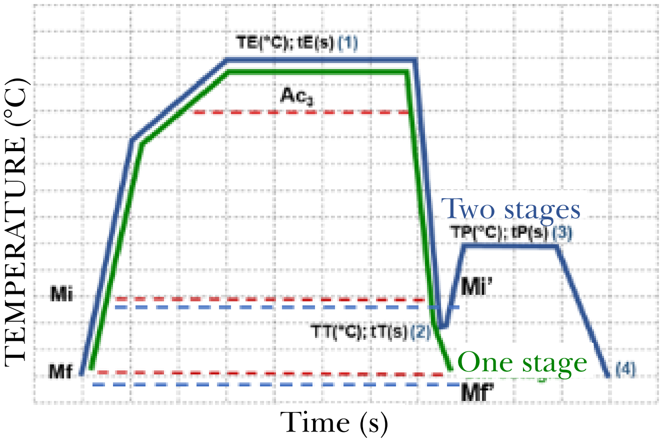
[1] Observações iniciais ou contextualização, se necessárias, devem ser incluídas como nota de rodapé na in
Os títulos dos itens e subitens são apenas exemplos. No entanto, sugerimos usar o formato 1.2.3 para sua subdivisão.

to their excellent combination of properties with minimal alloying element addition (SEO, 2019). A good alloy design for producing quenching and partitioning steel with a microstructure consisting of martensite laths separated by thin films of retained austenite should consider the addition of alloying elements that, combined with the high cooling rates employed in the quenching process, allow what follows. These are: preventing of ferrite and pearlite phases during quenching; the delaying or inhibiting bainite formation; avoiding or minimizing carbide precipitation (which would consume the carbon needed to stabilize the austenite); and achieving a sufficiently high carbon content to stabilize a considerable fraction of retained austenite at room temperature (SANTOFIMIA, 2011). For this purpose, in addition to C, these steels contain other alloying elements such as Mn, Ni, and Cr, which delay the formation of ferrite, pearlite, and bainite; and Si, Al, or P, which, added individually or in combinations, inhibit the precipitation of epsilon carbides and cementite, which would consume the carbon necessary for austenite stability. Carbon enrichment of austenite is considered beneficial because the TRIP effect during deformation can significantly contribute to the steel formability and energy absorption capacity (SANTOFIMIA, 2008). Additionally, carbon and Mn, along with other gamma-stabilizing alloying elements, reduce the A_{c3} or A_{r3} temperature (i.e., eutectoid transformation or decomposition temperature [A_{e3}] outside equilibrium during cooling or heating, respectively), increasing the austenite stability field.

The Q&P process, shown in Figure 1, was proposed to create steel microstructures containing retained austenite and has been the subject of various research studies since then (SPEER, 2003). It consists of an initial soaking heat treatment, starting from either full austenitization (austenitic field) or partial austenitization (intercritical field) of the steel, followed by cooling at a rate that inhibits any diffusional or mixed transformation, interrupted at a temperature between M_i and M_f, i.e., the martensitic transformation start and finish temperatures, as shown in Figure 1. At the end of this stage of the process, the microstructure consists of controlled fractions of martensite and retained austenite. During the quenching time (Qt) at this temperature range (one-stage treatment) or at a temperature slightly above M_i (two-stage treatment), carbon partitioning occurs from the carbon-supersaturated martensite to the untransformed austenite, aiming to stabilize it by lowering its M_i' and M_f' temperatures (as shown in Figure 1), meaning that after final cooling, carbon-enriched austenite remains present down to room temperature.

Thus, the final volumetric fraction of austenite can be controlled by the interruption temperature of the quenching cooling between M_i and M_f. At the end of the treatment, microstructures consisting of martensite and retained austenite (if soak was conducted from the austenitic field) or free ferrite, martensite, and retained austenite (if soak was conducted from the intercritical field) are obtained (SPEER, 2005; SPEER, 2011).

Figure 1 - single-stage and two-stage Q&P process applied to Q&P-suitable steels. M_i and M_f = refer to retained austenite during quenching; $M_{i'}$ and $M_{f'}$ = refer to carbon-enriched austenite during partitioning.



The factors influencing the volume fraction of retained austenite in Q&P steels primarily include the chemical composition, which determines the M_i temperature of austenite; the quenching temperature

(QT), which defines the initial volume fraction of martensite; and the partitioning time, which determine the final volume fraction of retained austenite. If the partitioning time is too short, the carbon diffusion from martensite to austenite will be insufficient, causing the retained austenite to become unstable and transform into martensite during the subsequent cooling process. If the partitioning time is too long, the retained austenite will become enriched in carbon but may decompose into ferrite and carbide or into bainite at the partitioning temperature (PT), reducing the volume fraction of retained austenite (WANG, SPEER, 2013; TOJI, MIYAMOTO, RAABE, 2015; KIM et al., 2009; LI et al., 2010; DE MOOR et al., 2006).

Other factors include the austenitic grain size and austenite morphology. The carbon percentage of austenite is affected by the partitioning time and temperature, such that increasing both the time and temperature of partitioning will increase the carbon diffusion rate and, consequently, the stabilization of retained austenite (KNIF et al., 2014; XIE et al., 2019; ZINSAS-BORUJERDI et al., 2018; ZHAO et al., 2014).

Regarding mechanical properties, according to Zinsaz-Borujerdi et al. (2018), the volume fraction of retained austenite in steels subjected to the Q&P process is considered the main factor affecting ductility, i.e., the total elongation evaluated in a uniaxial tensile test. The primary reason for this is that the total elongation of the steel depends not only on the percentage of retained austenite but also on its carbon content. This is because the stability of carbon-enriched austenite at the end of the soak at partitioning temperature determines whether, in the subsequent cooling, there will be only retained austenite alongside tempered martensite or retained austenite combined with additional martensite with higher carbon content along with some retained austenite. Increasing the partitioning time or temperature raises the carbon diffusion rate and, consequently, the stability of retained austenite (ZHAO et al., 2014). Matsumura, Sakuma, and Takechi (1987) concluded that the higher the percentage and stability of retained austenite, the greater the total elongation obtained.

Various studies have shown the influence of processing conditions of Q&P steels and their microstructures on the results of mechanical properties (JUNG, 2011; ARLAZAROV et al., 2016; KICKINGER et al., 2021; CHENG et al., 2022; JING, 2014; HE et al., 2019).

The objective of this work was to evaluate two AHSS microalloyed steels with different compositions, regarding their suitability for producing Q&P steel when subjected to a one-step Q&P process. After obtaining the microstructures resulting from the proposed Q&P process, their characterization was performed using optical and scanning electron microscopy, based on prior contrast using Klemm's color etching solution and the conventional 3% Nital reagent, respectively. This allowed the identification of the microconstituents through image analysis of the optical microscopy micrographs, as well as the quantification of retained austenite through the estimation of the MA constituent percentage using ImageJ software. These results were further complemented by mechanical characterization through uniaxial tensile testing.

2. Materials and Methods Method

2.1. Alloy design determination

The selection of alloys, aiming to evaluate the suitability of the steels for the Q&P process, considered a basic chemical composition of Fe-C-Mn-Si for two microalloyed Nb steels, with Alloy A containing additions of Ti and Mo, and Alloy B with additions of Cr and Mo, as shown in Table 1. The alloys were cast on an industrial scale in an LD converter, and after the hot rolling, pickling, and cold rolling processes, samples were taken from the sheets for this study.

Table 1 - Chemical composition (%wt) of the alloys selected for the study. Source: X

Alloys	C	Mn	Cr	Si	Mo	Ti+nB
Alloy A	0.15	2.0	-	0.5	0.3	.1
Alloy B	0.20	2.0	0.3	0.5	0.3	0.050
(maximum values).						

2.2. Determination of soaking temperatures for quenching treatment

The ThermoCalc software, using the TCFE9 database version, was used to determine temperatures that define the intercritical region in equilibrium (Ae1 and Ae3), evaluate the chemical composition of austenite (carbon percentage), and the precipitation temperatures of phases from the austenitic region.

2.2.1. Empirical equations

Empirical equations (GORNI, 2019) for determining A_{c1}, A_{c3}, and M_i temperatures were adopted as indicated in equations 1, 2, and 3, They were then compared with the obtained values using Thermo-Calc. These temperatures were used as the basis for the proposed quenching heat treatments in this work for the one-step Q&P process.

$$Ac_1 = 739 - 22{,}8C - 6{,}8Mn + 18{,}2Si + 11{,}7Cr - 15Ni - 6{,}4Mo - 5V - 28Cu \qquad \text{Equation 1.}$$

$$Ac_3 = 937{,}3 - 224{,}5C - 17Mn + 34Si - 14Ni + 21{,}6Mo + 41{,}8V - 20Cu \qquad \text{Equation 2.}$$

$$M_i = 539 - 423C - 30{,}4Mn - 17{,}7Ni - 12{,}1Cr - 11Si - 7{,}5Mo \qquad \text{Equation 3.}$$

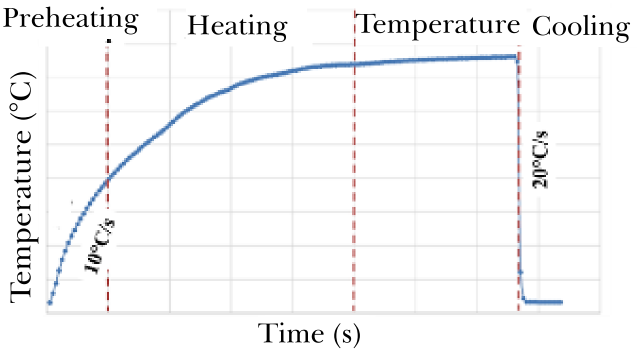
2.3 Sampling for heat treatments

Cold-rolled sheets with a thickness of 1.50 mm, produced on an industrial scale, were collected, and test specimens prepared with dimensions of 90 x 270 mm (longitudinal direction – LD in relation to the rolling direction) for conducting heat treatments. The one-step Q&P process with different soaking temperatures for each steel was carried out in a muffle furnace, with temperature monitoring through K-type thermocouples attached to the sheets and connected to a Fluke device.

2.4. One-step Q&P process

For the quenching and partitioning heat treatments, two soaking conditions were used: soaking in the austenitic field and soaking in the intercritical field, due to the proposed one-step Q&P process. For soaking in the austenitic field, a temperature of 840°C was adopted for both alloys of the study, as the A_{c3} temperatures of the alloys were below this temperature. For the intercritical field heat treatment, based on the A_{c1} and A_{c3} temperatures obtained, temperatures of 780°C and 820°C with a soaking time of 120 s and a hold time of 20 s were used. The cooling medium was room temperature oil, and the cooling rate achieved was over 20°C/s. An example of a graph obtained for monitoring the thermal cycle profile during the heat treatments can be seen in Figure 2. It shows that heating occurred with a non-linear profile and a decreasing rate as temperature increased, subdivided into two regions: preheating (average rate of 10°C/s) and heating (lower rate with more variation). Thus, the sheets required a hold time of approximately 90 seconds in the preheated furnace at the proposed soaking temperature before the soaking time could be considered.

Figure 2 - Heat treatment scheme used for Alloys A and B.



2.5. Metallographic characterization

Two techniques were used for metallographic characterization, as follows.

a. Optical microscopy (OM) – using a Zeiss microscope – with conventional reagents based on Nital (3 mL of nitric acid diluted in 97 mL of ethanol) and Klemm (aqueous saturated solution of sodium thiosulfate: 50 mL (solubility of anhydrous sulfate (Na2S2O3) 50 g/100 mL at 20°C, hydrated sulfate solubility 291.1 g/100 mL at 45°C, and 1 g of potassium metabisulfite). Etching time varied depending on the microstructure;

b. Scanning electron microscopy (SEM) – QUANTA 3D FEG – based on the microstructures revealed with the Nital 3% solution.

2.6. Quantitative metallography

To conduct an estimated quantitative evaluation of the retained austenite fraction present in the microstructure and associated with the MA constituent percentage, the ImageJ software was used as a tool to aid the characterization. Two test specimens from each alloy, subjected to the same treatment conditions, were used for a comparative analysis of the microstructures resulting from the Nital 3% etching.

2.7. Mechanical characterization

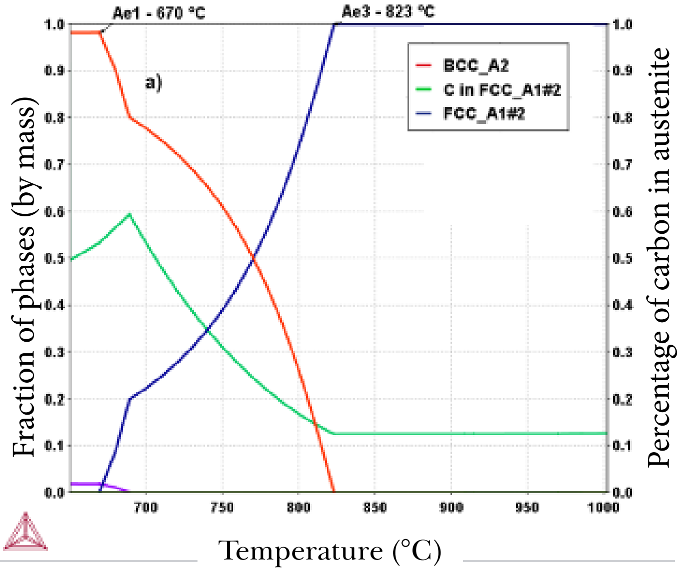
Uniaxial tensile tests were conducted on an Instron 5585H 25t machine, with total elongation determined by an AVE video extensometer. A total of two test specimens per condition under study for each alloy were taken from the longitudinal position and followed the ASTM-E-290 standard, being of the ISO I type with a 50 mm gauge length.

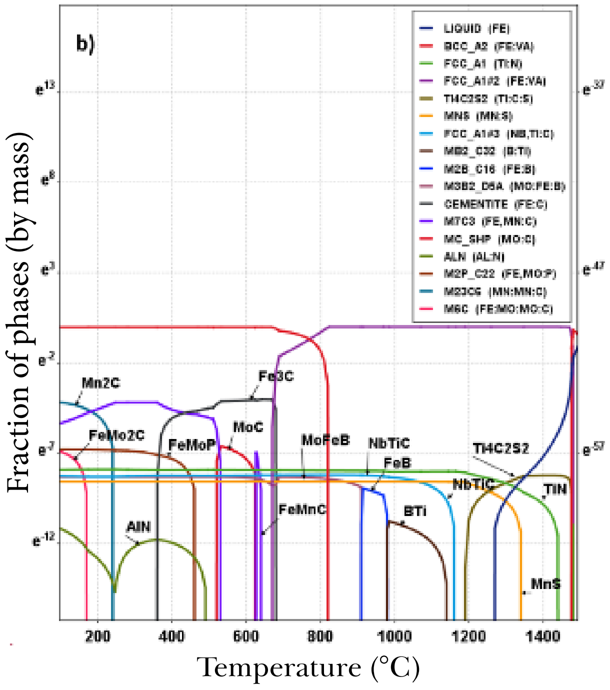
3. Results and Discussions

3.1 Thermodynamic aspects

In Figure 3 (a,b), the A_{e1} and A_{e3} temperatures, calculated via Thermocalc for alloys A and B, can be observed, showing a larger stability field for austenite in the first alloy compared to the second. In Figure 4 (a,b), the precipitation temperatures of carbides, nitrides, and carbonitrides for the respective alloys are shown. It is noted that, under equilibrium conditions, alloy A may consume more carbon in the formation of precipitates (08 carbon-based precipitates) than alloy B (05 carbon-based precipitates), indicating that in the latter, there is a greater likelihood of excess carbon available to stabilize austenite at room temperature compared to the former.

Figure 3 - Graphs related to thermodynamic data of Alloy A under equilibrium conditions, highlighting: (a) the phase transformation temperatures, A_{e1} and A_{e3}, and the distribution of carbon in austenite; and (b) the phase maps.





3.2 Carbon Percentage in Austenite and CE

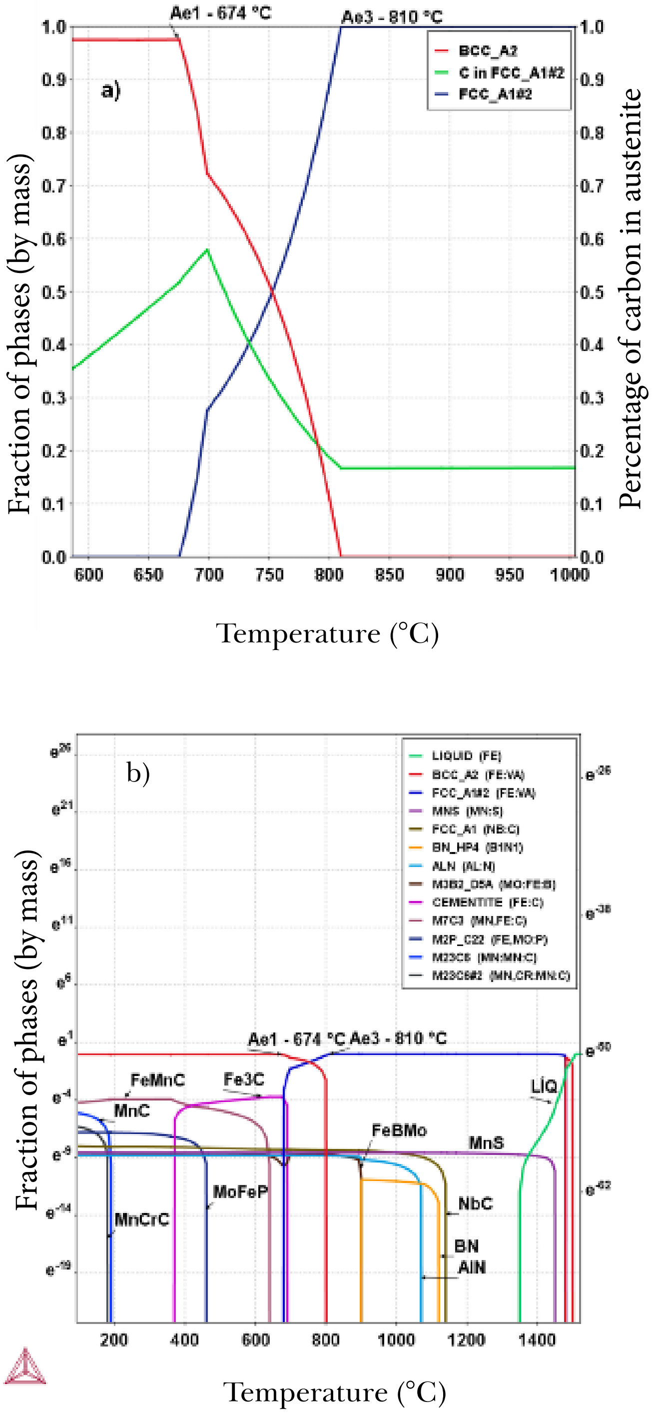
For alloy A it can be seen in Figure 3(a) that the carbon percentages in austenite (under equilibrium conditions) are above 0.20 %C at 780 °C and below 0.15 %C at 820 and 840 °C. And for alloy B it can be seen in Figure 4 (a) that these values remained above 0.20 %C at the temperature of 780 °C but remained above 0.15 %C at the temperatures of 820 and 840 °C, respectively.

In Table 2, it can be observed that the carbon equivalent content of alloy A (0.71; 0.52; 0.52%) is lower than that of alloy B (0.73; 0.60; 0.60%), respectively at the temperatures 780°C, 820°C, and 840°C, indicating that alloy B may exhibit greater hardenability compared to alloy A.

Table 2 - Carbon equivalent percentage of alloys A and B at different soaking temperatures.

	Carbon equivalent percentage of alloys	
T(°C)	Alloy A	Alloy B
780	0.71	0.73
820	0.52	0.60
840	0.52	0.60

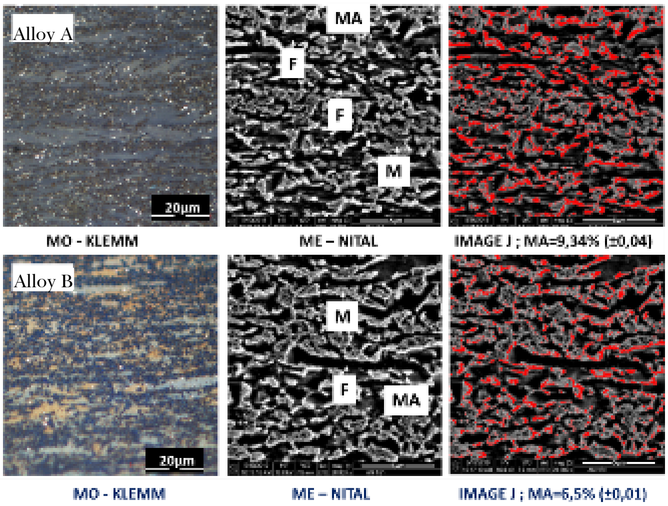
Figure 4 - Graphs related to thermodynamic data of Alloy B under equilibrium conditions, highlighting: (a) the phase transformation temperatures, Ae1 and Ae3, and the distribution of carbon in austenite; and (b) the phase maps.



3.3 Microstructural evaluation

Figures 5, 6, and 7 present the microstructural aspects resulting from the quenching heat treatment, performed in the intercritical field at soaking temperatures of 780°C and 820°C, and in the austenitic field at a soaking temperature of 840°C, respectively. The soaking time was of two minutes, based on analyses performed with optical microscopy (OM) and scanning electron microscopy (SEM).

Figure 5 - Microstructural characterization: OM (500x) and SEM (20,000x), and their results (estimate) of retained austenite quantification (related to the MA constituent) for alloys A and B subjected to the Q&P process, with soak at 780 °C.



The micrographs from the microstructural analysis are displayed in Figures 5, 6, and 7, from left to right as follows. Optical microscope micrograph resulting from the colored etching with Klemm's solution (first column). Scanning electron microscope micrograph resulting from the etching with 3% Nital solution (second column). And the respective image analysis for retained austenite quantification using ImageJ (third column), based on the selection of lighter regions highlighted in the SEM micrographs.

Figure 6 - Microstructural characterization: OM (500x) and SEM (20,000x), and their results (estimate) of retained austenite quantification (related to the MA constituent) for alloys A and B subjected to the T&P process, with soak at 820 °C.

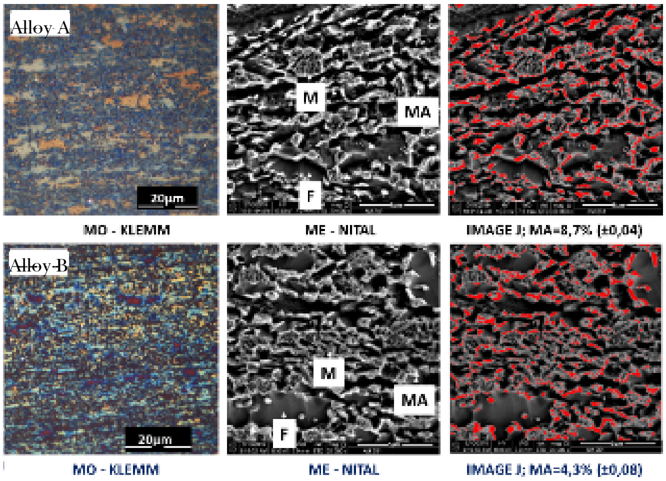
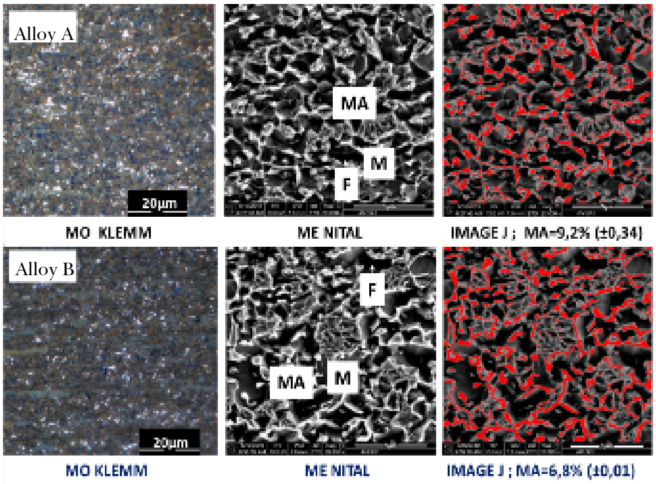


Figure 7 - Microstructural characterization: OM (500x) and SEM (20,000x), and their results (estimate) of retained austenite quantification (related to the MA constituent) for alloys A and B subjected to the T&P process, with soaking at 840 °C.



When comparing the quantification of samples treated at 780 °C, 820 °C, and 840 °C, a slight reduction in the retained austenite percentage (related to the MA constituent) was observed as the soaking temperature increased from 780 °C to 820 °C. This may be associated with the decrease in the carbon percentage in the alloys from 0.20%C at 780°C to 0.14%C at 820°C for alloy A, and from 0.25%C at 780°C to <0.15%C for alloy B, as seen with the increasing treatment temperature. This reduces the possibility of retained austenite stability at room temperature due to the decrease in M_f temperature and a higher proportion of austenite in the MA constituent.

It was also observed that alloy A exhibited a higher percentage of retained austenite (estimated by the MA constituent percentage) compared to alloy B at 840 °C. In this temperature range, the carbon percentage in the austenite matches the carbon percentage in the alloys, i.e., 0.13%C for alloy A and >0.15%C for alloy B.

3.4. Tensile test

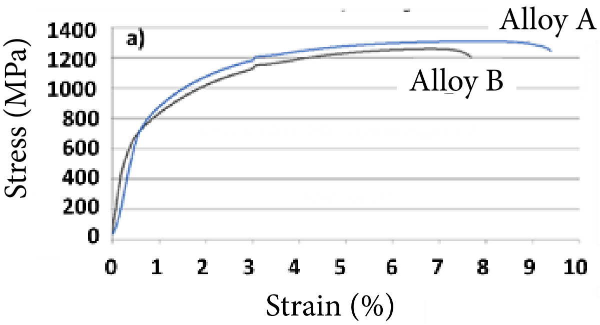
Table 3 presents the mechanical properties (average of two test specimens) for the variables of yield strength, YS (MPa), ultimate tensile strength, UTS (MPa), total elongation, TE (%), and strain hardening exponent, n-value (measured up to a maximum of 6% strain). Alloy B, treated at 820 ° was an exception, in which measurement was taken up to 4% strain), extracted from uniaxial tensile tests conducted on specimens for the two alloys studied, treated at temperatures of 780, 820, and 840 °C (as shown in Figure 8), with these values correlated with the MA constituent percentages quantified in the metallographic analysis.

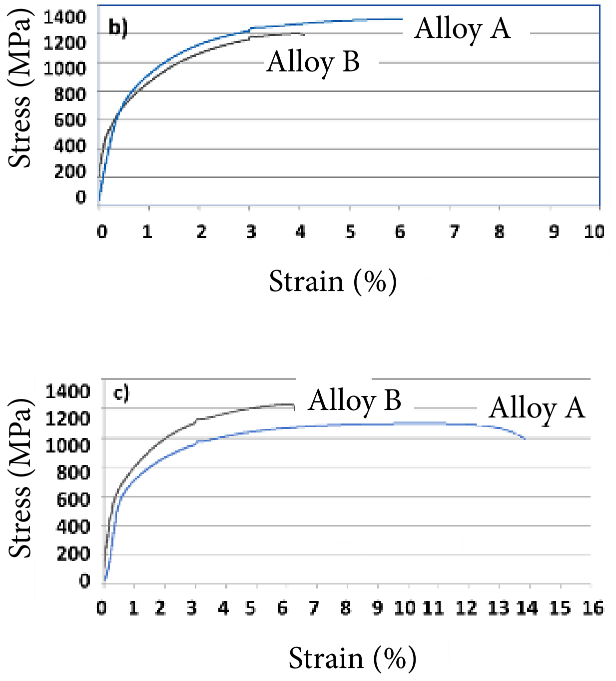
Table 3 - Results of mechanical properties of alloys A and B, correlated with the MA constituent fraction as a function of the heat treatment temperature.

Ligas	Temp°C) LE	(Mpa) LR	(Mpa)	Al (%)	$N_{4-6\%}$	MA (%)
A	780	804	1309	9.4	0.15	9.34 ± 0.2
B	780	670	1270	7.7	0.17	6.5 ± 0.3
A	820	760	1306	6.5	0.12	8.7 ± 0.1
B	820	592	1199	4.1	0,09	4.3 ± 0.2
A	840	639	1100	13.9	0.19	9.2 ± 0.25
B	840	631	1152	6.5	0.19	6.8 ± 0.3

The results obtained highlight that Alloy A, by presenting higher values of MA constituent fraction (average of 10 measurements), also showed superior values for yield strength (LE, MPa), tensile strength (LR, MPa), and total elongation (AL, %) when treated at temperatures of 780 °C and 820 °C. However, the specimens resulting from heat treatment at 840 °C for both alloys showed a decrease in LE (MPa) and LR (MPa) values, while maintaining higher AL (%) for Alloy A. Finally, the n4-6% values did not show a conclusive behavior based on the alloys under study, soak temperatures during the quenching heat treatment, and the observed MA constituent fraction

Figure 8 - Engineering stress-strain curves for alloys A and B subjected to the Q&P process with soak temperatures of 780 °C (a), 820 °C (b) (intercritical range), and 840 °C (c) (austenitic range).





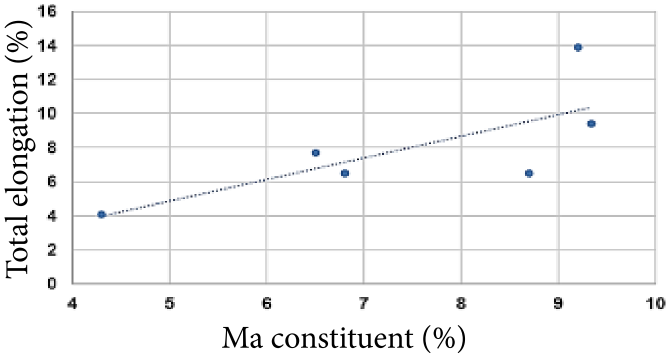
From the analysis of the engineering stress-strain curves shown in Figure 8, it is more evident that the increase in soak temperature used in the treatments resulted in a continuous improvement in total elongation and toughness for Alloy A compared to Alloy B, for the soak time applied in the treatment (2 minutes).

3.5. Influence of retained austenite percentage (related to the ma constituent) on total elongation

Figure 9 shows the quantified values of the MA constituent, related to retained austenite, with the total elongation observed in the specimens submitted to the uniaxial tensile test, after being subjected to the quenching heat treatments at the soaking temperatures of 780, 820 and 840 °C.

The trendline in the graph in Figure 9 shows that higher total elongation values tend to be obtained as the percentage of retained austenite, associated with the MA constituent, increases.

Figure 9 - Influence of retained austenite percentage (related to the % of MA constituent) on the total elongation of alloys A and B when subjected to a one-step T&P process.



4. Conclusion

Regarding the suitability for the Q&P process (with soak in the intercritical and austenitic fields during quenching), it is concluded that the factors that may justify the microstructural evolution and distinct mechanical behavior between the alloys evaluated in this study can be associated with what follows.

a. The action of Ti combining with N at high temperatures, forming the TiN compound and likely leaving B in solid solution, which delays the austenite transformation process during cooling. This can be more clearly observed at the temperatures practiced in the intercritical field (780°C and 820°C).

b. Although during cooling the percentage of carbon in the alloy with Cr addition was more favorable for the stabilization of austenite, during heating, the dissolution of precipitates in this alloy may not have occurred throughout the soaking phase, alongside the recrystallization process of this alloy.

c. It is also likely that, from a kinetic perspective, there was less carbide formation in the alloy with Ti addition compared to the alloy with Cr addition, leaving more carbon free for austenite stabilization during cooling.

d. Therefore, it can be inferred that the alloy with Ti addition possibly indicates greater suitability for the Q&P process compared to the alloy with Cr addition, under the treatment conditions adopted in this study. This is due to the presentation of higher values in mechanical properties (YS, UTS, and El) and a greater percentage of MA constituent, as well as likely greater austenite retention at the end of the quenching cooling phase.

Acknowledgements

The authors would like to thank CSN, IME and PUC-RJ for the support received with the samples, laboratory tests and characterization, as well as PRO-EX/CAPES for the funding subsidies provided to PPGCM/IME.

References

[1] ARLAZAROV, A.; OLLAT, M.; MASSE, J. P.; BOUZAT, M. Influence of partitioning on mechanical behavior of Q&P steels. *Materials Science and Engineering: A*, [*s. l.*], v. 661, p. 79-86, 2016. DOI: 10.1016/j.msea.2016.02.071.

[2] ASTM INTERNATIONAL. *E290-14: Standard test methods for bend testing of material for ductility*. [*s. l.*]. 2014.

[3] BLECK, W.; BRÜHL, F.; MA, Y.; SASSE, C. Materials and processes for the third-generation advanced high--strength steels. *Berg- und Hüttenmännische Monatshefte*, [*s. l.*], v. 164, p. 466-474, 2019. DOI: 10.1007/s00501-019-00904-y.

[4] CHENG, Y. Y.; ZHAO, G.; XU, D. M.; MAO, X. P.; BAO, S. Q.; YANG, G. W. Comparative study on microstructures and mechanical properties of Q&P steels prepared with hot-rolled and cold-rolled C–Si–Mn sheets. *Journal of Materials Research and Technology*, [*s. l.*], v. 20, p. 1226-1242, 2022. DOI: 10.1016/j.jmrt.2022.07.139.

[5] CLARKE, A. J.; SPEER, J. G.; MATLOCK, D. K.; RIZZO, F. C.; EDMONDS, D. V.; SANTOFIMIA, M. J. Influence of carbon partitioning kinetics on final austenite fraction during quenching and partitioning. *Scripta Materialia*, [*s. l.*], v. 61, n. 2, p. 149-152, 2009. DOI: 10.1016/j.scriptamat.2009.03.02.1.

[6] DE MOOR, E.; LACROIX, S.; SAMEK, L.; PENNING, J.; SPEER, J. Dilatometric study of the quench and partitioning process. *In*: INTERNATIONAL CONFERENCE ON ADVANCED STRUCTURAL STEELS, 3., 2006, Gyeongju. *Proceedings Of The 3rd International Conference On Advanced Structural Steels*. Gyeongju: Korean Institute of Metals and Matereials, 2006.

[7] EDMONDS, D. V.; HE, K.; RIZZO, F. C.; DE COOMAN, B. C.; MATLOCK, D. K.; SPEER, J. G. Quenching and partitioning martensite — A novel steel heat treatment. *Materials Science and Engineering: A*, [*s. l.*], v. 438-440, p. 25-34, 2006. DOI: 10.1016/j.msea.2006.02.133.

[8] GORNI, A. A. *Steel forming and heat-treating handbook*. [*s. l.*]: [*s. n.*], 2019.

[9] GRAJCAR, A.; KUZIAK, R.; ZALECKI, W. Third generation of AHSS with increased fraction of retained austenite for the automotive industry. *Archives of civil and mechanical engineering*, [s. l.], v. 12, n. 3, p. 334-341, 2012.

[10] HE, J.; HAN, G.; LI, S.; ZOU, D. To correlate the phase transformation and mechanical behavior of QP steel sheets. *International Journal of Mechanical Sciences*, [*s. l.*], v. 152, p. 198-210, 2019.

[11] KICKINGER, C.; SUPPAN, C.; HEBESBERGER, T.; SCHNITZER, R.; HOFER, C. Microstructure and mechanical properties of partially ferritic Q&P steels. *Materials Science and Engineering: A*, [*s. l.*], v. 815, p. 141-296, 2021.

[12] KIM, D. H.; SPEER, J. G.; KIM, H. S.; DE COOMAN, B. C. Observation of an isothermal transformation during quenching and partitioning processing. *Metallurgical and Materials Transactions A*, [*s. l.*], v. 40, p. 2048-2060, 2009.

[13] KNIF, D. D.; PETROV, R.; FÖJER, C.; KESTENS, L. A. I. Effect of fresh martensite on the stability of retained austenite in quenching and partitioning steel. *Materials Science Engineering: A*, [*s. l.*], v. 615, p. 107-115, 2014. DOI: 10.1016/j.msea.2014.07.054.

[14] LI, H. Y.; LU, X. W.; LI, W. J.; JIN, X. J. Microstructure and mechanical properties of an ultrahigh-strength 40SiMnNiCr steel during the one-step quenching and partitioning process. *Metallurgical and Materials Transactions A*, [*s. l.*], v. 41, n. 5, p. 1284-1300, 2010. DOI: 10.1007/s11661-010-0184-8.

[15] LIU, H.; JIN, X.; DONG, H.; SHI, J. Martensitic microstructural transformations from the hot stamping, quenching and partitioning process. *Materials characterization*, [*s. l.*], v. 62, n. 2, p. 223-227, 2011.

[16] MATLOCK, D. K.; SPEER, J. G. Third generation of AHSS: microstructure design concepts. *In:* HALDAR, A.; SUWAS, S.; BHATTACHARJEE, D. (eds). *Microstructure and Texture in Steels and Other Materials*. Londres: Springer, 2009. p. 185-205.

[17] MATSUMURA, O.; SAKUMA, Y.; TAKECHI, H. Enhancemente of elongation by retained austenite in intercritical annealed 0,4C-1,5Si-0,8Mn steel. *Transaction ISIJ*, [*s. l.*], v. 27, n. 7, p. 570-579, 1987.

[18] SANTOFIMIA, M. J.; ZHAO, L.; PETROV, R.; SIETSMA, J. Characterization of the microstructure obtained by the quenching and partitioning process in a low-carbon steel. *Materials Characterization*, [*s. l.*], v. 59, n. 12, p. 1758-1764, 2008.

[19] SANTOFIMIA, M. J.; ZHAO, L.; PETROV, R.; KWAKERNAAK, C.; SLOOF, W.G.; SIETSMA, J. Microstructural development during the quenching and partitioning process in a newly designed low-carbon steel. *Acta Materialia*, [*s. l.*], v. 59, n. 15, p. 6059-6068, 2011.

[20] SEO, E. J.; CHO, L.; KIM, J. K.; MOLA, J.; ZHAO, L.; DE COOMAN, B. C. Constituent-specific properties in quenching and partitioning (Q&P) processed steel. *Materials Science and Engineering: A*, [*s. l.*], v. 740-741, p. 439-444, 2019.

[21] SPEER, J. G.; ASSUNÇÃO, F. C. R.; MATLOCK, D. K.; EDMONDS, D. V. The "quenching and partitioning" process: background and recent progress. *Materials Research*, [*s. l.*], v. 8, n. 4, p. 417-423, 2005. DOI: 1590/S1516-14392005000400010.

[22] SPEER, J. G.; DE MOOR, E.; FINDLEY, K. O.; MATLOCK, D. K.; DE COOMAN, B. C.; EDMONDS, D. V. Analysis of microstructure evolution in quenching and partitioning automotive sheet steel. *Metallurgical and materials transactions A*, [s. l.], v. 42, p. 3591-3601, 2011.

[23] SPEER, J. G.; STREICHER, A. M.; MATLOCK, D. K.; RIZZO, F.; KRAUSS, G. Quenching and partitioning: a fundamentally new process to create high strength trip sheet microstructures. *In*: SYMPOSIUM ON THE THERMODYNAMICS, KINETICS, CHARACTERIZATION AND MODELING, 2003, Chicago. *Austenite formation and decomposition*. Warrendale: TMS, 2003. p. 505-522.

[24] SUN, J.; YU, H. Microstructure development and mechanical properties of quenching and partitioning (Q&P) steel and an incorporation of hot-dipping galvanization during Q&P process. *Materials Science and Engineering: A*, [*s. l.*], v. 586, p. 100-107, 2013.

[25] SUN, J.; YU, H.; WANG, S.; FAN, Y. Study of microstructural evolution, microstructure-mechanical properties correlation and collaborative deformation-transformation behavior of quenching and partitioning (Q&P) steel. *Materials Science and Engineering A*, [*s. l.*], v. 596, p. 89-97, 2014.

[26] TOJI, Y.; MIYAMOTO, G.; RAABE, D. Carbon partitioning during quenching and partitioning heat treatment accompanied by carbide precipitation. *Acta Materialia*, [*s. l.*], v. 86, p. 137-147, 2015.

[27] WANG, L.; SPEER, J. G. Quenching and partitioning steel heat treatment. *Metallography, Microstructure, and Analysis*, [*s. l.*], v. 2, p. 268-281, 2013. DOI: 10.1007/s13632-013-0082-8.

[28] ZHAO, C. *et al*. Effect of annealing temperature and time on microstructure evolution of 0,2C-5Mn steel during intercritical annealing process. *Materials Science Technology*, [*s. l.*], v. 30, n. 7, p. 791-799, 2014. DOI: 10.1179/1743284713Y.0000000416.

[29] ZINSAZ-BORUJERDI, A.; ZAREI-HANZAKI, A.; ABEDI, H. R.; KARAM-ABIAN, M.; DING, H.; HAN, D.; KHERADMAN, N. Room temperature mechanical properties and microstructure of a low alloyed TRIP-assisted steel subjected to one-step and two-step quenching and partitioning process. *Materials Science and Engineering: A*, [*s. l.*], v. 725, p. 341-349, 2018.

[30] XIE, Z. J.; HAN, G.; YU, Y. S.; SHANG, C. J.; MISRA, R. D. K. The determining role of intercritical annealing condition on retained austenite and mechanical properties of a low carbon steel: Experimental and theoretical analysis. *Materials Characterization*, [*s. l.*], v. 153, p. 208-214, 2019. DOI: 10.1016/j.matchar.2019.05.010

VOL.40 Nº4 2023
https://doi.org/10.22491/rmct.v40i4.12122.pt

Detection of Novelties in Malware Families

Ricardo Sant'Ana[1], Julio Cesar Cardoso Tesolin[1] e Julio Cesar Duarte[1]
[1]Military Institute of Engineering (IME)
Praça General Tibúrcio, 80, 22290-270, Praia Vermelha, Rio de Janeiro, RJ, Brazil
***ricardo.santana@ime.eb.br**

ABSTRACT: Many researches have already presented approaches to the malware detection task. Classifying them into families provides a better understanding of their behavior, allowing companies and researchers to optimize their efforts. Nevertheless, an issue still needs to be proper addressed: how to verify if an artifact detected as a malware belongs to a known family? This work proposes the use of two widely known classifiers - GMM and SVM - for a novelty detection task in malware analysis to redirect proper human and computational efforts for a quick counter measure. The main contribution of this work is the use of features directly extracted from the detected malwares's binary file such as entropy and image's texture for novelty detection.

KEYWORDS: Gaussian Mixture Model. Malware Family Detection. Novelty Detection. Support Vector Machine. Malware as an image. Entropy.

RESUMO: Muitas pesquisas já apresentaram abordagens para a tarefa de detecção de malware. Classificá-los em famílias fornece uma melhor compreensão de seu comportamento, permitindo que empresas e pesquisadores otimizem seus esforços. No entanto, um problema ainda precisa ser tratado corretamente: como verificar se um artefato computacional detectado como malware pertence a uma família já conhecida? Este trabalho propõe o uso de dois classificadores amplamente conhecidos – GMM e SVM – para uma tarefa de detecção de novidade em análise de malware, em que o objetivo é direcionar esforços humanos e computacionais adequados para fornecer uma rápida contramedida. A principal contribuição deste trabalho está no uso de características diretamente extraídas do arquivo binário do malware detectado, como entropia e textura de imagem para a tarefa de detecção de novidade.

PALAVRAS-CHAVE: Modelo de Misturas Gaussianas. Detecção de Família de Malware. Detecção de Novidade. Máquina de Vetores de Suporte. Malware como imagem. Entropia.

1. Introduction

Detecting and protecting against malicious computer artifacts constitutes a high-profile topic in cybersecurity. The literature includes various definitions of malicious artifacts (or malware) [12]. In general, these software can damage users, systems, and networks, corrupting their codes, harming their operations, and/or stealing their information.

A 2017 survey by a large company that produces software to combat malicious artifacts [1] involving 1,300 professionals from small and large information technology companies showed that 91% of them were attacked by some type of malware, 45% were ill-prepared for dedicated cyberattacks, and 17% lost financial data as a result of these attacks.

New malicious artifacts usually consist of versions of existing ones, as in [21], in which less than 10% of the malware in 2019 are new. Thus, characterizing them into families accelerates the identification of their behavior, of vital importance for allocating computational and human resources to promote a countermeasure as soon as possible. However, cases in which the malicious artifact configures a novelty in malware sets (rather than a variation of existing software) eludes this characterization. Thus, research initiatives must seek to automatically identify these novelties to accelerate the fight against these threats.

This research aims to evaluate the capacity of models to classify artifacts that configure novelties in malware families by using two common machine learning processes to detect novelties: Gaussian Mixture Models (GMM) and Support Vector Machines (SVM). This task entailed (a) defining a minimum set of characteristics to enable rapid detection; (b) proposing a method to select data sets; and (c) choosing criteria for model hyperparameters considering novelty detection. These constitute the main contributions of this study.

This study is structured as follows: Section 2 describes the theoretical foundations of novelty detection, the used machine learning models, and the publicly available malware. Section 3 reports the main studies on machine learning for malware analysis. Section 4 described the method to define the datasets and choose the most appropriate hyperparameters to fit the novelty detection models. Section 5 shows the experiments and the obtained results. Finally, Section 6 describes the conclusions and suggestions for future research.

2. Theoretical Framework

This section shows the theoretical foundations of malware analysis, anomaly and novelty detection, the used machine learning methods, and the dataset for the experiment.

2.1 Static and dynamic analysis

Malware analysis has two fundamental methods: the static and the dynamic approaches. Static analysis examines malware without running it, whereas dynamic analysis runs it [22]. Both techniques have advantages and disadvantages and can be used to complement analysis.

2.2 Novelty detection

According to [13], anomalies constitute data patterns that fail to fit a well-defined notion of normal behavior. Bringing this definition to the detection of malware, a malicious artifact that fails to fit into any known family is an anomaly. However, unprecedented malicious artifacts (anomalies) are quite common. Thus, following [13], this type of anomaly is defined as a novelty.

The use of one term or another exceeds mere semantics: the methods to identify anomalies presuppose anomalous samples (albeit in a very small quantity) in the training set. Novelty detection methods lack these anomalous samples for parametrization, making the detection of an unprecedented malware even more challenging. The anomaly detection problem can obviously use supervised machine learning methods, whereas the novelty detection problem commonly uses unsupervised machine learning methods.

2.3 Gaussian mixture models

The Gaussian mixture model (GMM) is a probabilistic model that assumes that all data points stem from a mixture of a finite number of Gaussian distributions with unknown parameters. Thus, a GMM-1 uses only one Gaussian distribution to model the distribution of data, whereas a GMM-32 uses 32 weighted Gaussian distributions. Traditional statistical techniques with Gaussian models to detect novelties may define their probability threshold in two ways: (i) as $\mu \pm 3\sigma$, in which μ is the mean and σ is the standard deviation and (ii), as per the Grubbs' test [13]. However, machine learning avoids establishing the threshold for novelties a priori.

Novelty detection by unsupervised learning mainly involves estimating densities. Clearly, prior knowledge of the density of the data points would enable us to solve any problem based on the data [14].

2.4 Support vector machine

Support vector machines (SVM) constitute a set of supervised learning methods to classify, regress, and detect anomalies. It can be specifically found for novelty detection in [14], which proposed a method to solve the following problem: once a dataset is extracted from an underlying probability distribution P, it is desired to estimate a subset S so the probability of a test sample from P lying outside S is equal to V. The use of SVM in this type of task is known as OneClassSVM. It has the advantage of avoiding assumptions about the form of the distribution of known data [15], i.e., it applies to any distribution. According to [23], the idea lies in finding a function that is positive for regions with high point density and negative for small densities.

A disadvantage of OneClassSVM refers to its discrete prediction whether the sample belongs to a modeled group or not. This limitation can be circumvented by changing OneClassSVM to enable the output of the SVM as conditional class probabilities, as per [16]. A probabilistic approach offers many advantages over the conventional method, including the ease of automatically selecting a probabilistic novelty threshold.

2.5 Database of malicious artifacts for novelty detection

Anomaly and novelty detection techniques based on machine learning clearly depend on databases with a representative volume of samples. Few such databases of malware are available to the public.

Also, according to [20], malicious artifacts for Windows represented, in August 2019, 74.49% of the existing malware universe, which motivated our choice for this platform.

3. Related studies

This section describes the main studies on machine learning and malware analysis.

The extensive use of machine learning to analyze malware includes several objectives, such as the detection of malicious artifacts [2, 3, 4, 5, 6], their variants [7, 8] and categories [9,10,12], among others.

The author in [2] proposed a new graph-based algorithm built from traces of dynamically collected instructions to classify artifacts into malicious or benign. Results shows that dynamic analysis, combined with n-grams, outperforms the antivirus tools at the time (2011). The authors in [3] proposed static and dynamic analyses of artifacts and similarity metrics with several kernels to detect malware. A weight is given to each kernel to find the weight/kernel combination with the best accuracy. Using this approach obtained 99.78% accurate results. The author in [4] used support vector machines to detect malware. The scope of that study totaled 398 samples, obtaining an accuracy from 94 to 95%. The authors in [5] use static and dynamic techniques to detect malicious artifacts. They used the C5.0 and random forest algorithms, which were implemented on the FAME framework. The obtained accuracy totaled 95.75% for binary classification and 93.02% for multiple categorization. An architecture to automatically analyze malicious artifacts (dynamic analysis) is proposed in [6]. The results obtained with random forest (based on ID3) showed an accuracy above 90%.

The author in [7] proposed a new use for hidden Markov models: sets of relevant opcodes are initially chosen, which are subjected to a hidden Markov model. Results improved from 8 to 42% when compared to studies with Hidden Markov models without opcode selection. In [8], a generic structure extracts structural information from malware as graphs of function calls that are encoded as function-level attributes. This approach is evaluated for 11 malware families, obtaining an accuracy of 86.67% in the detection of samples with previously chosen variants.

In [9], decision tree and machine support vector algorithms are used to categorize malicious artifacts into trojans, infectors, backdoors, and worms. Results obtained an accuracy above 98% for the decision tree algorithm and above 97% for the support vector machine algorithm. In [10], the authors propose two machine learning approaches to correctly classify the malicious artifacts in the Microsoft Malware Classification Challenge [11]. It extracted several artifact characteristics: referring to the binary content of the malicious artifact in the deconstructed file (disassembled). By an algorithm to select the best attributes, the system obtained a 99.76% accuracy. Finally, in [12], the authors used unsupervised machine learning (shared nearest neighbor) to classify and detect new malware families. Using a database of 20,000 samples, results show a 3% improvement using random forest and 18% using the Naive Bayes algorithm considering a combination of three attribute types in all cases (grayscale image and n-gram image attributes). In summary, these last articles show two different strategies: [10] classifies malware in existing families and [12] proposes classification and detection by an unsupervised approach.

Thus, despite the extensive research in malware analysis by machine learning, no analyzed article addresses novelty detection, the aim of this study.

4. Research development

This section describes the processes and methods to adjust the hyperparameters of machine learning models to detect malware novelties.

4.1 Analysis of the malware dataset

This study used the dataset provided by Microsoft and hosted on the Kaggle [11] platform to classify its malware families. Originally, two datasets are provided: a training dataset with 10,868 samples from known families and a test dataset with 10,873 samples from unknown families. This research only used the training subset since it was the only one with family labels for each sample. Table 1 shows the distribution of this training subset by family.

Table 1 - Distribution Of Malware Families On The Microsoft Database [11]

Order	Name of the family	Number of samples
1	Ramnit	1,541
2	Lollipop	2,478
3	Kelihos_ver3	2,942
4	Vundo	475
5	Simda	42
6	Tracur	751
7	Kelihos_ver1	398
8	Obfuscator,ACY	1,228
9	Gatak	1,013

The malware samples in this database are available in two formats: raw binary, which contains the hexadecimal representation of the binary content of the file; and the deconstructed code, the result of disassembling the original code (with the IDA tool [17, 10]). In both cases, the removal of the file header (PE header) made any attempt at dynamic analysis of the code impossible.

Since this research aims to define a classifier that can quickly identify novelties, only a minimum set of attributes from the raw binary data was considered based on two premises: (i) close-to-normal distribution and (ii) successful use in previous studies. Thus, four classes of attributes were chosen from [10]: entropy (ENT), metadata (MD1), Haralick (IMG1), and local binary patterns (IMG2).

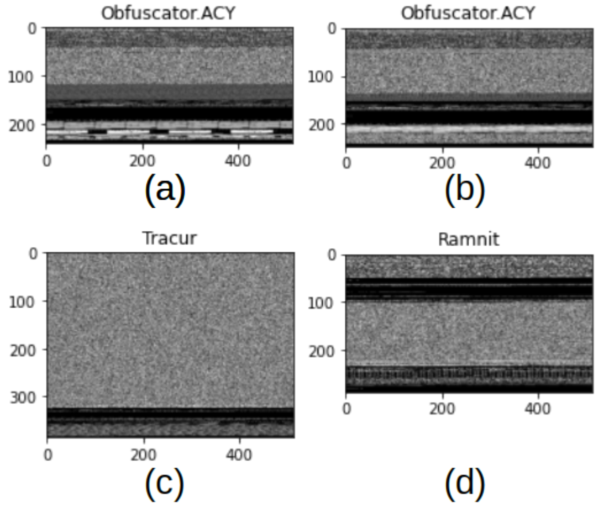
Entropy attributes (ENT) serve to measure the clutter in the raw binary file. Entropy is calculated by N sliding windows to represent the malware as a measure of entropy $E = E_i$ with $i = \{1, 2, 3 \ldots N\}$, in which E_i is the entropy measured in window i and N is the number of windows. Next, the statistics of the entropy vector obtained by the sliding window method were considered, i.e., the entropy was calculated for each 10,000-byte window and the statistical measures of the obtained quantiles, percentiles, mean, and variance of the distribution were then considered. The entropy of all malware bytes totaled 202 characteristics. The extracted metadata attributes (MD1) refer to file size and the address of the first byte string, totaling two characteristics.

An original way to represent a sample of malware refers to viewing them as a one-pixel grayscale in a 2D image [10]. Figure 1 shows a cutout of the grayscale image of (a) a Obfuscator.ACY malware, (b) another Obfuscator.ACY malware, (c) a Tracur malware, and (d) a Ramnit malware. (A) and (b) show that two malware of the same family have similar visual representations.

In total, two sets of attributes were extracted from the grayscale image of each sample that describe the textures: Haralick (IMG1), with 53 extracted traits; and the local binary standard attributes (IMG2), with 108 extracted characteristics, all of which are easy to process [10]. This extraction used Mahotas as its image processing and computer vision library [18].

Thus, the entropy (ENT), metadata (MD1), Haralick (IMG1), and local binary standard (IMG2) classes total 365 individual attributes. The source code of the extraction software for these attributes can be found in [10]. No initiatives were carried out to reduce dimensionality since these 365 attributes were selected from the 1,805 attributes in [10].

Fig. 1 - Representation of malware in their respective families by grayscale images.



4.2 Definition of the training and test sets

The definition of the training and testing sets entailed three steps to obtain suitable data subsets for novelty detection.

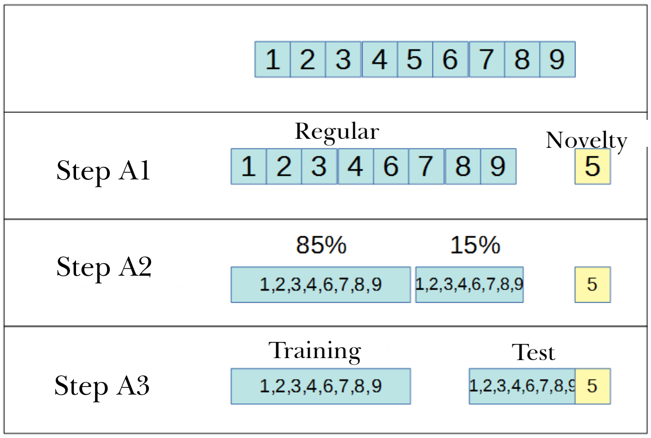
Initially, selecting the malware family to play the role of novelty proved to be challenging due to the absence of formal definition or method for this task to the best of our knowledge. It would be possible, for example, to choose both the family with the smallest number of samples and that with the greatest distance from the centroids of other families (considering that a set of attributes was selected). This research decided to adopt the family with the smallest number of samples as the novelty.

Thus, Step A1 chose samples belonging to family 5 (Simda) (Table 1). From this point on, this study refers to this family as the **novelty** class, grouping the other families into a single **regular** class.

Then, Step A2 divided the regular class into two groups, composed of 85 and 15% of all samples by random extractive partitioning, respectively. Finally, Step A3 defined the first group as the training set (85% of the regular class samples) and the test set as the union of the second (15% of the regular class samples) and the novelty classes (Simda).

As a result of this partitioning (Step 3), no samples of the novelty class belong to the training set. The ratio between the number of samples in the training and test group revolved around 6:1. Figure 2 shows the steps above, with the numbers 1 to 9 representing the nine malware families in Table 1.

Fig. 2 - Representation of the method to partition the training and test sets.



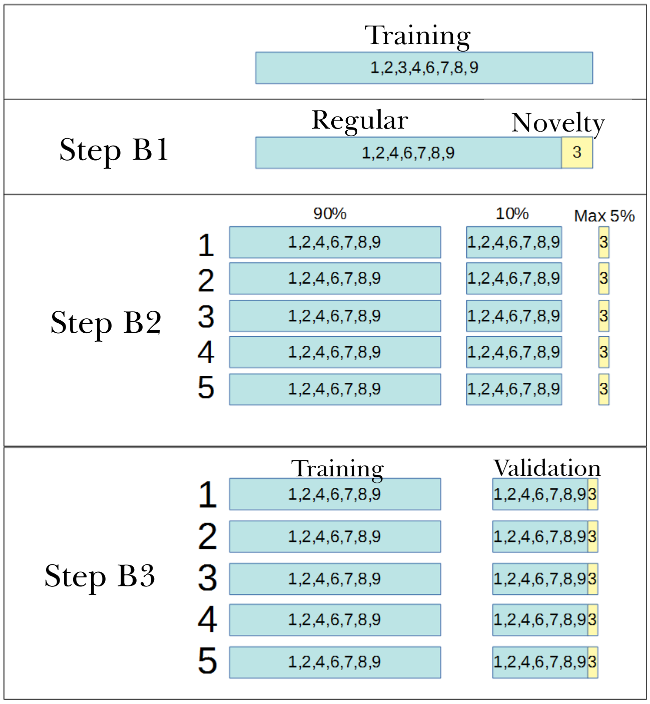
4.3 Definition of the training and validation set

As in Section 2.2, detecting a novelty differs from detecting an anomaly. While the latter includes samples to calibrate the machine learning process, the former offers no such possibility. Thus, the choice of the training and validation set should be adapted to novelty detection so the several families in the training set (Section 4.2) play the role of novelty class in the validation set. Thus, each family in the training set would play the role of novelty class in at least one iteration. As in the previous section, this process entailed three steps (Figure 3).

Step B1 chose a family from the training set to represent the novelty class, whereas the rest of the samples represented the regular class. In Figure 3, as an example, family 3 played the role of the novelty family. Step B2 divided the regular class into five equally sized subsets by random extractive partitioning. The idea of using five subsets is that, during the training and validation of the model, the performance will be the value of the average of the performances in each subset. Next, each subset was divided into two groups: the first with 90% of the samples and the second with 10% of the samples.

Finally, Step B3 designated the first group (90% of the samples) as the training set. On the other hand, the validation set consists of the union of the second group (10% of the samples) with a set of samples from the novelty family defined in Step B1 (family 3 in Figure 3), randomly limiting its volume to 5% of the total amount of the second group. Therefore, the **validation** set has regular and novelty samples. Again, this study sought a ratio between training samples and validation close to 6:1.

Fig. 3 - Method to select the training and validation sets for the novelty detection task. Family 3 served as the novelty class.



As each family will play the role of novelty, a total of eight experiments will be carried out to define the training/validation sets, as described above.

4.4 Classifier parameter selection

This study adjusted the hyperparameters of this model by simultaneously considering the result in the eight training/validation experiments in Section 4.3. Thus, this research created a criterion to accept a certain set of hyperparameters of a model.

Initially, P_i is defined as a set of hyperparameters in which i = {1, 2, 3 … n} and n is the number of attempts. To evaluate the learning process based on these hyperparameters, three metrics were established:

- ***Training Error Percentage*** **(TEP)**: Error in the training set, calculated by dividing the number of predictions outside the regular class and the total number of elements of the training class. The training dataset only has regular class elements.
- ***Validation Error Percentage*** **(VEP)**: Error in the validation set, calculated by dividing the number of errors in the prediction of validation set samples by the total number of validation set samples. The validation set has both regular and novelty samples.
- ***Novelty Error Percentage*** **(NEP)**: The novelty error is defined by dividing the number of errors in predicting only the validation set novelty class samples by the total number of novelty class samples in the validation set.

To accept a set of parameters P_i, the following criteria were defined:

$$(TEP < 30\%) \wedge (VEP < 30\%) \wedge (NEP < 40\%)$$

This criterion must be validated considering the average of the TEP, VEP, and NEP for each of the eight experiments in Section 4.2. Note that the percentage thresholds above were established after repeated attempts.

Case P_i meets the established criteria, it is selected. Otherwise, a new set of parameters P_{i+1} is set and the parameter selection process is restarted. This study showed no intention to choose the best possible model. It aimed to choose the one with hyperparameters set to P_i, capable of being used in all eight experiments. The adopted assumption is that the criterion for choosing hyperparameters P_i can generalize the final model and obtain a good result for unprecedented samples.

Finally, with the chosen set of hyperparameters P_i, a model with the entire original training set (families 1, 2, 3, 4, 6, 7, 8, and 9) is trained. Then, the test set (family 5) is introduced to this model. To estimate the performance of the model in the test suite, the following metrics are defined:

***Training Error Percentage* (TEP)**: Error in the training set, calculated by dividing the number of predictions outside the regular class and the total number of elements of the training class.

***Test Error Percentage* (TEPer)**: Error in the test set, calculated by dividing the number of errors in the prediction of training set samples by the total number of validation set samples. The test set features regular samples and novelty samples from family 5.

***Novelty Error Percentage* (NEP)**: novelty error, i.e., family 5, defined by dividing the number of errors in predicting only the test set novelty class samples by the total number of novelty class samples in the test set.

5. Results

This section describes the results of the hyperparameters that were chosen in the previous section for the support vector machine and Gaussian mixture machine models. As in sections 2.3 and 2.4, this study used the support vector machine and Gaussian mixture models to detect anomalies and novelties despite their rather different approaches, justifying the choice of these models to evaluate their behavior in a specific malware case. Note that this research evaluated their performance both in samples from previously identified families and in unprecedented samples.

5.1 Support Vector Machine Model

This study decided to use the RBF kernel for the SVM model, as per [19]. The rest of the hyperparameters that can be modified are (i) outliers (which define the number of anomalies in the training set) and (ii) gamma (γ) (which defines the chosen kernel). As a probabilistic implementation of OneClassSVM — as in [16] —, this study chose a conventional version of this algorithm.

Then, the algorithm obtained its best results with 13% outliers and a 0.01 γ, following the method in Section 4.4. Table 2 shows the results of the metrics for the various validation sets. In each row, the metric is averaged for five subsets, in which one of the eight families plays the role of novelty (see Section 4.3).

The parameter selection phase finds that the worst result occurs when family 6 (Tracur, with a 74.18% novelty error percentage) is novel. This may stem from the Tracur malware family differing from the others since they can primarily redirect online research, download and execute files (including other malware), and allow for command-and-control actions on the infected computer. Thus, Tracur family artifacts differ from the others for their greater capacity for malicious activities. As a result, the classifier may have erroneously classified them into other families.

Since the average metrics result meets the criteria in Section 4.4, this study chose this set of hyperparameters to create a model that can consider the entire training set in Section 4.2.

Table 2 - Training and validation error for SVM

Name of the family	TEP	VEP	NEP
Ramnit	13.14%	12.48%	6.08%
Lollipop	13.10%	13.97%	7.08%
Kelihos_ver3	13.21%	12.46%	0.00%
Vundo	13.17%	11.63%	1.41%
Tracur	13.18%	33.99%	74.18%
Kelihos_ver1	13.14%	12.48%	6.08%
Obfuscator,ACY	13.27%	12.61%	3.92%
Gatak	13.19%	16.44%	15.73%
Mean:	**13.18%**	**15.86%**	**13.43%**

Next, Table 3 shows the prediction in the test set and the result of the metrics in this set. The SVM model solidly detected samples from the regular set but shows a decreased performance in novelty detection.

Table 3 - SVM test error

Name of the family	TEP	PET	NEP
Simda	13.17%	17.29%	26.19%

5.2 Gaussian mixtures model

Initially, a Gaussian mixture model with only one Gaussian was used, which is equivalent to a simple Gaussian model to detect novelties. Several attempts were made to find the error threshold (ε) using the approach described in Section 4.4, in which values below this threshold indicate that the sample belongs outside the regular class.

After several experiments, an error threshold was reached $\varepsilon = \mu - 0.98\sigma$, in which μ is the average of the weighted logarithms of the probabilities for each training sample and σ is the standard deviation of the weighted logarithms of the probabilities for each training sample. Table 4 shows the results of the metrics for the various validation sets. Each row shows the metric average for the five subsets in which each of the eight families played the role of novelty.

Table 4 - Error training and validation for GMM-1

Name of the family	TEP	VEP	NEP
Ramnit	31.66%	30.91%	40.50%
Lollipop	21.64%	21.66%	10.97%
Kelihos_ver3	31.52%	32.51%	0.01%
Vundo	28.70%	29.04%	98.12%
Tracur	33.76%	35.16%	38.54%
Kelihos_ver1	28.30%	28.72%	13.31%
Obfuscator.ACY	29.40%	28.79%	88.71%
Gatak	26.25%	27.37%	24.37%
Mean:	**28.95%**	**18.97%**	**39.34%**

continues

The average result of the metrics met the criteria in Section 4.4. The families that obtained the worst result were Vundo and Obfuscator.ACY. A possible reason for this error is that Gaussian distributions are unable to approximate the distribution of the attributes of these families. Thus, this set of hyperparameters were selected, creating a model that considers the entire training set, and then the prediction was performed in the set of tests. Table 5 shows the results for the test set.

Table 5 - Test error for GMM-1

Name of the family	TEP	VEP	NEP
Simda	27.79%	27.91%	19.05%

It can be preliminarily concluded that the Gaussian model could detect novelties according to the criteria to select parameters.

In any case, several experiments were carried out varying the number of Gaussians in the GMM. Models with 2, 4, 8, 16, 32, and 64 Gaussians were tested. From 32 Gaussians onward no improvement in training/validation occurred. Thus, Table 6 shows the results of the metrics for the various validation sets for 32 Gaussians.

Table 6 - Training error for GMM-32

Name of the family	TEP	VEP	NEP
Ramnit	16.00%	30.25%	0.08%
Lollipop	8.57%	20.25%	14.98%
Kelihos_ver3	8.14%	21.69%	0.19%
Vundo	8.54%	18.86%	97.92%
Tracur	11.27%	21.09%	29.97%
Kelihos_ver1	18.04%	27.26%	2.84%
Obfuscator.ACY	9.18%	18.84%	88.77%
Gatak	7.87%	17.74%	26.40%
Mean:	**10.95%**	**22.00%**	**39.33%**

Again, a large variation in the percentage of novelty error occurred, as the case of GMM-1. The average error percentage of the model with only one Gaussian resembles the 32-Gaussian model.

The process of choosing and selecting hyperparameters for the models with 2, 4, 8, 16, and 32 Gaussians was carried out and the set of tests was applied to each of these models. Table 7 shows these results.

Table 7 - Test error for 2,4,8,16, and 32 GMM

GMM	Name of the family	TEP	PET	NEP
2	Simda	17.09%	18.13%	30.95%
4	Simda	27.22%	30.79%	28.57%
8	Simda	26.58%	28.09%	28.57%
16	Simda	21.00%	27.01%	28.57%
32	Simda	15.52%	25.15%	26.19%

If we only consider the number of errors in the detection of novelties, GMM-2 showed 13 errors; GMM-4, GMM-8, and GMM-16, 12 (28.57%); and GMM-32, 11 errors (26.19%). Thus, model training and testing evince the slight outperformance of the GMM-32 model. This was expected since the greater the number of Gaussians, the better the model fits the distribution of characteristics.

It is interesting to note that the model with a single Gaussian better detected novelties, i.e., a model that best fits the attribute distribution of the regular set fails to necessarily offer the best model for novelty detection. Also, both models (GMM-1 and GMM-32), during parameter selection, showed the worst results for classes 4 (Vundo) and 8 (Gatak) as novelties.

SVM, GMM-1, and GMM-32 results show that all models adequately detected novelties according to the criteria for choosing parameters and the difficulty of detecting unprecedented samples (novelty).

In other words, SVM obtaining 26.19% NEP and GMM-1, a 19.05% NEP equals claiming that SVM and GMM-1 detected 73.81 and 80.95% of all novelties, respectively, which is far better than no detection.

The SVM model seems to have a greater potential to detect novelties than the GMM model but both showed compatible results.

6. Conclusion

Novelty detection is a complex task as it lacks an obvious relation between parameter selection (by a training and validation set) and testing results (by a test set). Thus, this study used two classifier algorithms (GMM and SVM) that satisfactorily detected novelties. Defining training, validation, and testing sets and criteria for accepting hyperparameters was essential for these results. This study employed an adequate method to choose parameters for each model, although more in-depth studies are needed since some experiments showed poor results during selection (the Vundo and Gatak novelty classes for GMM and the Tracur novelty class for SVM).

As a suggestion for future research, this study proposes other attributes, a further development of poor results in some experiments during parameter selection, datasets with larger numbers of malware families, and a probabilistic novelty SVM classifier, as suggested in [16].

References

[1] – KASPERKY LABS. Ready… or Not? Balancing future opportunities with future risks. Moscou: Kaspersky Labs, 2017. Disponível em: https://media.kaspersky.com/documents/business/brfwn/en/The-Kaspersky-Lab-Global-IT-Risk-Report_Kaspersky-Endpoint-Security-report.pdf. Acesso em: dd mmm aaaa.

[2] - ANDERSON, B.; QUIST, D.; NEIL, J.; STORLIE, C.; LANE, T. Graph-based malware detection using dynamic analysis. Journal in Computer Virology, v. 7, n. 4, p. 247-258, 2011. DOI: 10.1007/s11416-011-0152-x

[3] – ANDERSON, B.; STORLIE, C.; LANE, T. Improving malware classification: Bridging the static/dynamic gap. In: Proceedings of the workshop on security and artificial intelligence, 5., [S.l.: s.n.], 2012. p. 3-14.

[4] - KRUCZKOWSKI, M.; SZYNKIEWICZ, E. N. Support vector machine for malware analysis and classification. In: Proceedings Of The 2014 Ieee/Wic/Acm International joint conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). Washington, DC: IEEE Computer Society, 2014. p. 415-420. DOI: 10.1109/WI-IAT.2014.127

[5] - MANGIALARDO, R. J.; DUARTE, J. C. Integrating static and dynamic malware analysis using machine learning. IEEE Latin America Transactions, v. 13, n. 9, p. 3080–3087, 2015.

[6] - ANDRADE, C. A. B.; MELLO, C. G.; DUARTE, J. C. Malware automatic analysis. In: 2013 BRICS Congress On Computational Intelligence And 11th Brazilian Congress on Computational Intelligence. [S.l.: s.n.], 2013. p. 681-686.

[7] - GHARACHEH, M.; DERHAMI, V.; HASHEMI, S.; HAZRATI FARD, S. M. Detection of metamorphic malware based on hmm: A hierarchical approach. International Journal of Intelligent Systems and Applications, v. 8, n. 4, p. 18-25, 2016.

[8] - KONG, D.; YAN, G. Discriminant malware distance learning on structural information for automated malware classification. In: Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining, 19., New York: ACM, 2013. p. 1357-1365. DOI: 10.1145/2487575.2488219

[9] - CHEN, Z.; ROUSSOPOULOS, M.; LIANG, Z.; ZHANG, Y.; CHEN, Z.; DELIS, A. Malware characteristics and threats on the internet ecosystem. Journal of Systems and Software, v. 85, n. 7, p. 1650-1672, 2012. DOI: 10.1016/j.jss.2012.02.015

[10] - AHMADI, M.; ULYANOV, D.; SEMENOV, S.; TROFIMOV, M.; GIACINTO, G. Novel feature extraction, selection and fusion for effective malware family classification. In: ACM Conference on Data and Application Security and Privacy (CODASPY), 16., 2016. p. 183-194

[11] – PANCONESI, A.; MARIAN; CUKIERSKI, W. Microsoft Malware Classification Challenge (BIG 2015). [S. l.]: Kaggle, 2015. Disponível em: https://www.kaggle.com/c/malware-classification. Acesso em: 17 set. 2018.

[12] - LIU, L.; WANG, B.-S.; YU, B.; ZHONG, Q.-X. Automatic malware classification and new malware detection using machine learning. Frontiers of Information Technology & Electronic Engineering, v. 18, p. 1336-1347.

[13] CHANDOLA, V.; ARINDAM, B.; VIPIN, K. Anomaly Detection: A Survey. ACM Computing Surveys, v. 41, n. 3, p. 1-58.

[14] – SOLLA, S. A.; LEEN, T. K.; MÜLLER, K.-R. Support vector method for novelty detection. In: Advances in Neural Information Processing Systems, 12., Cambridge: MIT Press, 2000. p. 582-588.

[15] - SHAWE-TAYLOR J.; ŽLIČAR, B. Novelty Detection with One-Class Support Vector Machines. In: MORLINI, I.; MINERVA, T.; VICHI, M. (Ed.). Advances in Statistical Models for Data Analysis. Berlin: Springer, 2015. p. 231-257.

[16] – CLIFTON, L.; CLIFTON, D. A.; ZHANG, Y.; WATKINSON, P.; TARASSENKO, L.; YIN, H. Probabilistic novelty detection with support vector machines. IEEE Transactions on Reliability, v. 63, n. 2, p. 455-467, 2014.

[17] - RONEN, R.; RADU, M.; FEUERSTEIN, C.; YOM-TOV, E.; AH-MADI, M. Microsoft malware classification challenge. [S. l.: s. n.], 2018.

[18] - COELHO, L. P. 2013. Mahotas: Open source software for scriptable computer vision. Journal of Open Research Software, v. 1, n. 1, 2013.

[19] - CHANG, C. C. A practical guide to support vector classification, Technical report, n. 5. Taipei: National Taiwan University, 2005.

[20] – Heightened threat scenario: all the facts in the AV-TEST Security Report 2018/2019. Magdeburg. AVTEST – The Independent IT-Security Institute, 2019. Disponível em: https://www.av-test.org/en/news/heightened-threat-scenario-all-the-facts-in-the-av-test-security-report-2018-2019/. Acesso em: 23 ago. 2019.

[21] – Malware. AVTEST – The Independent IT-Security Institute, 2019. Disponível em: https://www.av-test.org/en/statistics/malware//. Acesso em: 10 set. 2019.
[22] SIKORSKI, M.; HONIG, A. Practical malware analysis: the hands-on guide to dissecting malicious software. San Francisco: No Starch Press, 2012.
[23] SCHÖLKOPF, B.; WILLIAMSON, R.; SMOLA, A.; SHAWE-TAYLOR, J.; PLATT, J. Support vector method for novelty detection. NIPS, v. 12. 1999.

VOL.40 Nº4 2023
https://doi.org/10.22491/rmct.v40i4.12301.en

Analysis and characterization of titanium surfaces in different stages of the silanization process

André Aguiar Marques[1], andre.marques@ime.eb.br, ORCID: 0000-0003-1511-4038
Marcelo Henrique Prado da Silva[1], marceloprado@ime.eb.br, ORCID: 0000-0002-1182-5345
Álvaro José Boareto Mendes[1], boareto@ime.eb.br, ORCID: 0000-0003-1432-9045
Carlos Nelson Elias[1], elias@ime.eb.br, ORCID: 0000-0002-7560-6926
[1]Instituto Militar de Engenharia, Rio de Janeiro, RJ, Brasil

ABSTRACT: The main methods of physical, chemical and topographical modification of the surfaces of metallic implants have already been widely studied and documented. With advances in tissue engineering, new approaches to replacing or repairing organs and tissues are made possible by a new era of bioactive materials with the combination of biomaterials with biomolecules. The association of biomolecules with titanium surfaces can be carried out through the silanization method, which enables a covalent bond between the biomolecule and the silane anchored on the previously functionalized titanium surface. This study proposes the analysis, through scanning electron microscopy and characterization in laser interferometry, of titanium surfaces in three stages that make up the silanization process and precede the association of the biomolecule. For this, Grade 4 titanium discs with 6 mm in diameter and 2 mm in thickness were used and studied after the acid etching and alkaline etching stages and after the incorporation of silane. The titanium surfaces of the discs from the three mentioned stages of the preparation process were analyzed and characterized, comparing the uniformity standard and roughness and wettability parameters found. The results showed that all stages of the treatment maintained a micro-rough and uniform surface with adequate roughness parameters compatible with the parameters considered ideal. However, there was an important variation regarding wettability in the groups studied.

KEYWORDS: *Titanium, biomolecule, silanization, roughness, wettability.*

RESUMO: As principais modalidades de modificação física, química e topográfica das superfícies dos implantes metálicos já são amplamente estudadas e documentadas. Com os avanços da engenharia tecidual, novas abordagens de substituição ou reparo de órgãos e tecidos são possibilitadas por uma nova era de materiais bioativos com a combinação de biomateriais com biomoléculas. A associação de biomoléculas às superfícies de titânio pode ser realizada por meio do método de silanização, que viabiliza uma ligação covalente entre a biomolécula e o silano ancorado na superfície de titânio previamente funcionalizada. Este estudo propõe a análise, através de microscopia eletrônica de varredura e caracterização em interferometria a laser, de superfícies de titânio em três etapas que compõem o processo de silanização e precedem a associação da biomolécula. Para isso, foram utilizados discos de titânio Grau 4 com 6 mm de diâmetro e 2 mm de espessura e estudadas após as etapas de condicionamento ácido, de condicionamento alcalino e após a incorporação do silano. As superfícies de titânio dos discos das três etapas citadas do processo de preparação foram analisadas e caracterizadas, comparando o padrão de uniformidade e parâmetros de rugosidade e molhabilidade encontrados. Os resultados mostraram que todas as etapas do tratamento mantiveram uma superfície microrrugosa e uniforme, com parâmetros de rugosidade adequados e compatíveis com os parâmetros considerados ideais. No entanto, houve uma variação importante quanto à molhabilidade nos grupos estudados.

PALAVRAS-CHAVE: Titânio, biomolécula, silanização, rugosidade, molhabilidade.

1. Introduction

Although the success rate of titanium implants is in the order of 95-98%, new surface modification techniques are continuously developed to improve roughness, wettability, adhesion, cell fixation, cell proliferation, and osseointegration [1]. The treatments aim to obtain the appropriate roughness of the implants to increase bone anchoring and improve the biocompatibility of the implants [2]. At the same time, changes in the oxide layer of metallic surfaces, such as titanium, and surface functionalization techniques with bioactive materials have been investigated and developed [3]. The functionalization of metal surfaces is very relevant in the area of biomaterials, since it is able to control the wettability of the material [4], surface energy [5], protein adsorption [6], drug release [7], and the interaction of cells with the implant [8].

Surface functionalization by silanes is one of the most commonly used methods to prepare monolayers. The main advantage of using silanes on metal surfaces is the rapid formation of a covalent bond between the substrate and the anchoring group. This covalent bond stabilizes the monolayer and allows subsequent chemical modifications, such as the association of biomolecules, without compromising the integrity of the monolayer [9].

Faced with the possibility of conferring antimicrobial activity to the titanium surface, or even some level of bioactivity in a metal initially considered inert, this study performed field emission gun scanning electron microscope analysis (FEG-SEM) and laser interferometry characterization of titanium surfaces in three different stages of the silanization process, to analyze and characterize the uniformity pattern and roughness parameters found.

2. Literature review

several surface modifications of metal implants have been developed using subtractive and additive methods. Among these methods, the most commonly used are particle blasting, acid etching, and anodizing. These techniques increase roughness and improve clinical success, with faster healing rates and potentially shorter time intervals for loading. However, each procedure generates a rough surface with slightly different topographic characteristics, even when they have equal values in their arithmetic means of roughness (Ra). Moreover, as a result of these modifications, the wettability and chemical characteristics of the implant surface are also often altered and can dramatically change the initial cellular response to an implanted material [10].

Most of the studies available in the literature that analyze the influence of roughness on osseointegration conventionally use only Ra values. Some studies correlate the Ra value with other properties of surfaces, such as wettability, *in vitro* cell adhesion, and protein adsorption. In addition to the fact that the choice of Ra as a parameter for roughness analysis is not justified in these studies, the isolated evaluation of an implant roughness parameter is not the ideal analysis for its characterization, since surfaces with similar values of Ra may present different morphologies. Therefore, there is a need to correlate the various parameters that characterize roughness, as well as analyze interference with other surface properties of implants [11].

The parameters for roughness analysis and characterization determine the implant surface. Nevertheless, the specific role of each roughness parameter in osseointegration is not yet conclusive [11]. A better evaluation of roughness and cell adhesion can be obtained by associating the values of at least one height parameter, one spatial parameter, and one hybrid parameter [12]. Surface roughness can be divided into three levels depending on the scale of the features: macro, micro, and nano-sized topologies [13]. The macro level is defined for topographic features as being in the range of millimeters to tens of microns. This scale is directly related to the implant geometry, with threaded screws and macroporous surface treatments resulting in surface roughness greater than 10 μm. The high roughness results in mechanical interlocking between the implant surface and the adjacent bone. However, a big risk with high surface roughness may be an increase in peri-implantitis [14].

The micrometer roughness of dental implants is in the range of 1-10 μm. This roughness range maximizes the interlocking between the mineralized bone and the implant surface. On these surfaces, the irregularities allow osteogenic cells to join and deposit bone, producing the bone-implant interface; therefore, it is considered that microrugosities act at the cellular level of osseointegration [15].

Surface profiles in the nanometer range play an important role in protein adsorption, osteoblast cell adhesion, and the rate of osseointegration. However, reproducible surface roughness in the nanometer range is difficult to produce with chemical treatments. Furthermore, optimal surface nano-topography for selective protein adsorption leading to osteoblast cell adhesion and rapid bone apposition is still unknown [14].

Several methods have been developed to create a rough surface and improve the osseointegration of titanium dental implants. These methods use plasma spraying of titanium, sandblasting with ceramic particles, acid etching, and anodizing [10, 14].

Surface modification techniques can be used individually or combined, and can be classified into three categories: physical, chemical, and biological [2].

Nano-rough surfaces with characteristics similar to those of cell membrane receptors and proteins play a crucial role in improving implant performance and osseointegration. In a nanoscale condition, cell membrane receptors, integrins, and proteins are involved and improve the overall quality of osseointegration and other biological reactions between bone and implant [16].

Although metallic biomaterials may differ in diverse physical and chemical properties, many of them share the potential for surface functionalization by the reactivity of surface bonded -OH groups as anchor points for the formation of densely packed monolayers. Activation can be achieved by wet etching, dry etching, and plasma activation. The surface terminated by hydroxyl groups makes it possible to bond with other molecules by condensation reactions [9].

Several functional groups react with the terminal hydroxyl, leading to the formation of layers known as self-assembled monolayers (SAMs) or self-organizing monolayers. By specific chemical interactions, organic molecules can spontaneously organize themselves on various types of surface [17].

The preparation of silane monolayers from solution is currently standardized to more reproducibly and rapidly obtain silane SAMs by immersing a metal in a solution of the precursor APTES (3-aminopropyl) triethoxysilane at different concentrations and temperatures. APTES is commonly used to obtain amine-terminated surfaces that are applied to promote protein adhesion and cell growth in biological implants [9].

The reaction has the potential to form a modified surface or work as an intermediate in binding organic ligands to titanium surfaces [4].

Materials and methods

the samples were prepared jointly in the Biomaterials Laboratory – SE/8 and Biotechnological Processes Laboratory – SE/5 of the Military Institute of Engineering, RJ.

A total of 15 samples were divided into three groups with five grade 4 titanium discs, measuring 6X2 mm and initially with polished surface. The groups formed were: titanium after the acid etching step (Ti-Etc), titanium after the alkaline etching step (Ti-Alk), and titanium after the silanization step (Ti--Sil).

All discs were subjected to the same acid etching with a solution composed of H_2SO_4 / HF / HCl to form a uniform microrough surface similar to the microtexture of osseointegrable implants available on the market, this being, therefore, the initial working surface for this study. After this treatment, the discs were washed with distilled water, dried in N2.

The samples of the Ti-Alk and Ti-Sil groups were activated with alkaline treatment by immersion in 20 ml of solution with 5M NaOH for 8h at a temperature of 60°C for the formation of the hydroxyl functional groups (OH^-) required for covalent bonding between titanium and silane.

After this step, they were washed with distilled water for 30 min and dried individually in N2. Although this step makes it possible to bond the silane on the titanium surface, it has as a possible consequence the formation of an amorphous sodium titanate layer on the metal surface, which, if immersed in fluid with ideal ionic conditions, may undergo crystallization in apatite, thus conferring some level of bioactivity to the material.

For the silanization procedure, a silane, a base, and a solvent are required. The silane chosen was CPTES of 3(chloropropyl)-triethoxysilane, due to the presence of the terminal group chloro in its organofunctional portion, which facilitates the interaction of silane with the peptide at a pH of 11. Anhydrous pentane was the chosen solvent and N,N-diisopropylethylamine (DIPEA) was the chosen base. All reagents were purchased from Sigma-Aldrich Brasil.

For creating the Ti-Sil group, five previously activated samples were immersed in a solution containing 7 ml of anhydrous pentane, 1.2 ml of 3(chloropropyl)--triethoxysilane, and 0.6 ml of the base N,N-diisopropylethylamine (DIPEA) under a saturated N2 atmosphere for 1h. At the end of this period, the samples were washed three times with distilled water and dried with nitrogen.

Samples from the three groups were analyzed by scanning electron microscopy (SEM) and characterized in Zygo laser interferometry at the Biomaterials Laboratory of the Military Institute of Engineering (IME).

Wettability was quantified by measuring the contact angle with distilled water. In this study, the FTA100 goniometer (First Ten Angstroms, Portsmouth, VA, USA) was used. The smaller the contact angle, the higher the wettability. Reducing the contact angle indicates that the surface is more hydrophilic.

3. Results and discussion

after acid etching with the solution composed of H_2SO_4 / HF / HCl, all discs formed a microrough surface with Ra parameter approximately between 1 and 2 μm, similar to the microtexture of osseointegrable implants available on the market, as verified in **Figures 1 and 2**, which confirmed the uniformity of the treatment carried out throughout their entire length.

The surface of the samples activated by etching with NaOH (**Figure 3**) and samples subjected to silane incorporation (**Figure 4**) were also analyzed by FEG-SEM and characterized by Zygo laser interferometer, confirming the uniformity of surface microtexture and roughness parameters achieved after acid etching.

The FEG-SEM analysis of the Ti-Alk and Ti-Sil subgroups in greater magnification, especially in 40000X (**Figures 5 and 6**), showed surfaces with nanometric characteristics throughout the surface, which may mean better protein adsorption and an improvement in implant osseointegration.

Fig. 1 - Micrograph (FEG-SEM) of the surface of the Ti-Etc group titanium sample after acid etching. 500X, 1000X, 2500X, 5000X magnification.

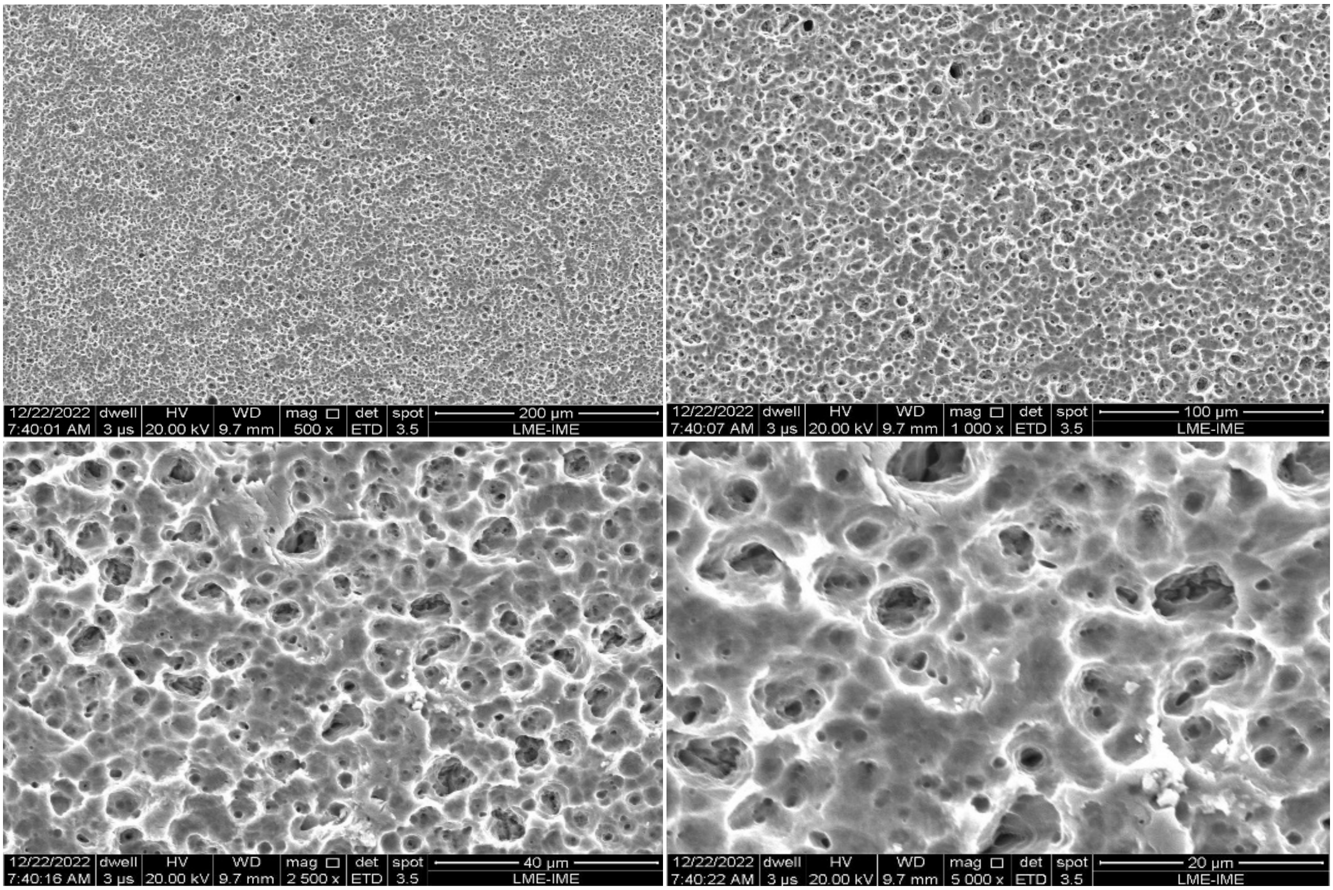


Fig. 2 - Micrograph (FEG-SEM) of the surface of the Ti-Etc subgroup titanium sample after acid etching. 10000X, 40000X magnification.

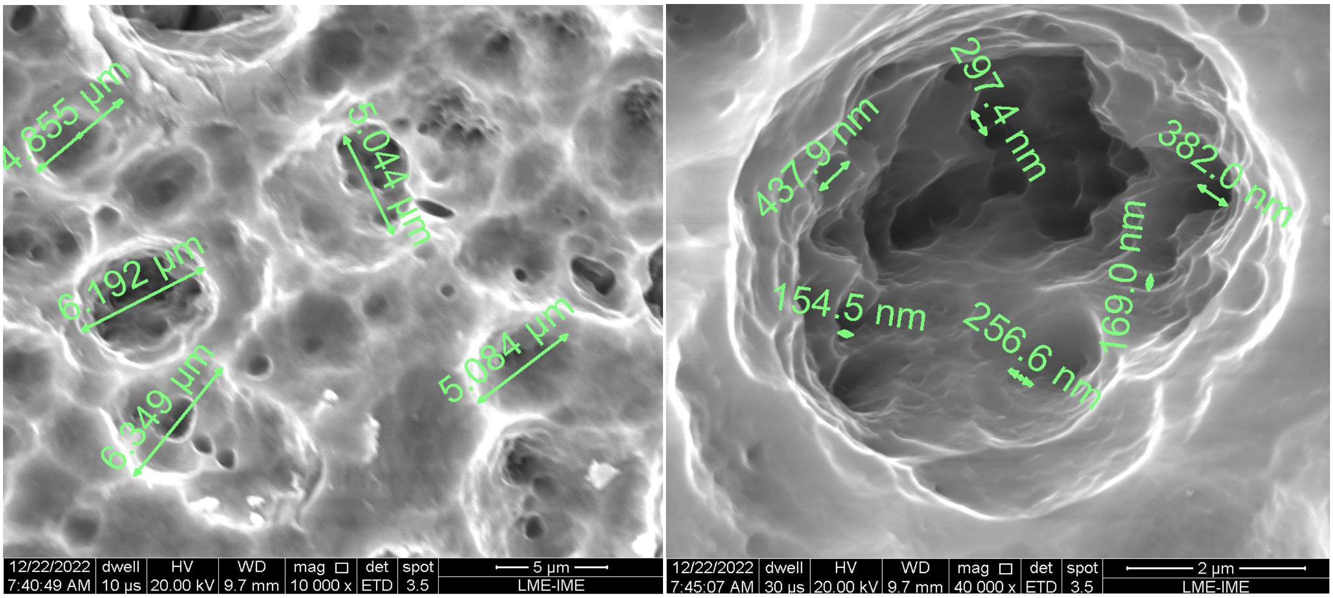


Fig. 3 - Micrograph (FEG-SEM) of the surface of the Ti-Alk group titanium sample after acid etching. 500X, 1000X, 2500X, 5000X magnification.

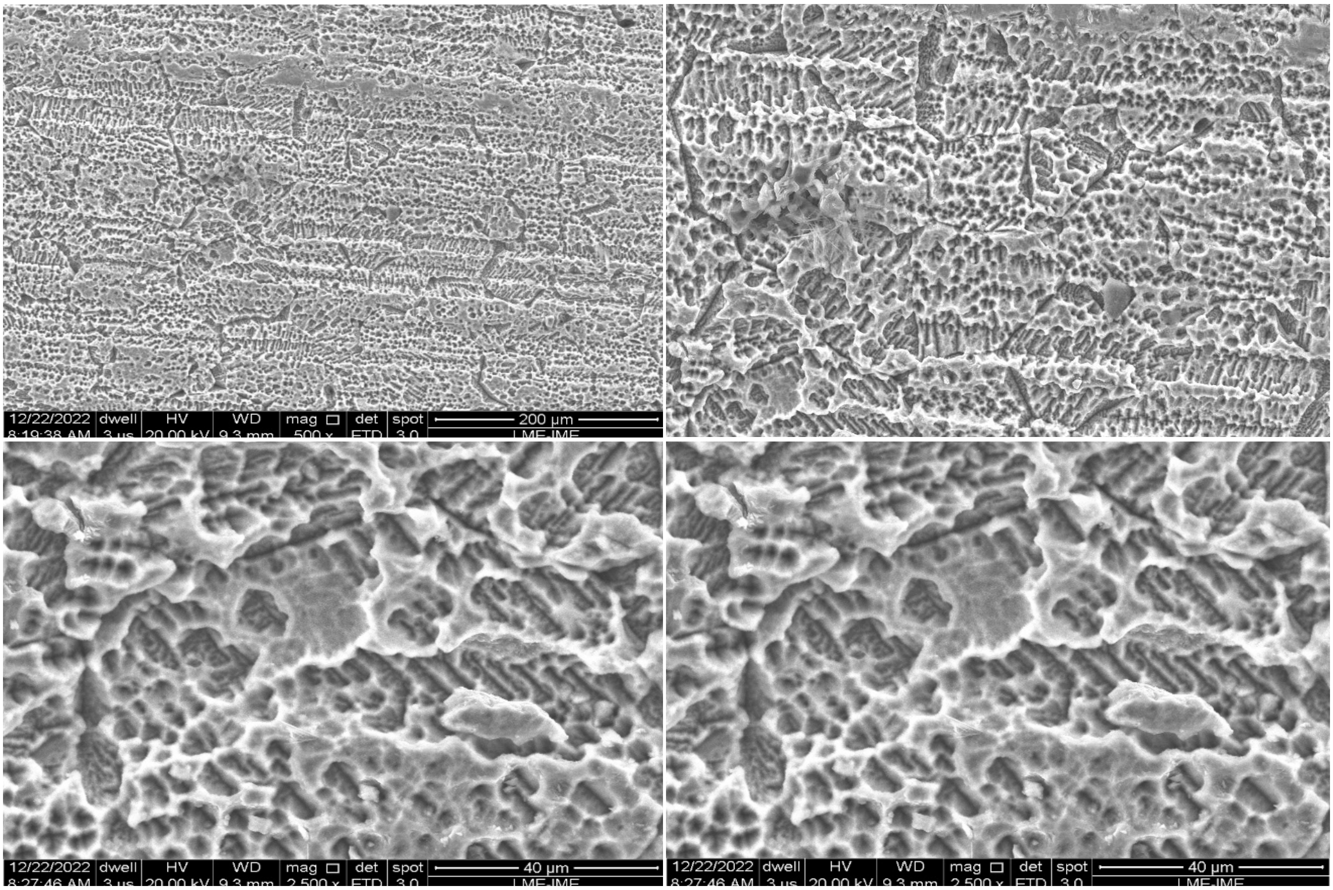


Fig. 4 - Micrograph (FEG-SEM) of the surface of the Ti-Sil group titanium sample after silanization. FEG--SEM. 500X, 1000X, 2500X, 5000X magnification.

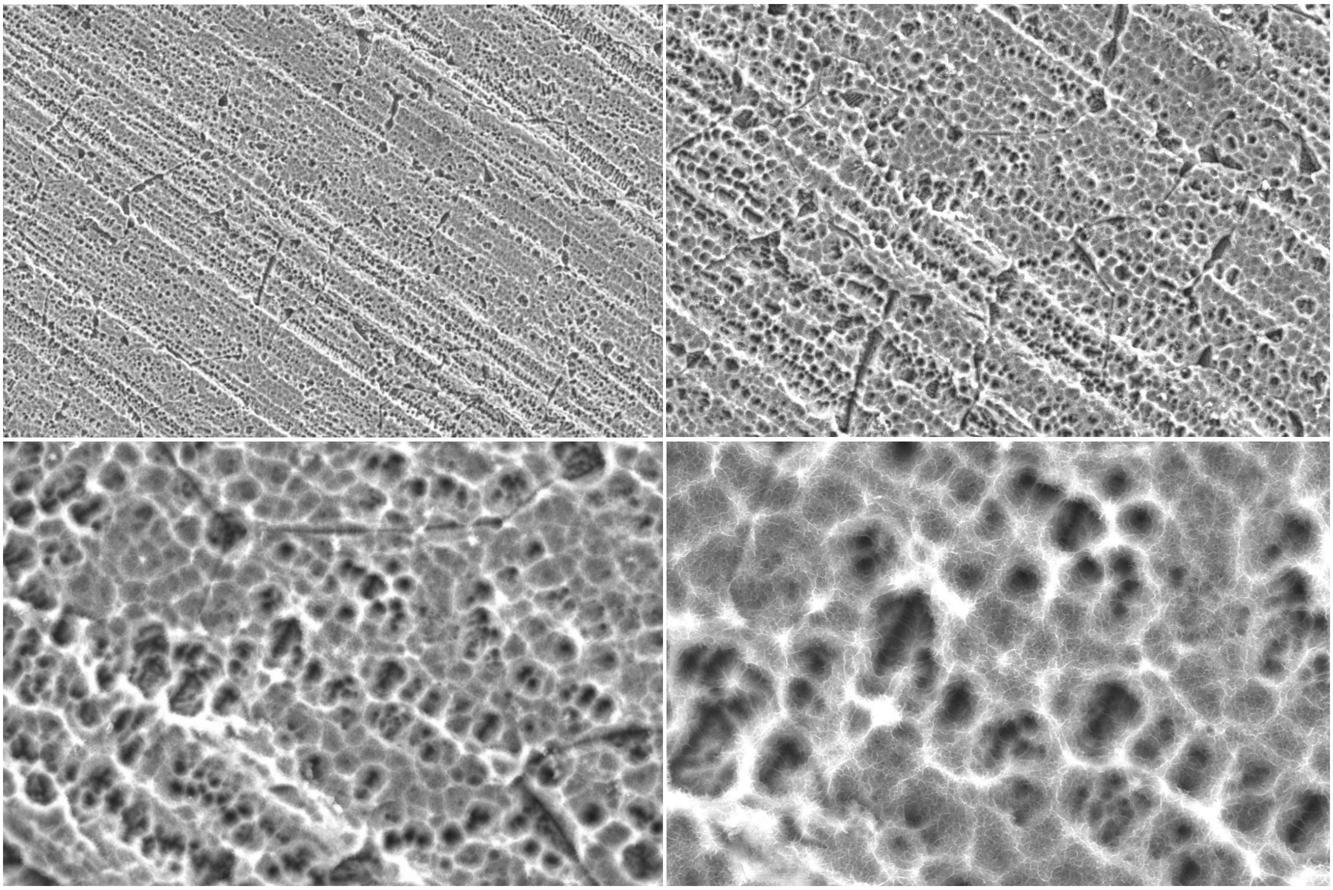


Fig. 5 - Micrograph (FEG-SEM) of the surface of the Ti-Sil group titanium sample after alkaline etching. FEG-SEM. 10000X, 40000X magnification.

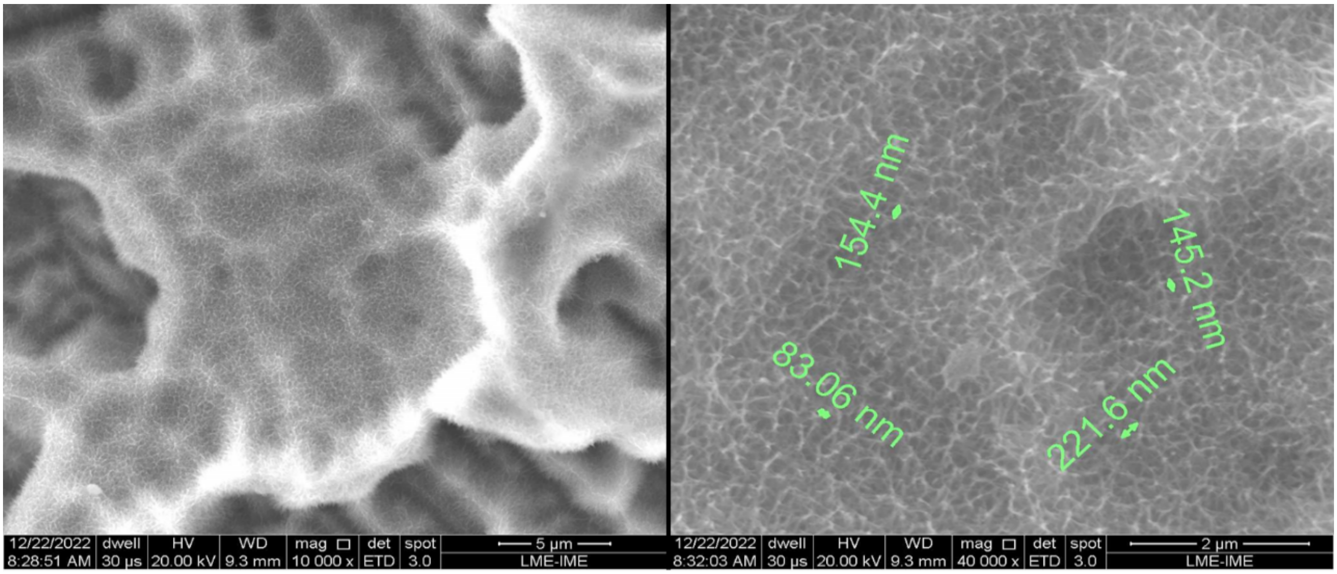
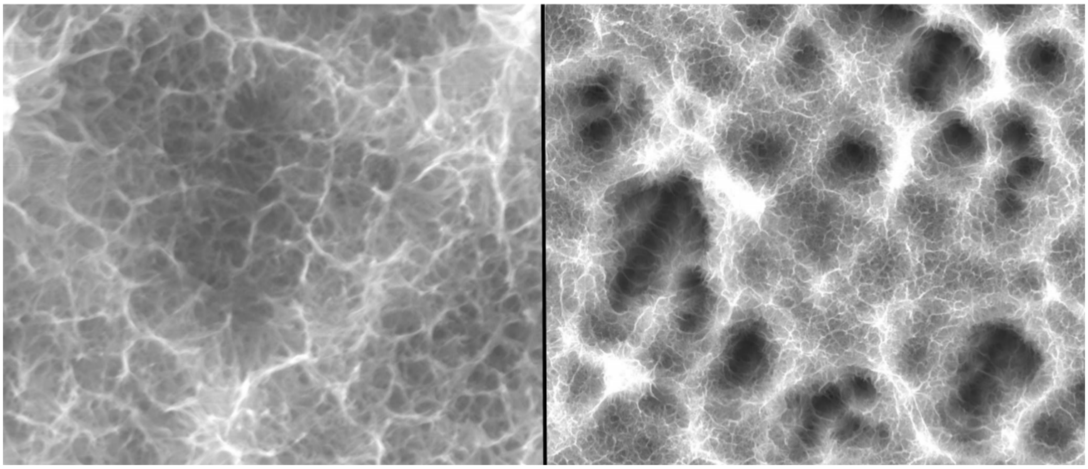


Fig. 6 - Micrograph (FEG-SEM) of the surface of the Ti-Sil subgroup titanium sample after silanization. FEG-SEM. 10000X, 40000X magnification.



The surface characterization in Zygo laser interferometer of the samples of the three groups studied (**Figures 7, 8, 9; Tables 1, 2, 3**) showed a surface with roughness parameters compatible with commercial osseointegrable implants, including an average Ra parameter between 1 and 2 μm.

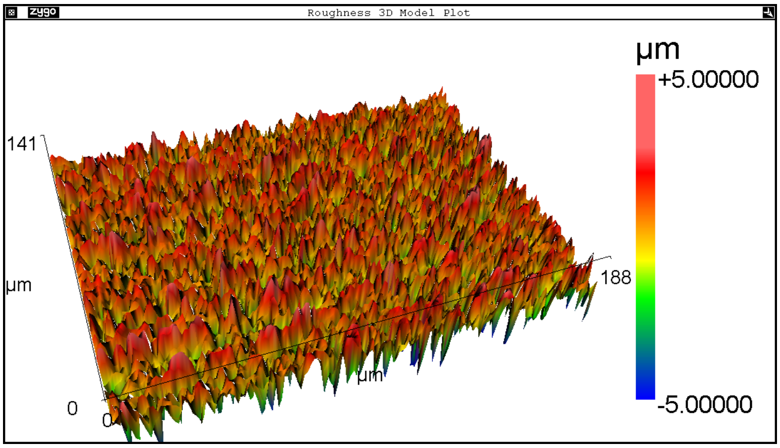
No significant difference was found between the means of the roughness parameters in the groups studied. However, statistical difference was observed between the other groups.

A comparative study in titanium samples with anodized surface, with acid etching and fluoride on the Ra value and contact angle, concluded that the contact angles were influenced by Ra in the sample groups. Despite the small difference between the Ra of the groups, the wettability had a tendency to reduce with the increase of Ra [18].

Different treatments applied to increase the surface roughness and wettability of titanium samples, and analyzed with profilometry and contact angle measurements, showed that wettability is not the critical parameter for cell adhesion and proliferation, and that surface topography plays the main role [19].

In this study, the highest mean Ra values were found in the Ti-Alk group, that is, after the alkaline treatment step, compared to the other two subgroups. Nevertheless, different from the results reported by Coutinho and Elias [18], the group that presented higher mean values of Ra also had lower values of contact angle and higher wettability.

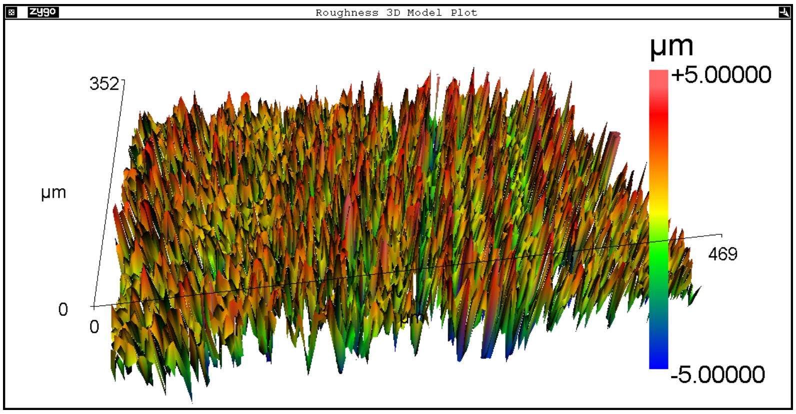
Fig. 7 - 3D image of sample surface microtopography after acid etching – Zygo Interferometer



Tab. 1 - Roughness parameters in five different regions on the sample surface after acid etching – Zygo Interferometer

	Ra	Rq	Peaks	Valleys	Peak density	Valley density	Peak distance
	μm	μm			1/mm^2	1/mm^2	μm
1	1.013	1.380	84	480	3,185.323	18,201.848	17.718
2	0.990	1.303	85	457	3,223.391	17,330.467	17.613
3	1.002	1.313	87	464	3,299.397	17,596.782	17.409
4	1.105	1.453	86	403	3,261.101	15,281.670	17.511
5	1.083	1.424	92	418	3,491.314	15,862.711	16.924
Mean	1.038	1.374	87	444	3,292.105	16,854.696	17.435
Deviation	0.052	0.066	3	32	119.188	1,229.869	0.308

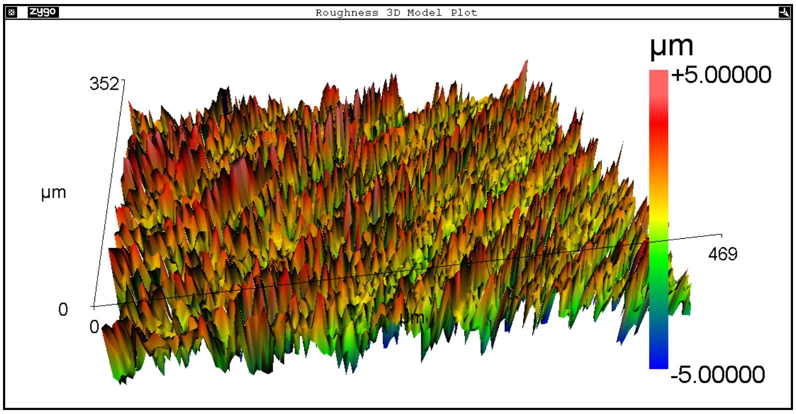
Fig. 8 - 3D image of sample surface microtopography after alkaline etching – Zygo Interferometer



Tab. 2 - Roughness parameters in five different regions on the sample surface after alkaline etching – Zygo Interferometer

	Ra	Rq	Peaks	Valleys	Peak density	Valley density	Peak distance
	μm	μm			1/mm^2	1/mm^2	μm
1	1.497	1.881	335	653	2,074.292	4,043.321	21.957
2	1.463	1.848	484	922	3,133.691	5,969.552	17.864
3	1.457	1.842	713	1374	4,980.133	9,597.060	14.170
4	1.522	1.895	396	569	2,429.805	3,491.310	20.287
5	1.550	1.927	295	626	1,797.977	3,815.368	23.583
Mean	1.498	1.879	445	829	2,883.180	5,383.322	19.572
Deviation	0.039	0.035	166	334	1274.508	2,546.225	3.687

Fig. 9 - 3D image of sample surface microtopography after silanization – Zygo Interferometer

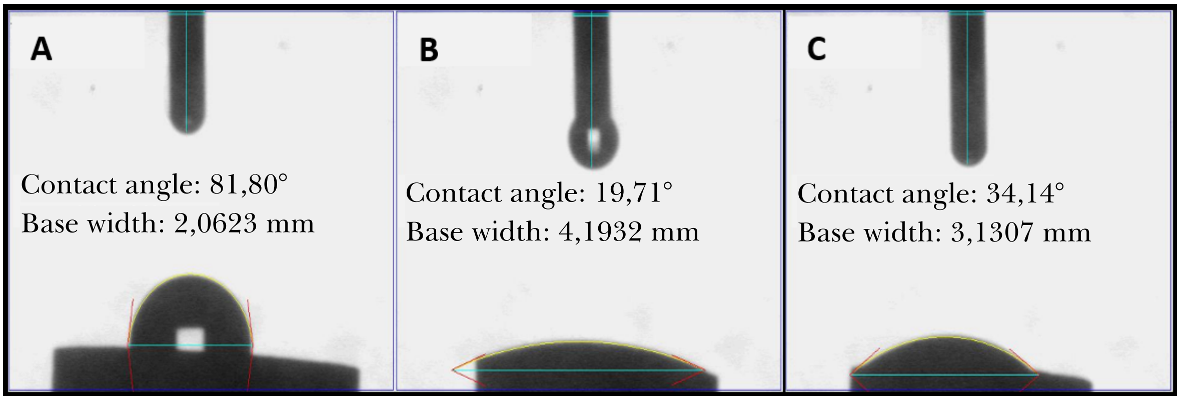


Tab. 3 - Roughness parameters in five different regions on the sample surface after silanization – Zygo Interferometer

	Ra	Rq	Peaks	Valleys	Peak density	Valley density	Peak distance
	µm	µm			1/mm²	1/mm²	µm
1	1.231	1.619	621	620	3,769.170	3,763.101	16.288
2	1.157	1.554	676	634	4,101.644	3,846.808	15.614
3	1.365	1.828	667	498	4,053.691	3,026.594	15.706
4	1.255	1.664	765	551	4,641.971	3,343.433	14.677
5	1.213	1.622	617	651	3,743.953	3,950.265	16.343
Mean	1.244	1.657	669	591	4,062.086	3,586.040	15.726
Deviation	0.076	0.103	60	64	362.253	388.538	0.673

The wettability quantified by measuring the contact angle with distilled water showed lower values on the surfaces of the Ti-Alk and Ti-Sil groups, causing the studied surfaces of these two steps to present more hydrophilic characteristics compared to the surfaces of the subgroup only with the acid treatment (**Figure 10**).

Fig. 10 - Wettability of the samples of the Ti-Etc (A), Ti-Alk (B), and Ti-Sil (C) groups.



5. Conclusion

According to the results found, it can be concluded that:

1. No significant difference was found between the means of roughness parameters in the surface samples of the three groups studied;
2. The samples of all groups studied presented mean roughness values, compatible with the parameters recommended in previous studies;
3. The samples of the Ti-Alk group showed significantly higher wettability than samples with only acid etching;
4. In the studied samples, the tendency of linear reduction of wettability with increasing roughness was not observed;
5. The characteristics found in the titanium surfaces after alkaline etching and silanization steps maintain adequate roughness parameters and also suggest an improvement in wettability, conferring greater hydrophilicity to these surfaces, compared to the initial group used as reference (Ti-Etc).

References

[1] JEMAT, A.; GHAZALI, M. J.; RAZALI, M.; OTSUKA, Y. Surface modifications and their effects on titanium dental implants. **BioMed Research International**, [*s. l.*], 2015. https://doi.org/10.1155/2015/791725

[2] LIU, Y.; RATH, B.; TINGART, M.; ESCHWEILER, J. Role of implants surface modification in osseointegration: A systematic review. **Journal of Biomedical Materials Research Part A**, [*s. l.*], v. 108, n. 3, p. 470-484. https://doi.org/10.1002/jbm.a.36829

[3] ZAVAGLIA, C. A. C.; PRADO-SILVA M. H. **Biomateriais**. São Paulo: Elsevier Inc., 2016.

[4] PESEK, J. J.; MATYSKA, M. T. Methods for the modification and characterization of oxide surfaces. **Interface Science**, [*s. l.*], v. 5, p. 103-117, 1997. https://doi.org/10.1023/A:1008648923962

[5] RYU, D. Y.; SHIN, K.; DROCKENMULLER, E.; HAWKER, C. J.; RUSSELL, T. P. A generalized approach to the modification of solid surfaces. **Science**, [*s. l.*], v. 308, n. 5719, p. 236-239, 2005. https://doi.org/10.1126/science.1106604

[6] LIU, X.; CHU, P. K.; DING, C. Surface nano-functionalization of biomaterials. **Materials Science and Engineering: R: Reports**, [*s. l.*], v. 70, n. 3-6, p. 275-302, 2010. https://doi.org/10.1016/j.mser.2010.06.013

[7] ZAFAR, M. S.; FAREED, M. A.; RIAZ, S.; LATIF, M.; HABIB, S. R.; KHURSHID, Z. Customized therapeutic surface coatings for dental implants. **Coatings**, [*s. l.*], v. 10, n. 6, p. 568, 2020. https://doi.org/10.3390/coatings10060568

[8] SAFONOV, V.; ZYKOVA, A.; SMOLIK, J.; ROGOWSKA, R.; LUKYANCHENKO, V.; KOLESNIKOV, D. Modification of implant material surface properties by means of oxide nano-structured coatings deposition. **Applied Surface Science**, [*s. l.*], v. 310, [*s. l.*], p. 174-179, 2014. https://doi.org/10.1016/j.apsusc.2014.04.110

[9] PUJARI, S. P.; SCHERES, L.; MARCELIS, A. T.; ZUILHOF, H. Covalent surface modification of oxide surfaces. **Angewandte Chemie International Edition**, [*s. l.*], v. 53, n. 25, p. 6322-6356, 2014. https://doi.org/10.1002/anie.201306709

[10] HOTCHKISS, K. M.; SOWERS, K. T.; OLIVARES-NAVARRETE, R. Novel in vitro comparative model of osteogenic and inflammatory cell response to dental implants. **Dental Materials**, [*s. l.*], v. 35, n. 1, p. 176-184, 2019. https://doi.org/10.1016/j.dental.2018.11.011

[11] LEITE, G. B.; FONSECA, Y. R.; GOMES, A. V.; ELIAS, C. N. Relação entre os parâmetros de rugosidade 3D e a molhabilidade do titânio com grãos micrométricos e sub-micrométricos. **Revista Matéria**, Rio de Janeiro, v. 25, n. 2, 2020. https://doi.org/10.1590/S1517-707620200002.1055

[12] WENNERBERG, A.; ALBREKTSSON, T. Effects of titanium surface topography on bone integration: a systematic review. **Clinical oral implants research**, [*s. l.*], v. 20, p. 172-184, 2009. https://doi.org/10.1111/j.1600-0501.2009.01775.x

[13] MATOS, G. R. M. Surface roughness of dental implant and osseointegration. **Journal of Maxillofacial and Oral Surgery**, [*s. l.*], v. 20, p. 1-4, 2021. https://doi.org/10.1007/s12663-020-01437-5

[14] LE GUÉHENNEC, L.; SOUEIDAN, A.; LAYROLLE, P.; AMOURIQ, Y. Surface treatments of titanium dental implants for rapid osseointegration. **Dental Materials**, [*s. l.*], v. 23, n. 7, p. 844-854, 2007. https://doi.org/10.1016/j.dental.2006.06.025

[15] WENNERBERG, A.; ALBREKTSSON, T.; ANDERSSON, B.; KROL, J. J. A histomorghometric study of screw-shaped and removal torque titanium implants with three different surface topographies. **Clinical Oral Implants Research**, [*s. l.*], v. 6, n. 1, p. 24-30, 1995. https://doi.org/10.1034/j.1600-0501.1995.060103.x

[16] LI, J.; ZHOU, P.; ATTARILAR, S.; SHI, H. Innovative surface modification procedures to achieve micro/nano-graded Ti-based biomedical alloys and implants. **Coatings**, [*s. l.*], v. 11, n. 6, p. 647, 2021. https://doi.org/10.3390/coatings11060647

[17] FREIRE, R. S.; PESSOA, C. A.; KUBOTA, L. T. Emprego de monocamadas auto-organizadas no desenvolvimento de sensores eletroquímicos. **Química Nova**, [*s. l.*], v. 26, n. 3, p. 381-389, 2003. https://doi.org/10.1590/S0100-40422003000300016

[18] COUTINHO, M. P.; ELIAS, C. N. Rugosidade e molhabilidade do titânio usado em implantes dentários com superfícies tratadas. **Revista Brasileira de Odontologia**, Rio de Janeiro, v. 66, n. 2, p. 234-238, 2010. http://dx.doi.org/10.18363/rbo.v66n2.p.234

[19] PORRELLI, D.; MARDIROSSIAN, M.; CRAPISI, N.; URBAN, M.; ULIAN, N. A.; BEVILACQUA, L.; MAGLIONE, M. Polyetheretherketone and titanium surface treatments to modify roughness and wettability–Improvement of bioactivity and antibacterial properties. **Journal of Materials Science & Technology**, [*s. l.*], v. 95, p. 213-224, 2021. https://doi.org/10.1016/j.jmst.2021.04.023

VOL.40 Nº4 2023
https://doi.org/10.22491/rmct.v40i4.12375.en

Direct energy deposition additive manufacturing (DED-AM) of surfaces with H13 tool steel powder atomized in Brazil

Rafael Ramos[1], Naiara V. Le Sénéchal[1], Helder P. Vicente[2], Paulo P. O. L. Dyer[2], Andersan S. Paula[1] e Getúlio Vasconcelos[2]
[1] Instituto Militar de Engenharia (IME), Graduate Program in Materials Science and Engineering (PPGCEM), Rio de Janeiro, RJ, Brazil
[2] Instituto de Estudos Avançados (IEAv), Division of Photonics, São José dos Campos, SP, Brazil
***ramos_rr@ime.eb.br**

Abstract: *H13 steel powder, for the first-time water atomized in a Brazilian company, was deposited by Direct Energy Deposition (DED) additive manufacturing (AM) over a microalloyed hot rolled steel sheet. This study aimed to evaluate microstructural, compositional, and mechanical aspects in the cross-section of the formed deposit layers, as well as at the substrate-first layer interface, using optical and scanning electron microscopy and microhardness analysis. The depositions resulted in a morphology with columnar grains close to the substrate (epitaxial growth) and a reduction in columnar extension, accompanied by changes in microhardness, in reheated areas at the intersections derived from layers stacking, which overlap involved partial layer remelting, and the in situ heat treatment of remelted zones adjacent volume. In addition, adequate geometric and chemical dilution depth at the first layer-substrate interface anchoring the deposit to the substrate was observed.*

*KEYWORDS***:** *Additive Manufacturing; Direct Energy Deposition; H13 Tool Steel; Microstructure; Dilution.*

Resumo: *Pó de aço H13, pela primeira vez atomizado em água em uma empresa brasileira, foi depositado por manufatura aditiva (MA) via deposição por energia direcionada (DED), sobre chapa laminada de aço microligado. Este estudo objetivou avaliar os aspectos microestruturais, composicionais e mecânicos ao longo da seção transversal das camadas de depósitos formadas, como também na interface substrato-primeira camada, por meio de microscopia óptica e eletrônica de varredura, assim como análise da microdureza. As deposições resultaram em uma morfologia com grãos colunares próximo ao substrato (crescimento epitaxial) e uma redução da extensão colunar, acompanhada de mudanças na microdureza, em áreas reaquecidas nas interseções decorrentes do empilhamento de camadas, cuja sobreposição envolveu a refusão parcial, e o tratamento térmico in situ do volume adjacente às zonas refundidas. Adicionalmente, verificou-se uma diluição química e geométrica com profundidade adequada na interface primeira camada-substrato, garantindo a ancoragem do depósito no substrato.*

PALAVRAS-CHAVE: Manufatura Aditiva; Deposição Por Energia Direcionada; Aço Ferramenta H13; Microestrutura; Diluição.

1. Introduction

H13 steel, widely defined in international standards AISI/SAE, whose classification and chemical composition (see Table 1) was adopted by ABNT NBR NM 122-1 [1], is very useful in the manufacture of molds and dies for casting and injection, due to the stability of its properties at high temperature [2, 3, 4]. However, traditional manufacturing methods are time-consuming, expensive [2], and rely on a sequence of several mechanical manufacturing steps and heat treatments [4].

Table 1 - Chemical composition of H13 steel.

	Elements (% by weight)							
	C	**Mn**	**Si**	**Cr**	**Mo**	**V**	**P**	**S**
Min	0.32	0.20	0.80	4.75	1.10	0.80	-	-
Max	0.45	0.50	1.20	5.50	1.75	1.20	0.03	0.03

Source: ABNT NBR NM 122-1 [1].

Additive manufacturing (AM), including directed energy deposition (DED), is gaining the attention of the tooling industry due to its potential to manufacture parts with customized geometries, complex cooling systems, and functionally graded materials (FGM) [2, 5]. Among the AM techniques, DED is capable of producing larger pieces [6] with the use of robotic arms that guarantee freedom of movement in various directions [7]. Nevertheless, understanding the peculiarities of the process and its relationship with the properties of the parts produced is a challenge [2, 7].

Parts produced by DED-AM normally have a cellular structure with equiaxial grains in reheated areas (intersection of layers) and columnar grain morphology in the lower portions (epitaxial growth), because of temperature gradients and heat flow during the process, with different properties associated with each of these regions [3].

The deposition of laterally or vertically overlapped tracks can lead to remelted or at least thermally affected zones (Heat Affected Zones – HAZ), in which heat treatments can occur [8]. In turn, fusion in deposition also generates a layer-layer or layer-substrate dilution depth ensuring the physical bond between them, as well as their chemical interaction. This situation is influenced by the process parameters, the convection in the melt pools, and the solubility between the elements (mainly at the layer-substrate interface) [9].

Thus, this study aimed to evaluate, from the cross-section of samples manufactured by DED-AM with fixed energy density, the interactions between the first layers of a deposit and the substrate.

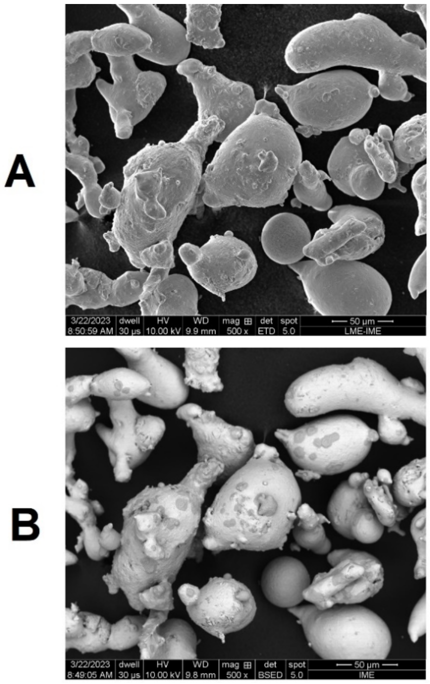
2. Material and Methods

The material manufactured by DED-AM is a H13 steel powder water atomized, produced in a Brazilian company for the first time, to the best of our knowledge [10]. The material was used in an attempt to stimulate national technological independence and reduce costs, considering the high prices involved in importing such powder manufactured by gas atomization in enterprises abroad.

Before being processed in the DED-AM equipment, the powder was sieved to be between 53 μm and 150 μm (mesh sieves #270 and #100) and had its particles analyzed by micrograph interpretation using ImageJ, generating results of $D_{50} = 68\ \mu$m and $D_{90} = 128\ \mu$m. Moreover, the morphology of the powder was irregular, with some spherical particles in the middle of several pear-like ones [11], as illustrated in Figure 1, aspects that apparently did not affect the feed flow or even the densification of the deposition in the DED-AM system.

The depositions were carried out on low carbon steel substrates, with relevant manganese content and molybdenum microaddition (Table 2 [12]), in a manufacturing laboratory of the Institute for Advanced Studies (IEAv), agency of the Aeronautics Technological Center (CTA) and partner in projects of the Military Engineering Institute (IME). A non-proprietary production apparatus was used, illustrated in Figure 2, resulting from the integration of several components whose initial parameterization was discussed in another study [10], which also aimed to stimulate national technological independence.

Figure 1 - SEM micrographs of H13 powder particles: (A) SE detector; (B) BSE detector.



Source: The author.

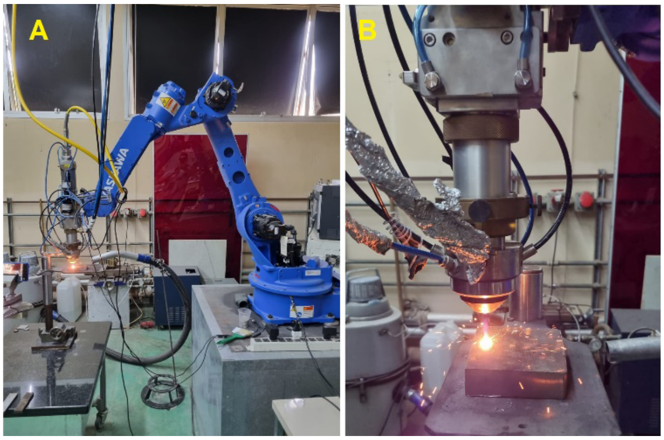
Table 2 - Chemical composition of the steel used as substrate.

Elements (% by weight)							
C	**Mn**	**P**	**S**	**Si**	**Cu**	**Ni**	**Cr**
0.082	1.606	0.016	0.008	0.005	0.014	0.009	0.020
Mo	**Sn**	**Al**	**N**	**Ti**	**Nb**	**V**	**B**
0.155	0.003	0.028	0.006	0.002	0.014	0.005	0.000

Source: Souza (2011) [12].

The metallic powder was conducted by inert argon gas (flow of 6 L/min, feed rate of 7.2 g/min) and melted by a laser beam (1282.5 W power and 5 mm/s scanning speed) during deposition, generating the superposition of several tracks and allowing the manufacture of samples with surfaces of one, two, and three layers stacked on the substrate, as illustrated in Figure 3. The mentioned parameters were defined after several tests with various combinations between scanning speeds, laser powers, and powder feed rates, among others [10].

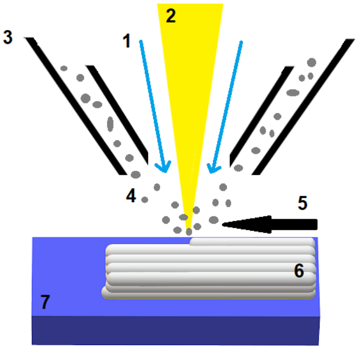
Figure 2 - Overview of deposition equipment: (A) robotic arm; and (B) laser head and powder feeders.



Source: The author.

The samples were cross-sectioned, as illustrated in Figure 4, prepared (sanded, polished, and chemically attacked with reagents Villela and Nital 5%), and evaluated by optical microscopy (OM) and scanning electron microscopy (SEM), the latter associated with the use of secondary (SE) and backscattered electron detectors (BSE), in addition to energy dispersive X-ray spectroscopy (EDS). Vickers microhardness tests were also performed (HV, 200 gf load, 3 to 5 indentations per horizontal line). The images from the SEM micrographies of the metal powder and the cross sections of the depositions were analyzed in the ImageJ software for particle size distribution and thickness, respectively.

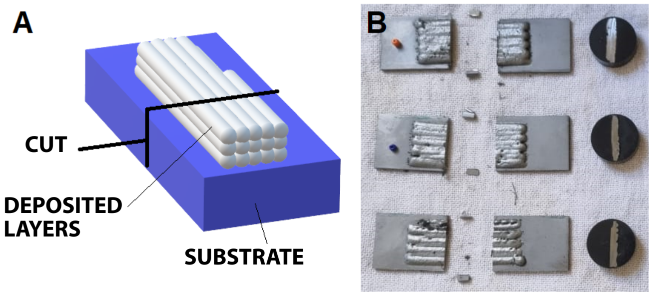
Figure 3 - Schematic diagram of the depositions performed.



Item	Description
1	Protection gas
2	Laser beam
3	Head structure
4	H13 metallic powder
5	Scanning direction
6	Deposited tracks
7	Substrate

Source: The author.

Figure 4 - View of the cut made in the deposited samples: (A) schematic drawing; and (B) samples with one, two, and three layers after cutting and embedding.



Source: The author.

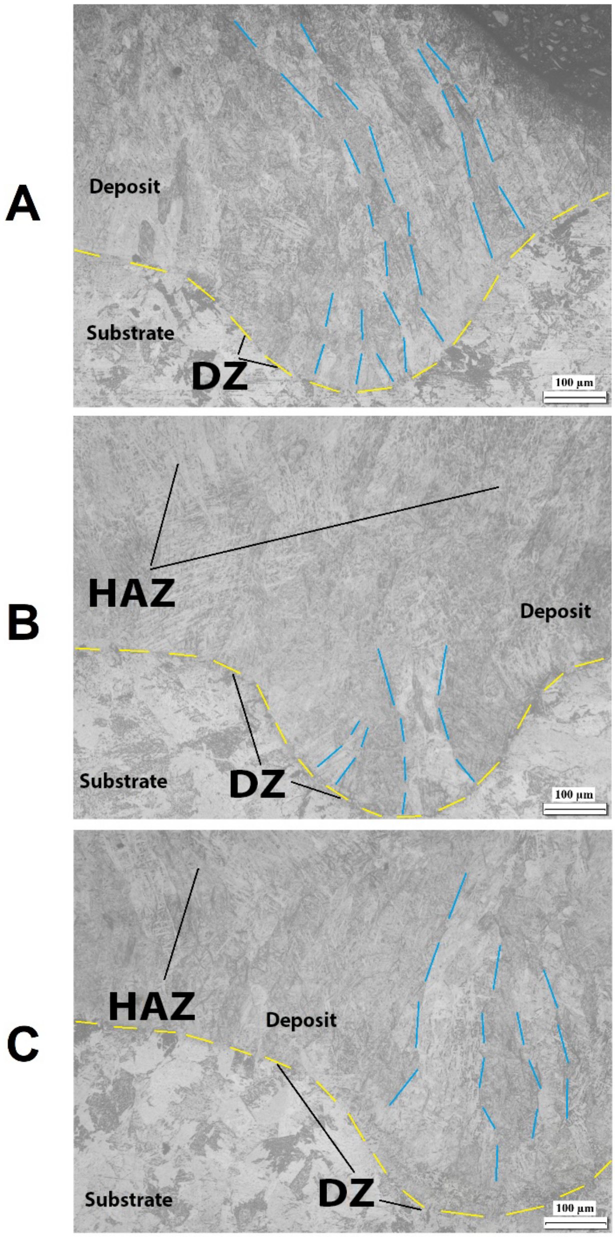
3. Results

Micrographs resulting from the OM analysis of the cross section of the depositions with one, two, and three layers were evaluated. Figure 4B hows three distinct depositions, and the areas in which the interaction between deposit and substrate or between layers was clear were chosen for micrographic analysis.

Figure 5 presents the dilution zone (DZ) between the deposited layers and the substrate, a region delimited by dashed lines with orientation close to horizontal, deflecting to follow the contour of the deposit bead observed, revealing a good level of anchoring and densification. Furthermore, the emergence of microstructural transition zones is visible in Figures 5B and 5C, compared to 5A. These heat affected zones (HAZ) occur due to the heat supply resulting from the deposition of laterally adjacent or overlapping tracks, in which the columnar aspect is reduced due to *in situ* heat treatments [5, 8].

In Figure 6, with a slightly higher magnification, the columnar aspect of the grains near the deposit-substrate interface, resulting from epitaxial growth, becomes even clearer than in Figure 5. In both cases, with indicative lines tending to be vertically oriented, remembering that there is no intention to demarcate all existing contours. Furthermore, when comparing the three depositions, one can notice the reduction of the columnar morphology of the grains, as well as their coalescence in samples with more layers of deposits and the predominance of the columnar aspect in higher positions (as in Figure 6C), in which new layers were deposited. Such changes are attributed to the greater heat flow in depositions with more layers, to the fact that the lower layers are already deposited and susceptible to the thermal action of the subsequent depositions and, finally, to the fact that such thermal action generates locally heterogeneous microstructures in the AM, depending on the thermal cycle developed in each position [13].

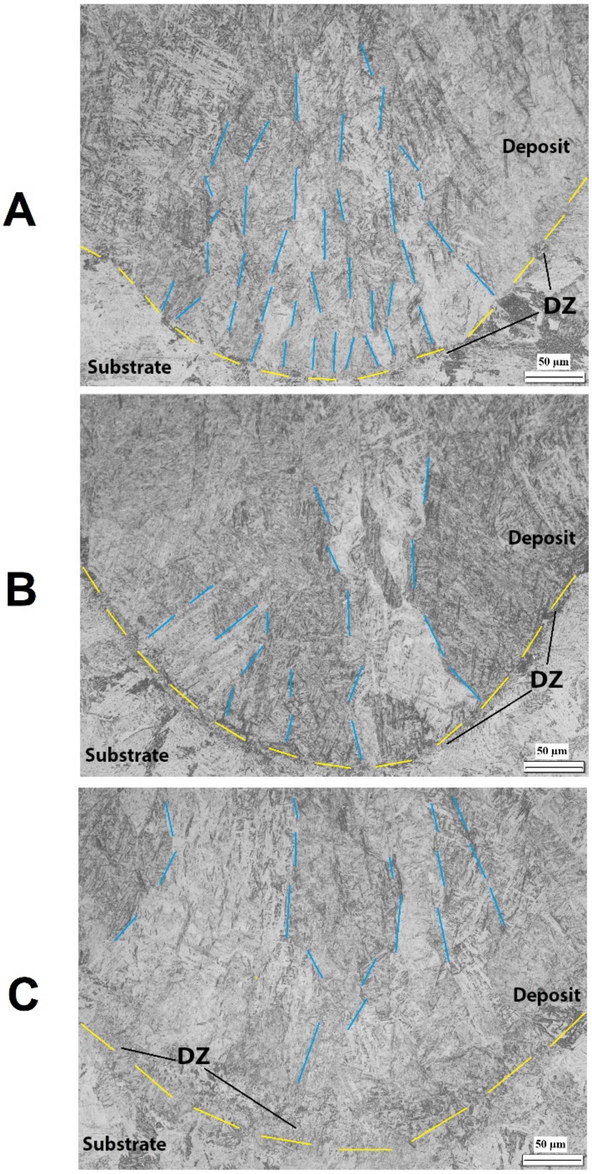
Figure 5 - Micrographs obtained via OM, bright field. Deposited layers: A, one; B, two; and C, three.



Source: The author.

In turn, the observation by dark field OM is shown in Figure 7, reinforcing the visualization of the columnar microstructure of the grains already mentioned (A) and the occurrence of thermal effects between adjacent layers; in (B), the upper portion of the bead of the first layer presents some morphological changes; in (C), the transition zone between the second and third layers appears in the sample in which three layers were deposited. In Figure 7C, the change in the direction of grain growth in the microstructure is also noticeable, as a result of changes in the heat flow due to the deposition of layers higher than previously deposited, since the grains grow in the opposite direction to the thermal gradient present in the solidifying melt pool [14].

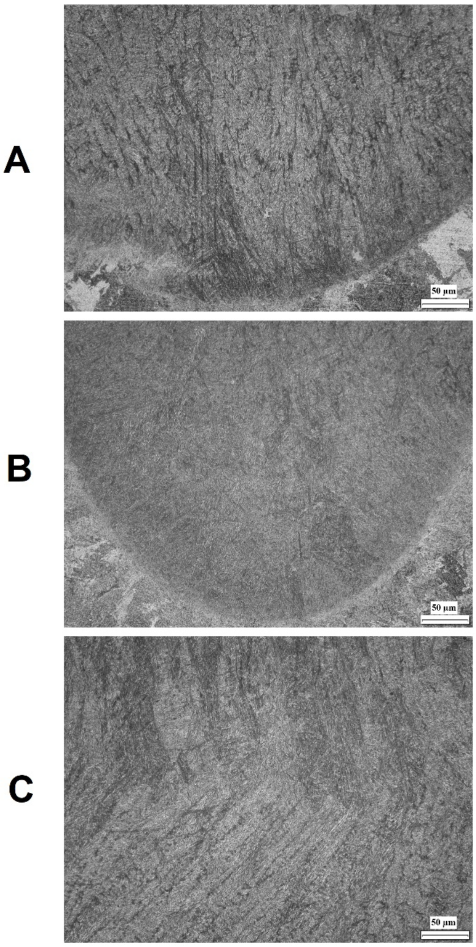
Figure 6 - Micrographs obtained via OM, bright field, high magnification. Deposited layers: A, one; B, two; and C, three.



Source: The author.

In the micrographs obtained via SEM with low magnification (Figure 8), it is possible to confirm the fact that the solidification is guided by the heat flow [14] and thus is modified by the number of layers and deposition strategy, as indicated by auxiliary lines marked with stars, drawn on the images following the direction of the solidification front in the microstructure. In addition, a good level of densification and anchoring in the substrate is confirmed, despite the presence of a few pores (Figures 8A and 8B). No evidence of the balling defect [6] was found, which had been observed in previous studies with the metal powder and the productive apparatus used, but in other configurations [10], indicating an improvement in the selection of the productive parameters used.

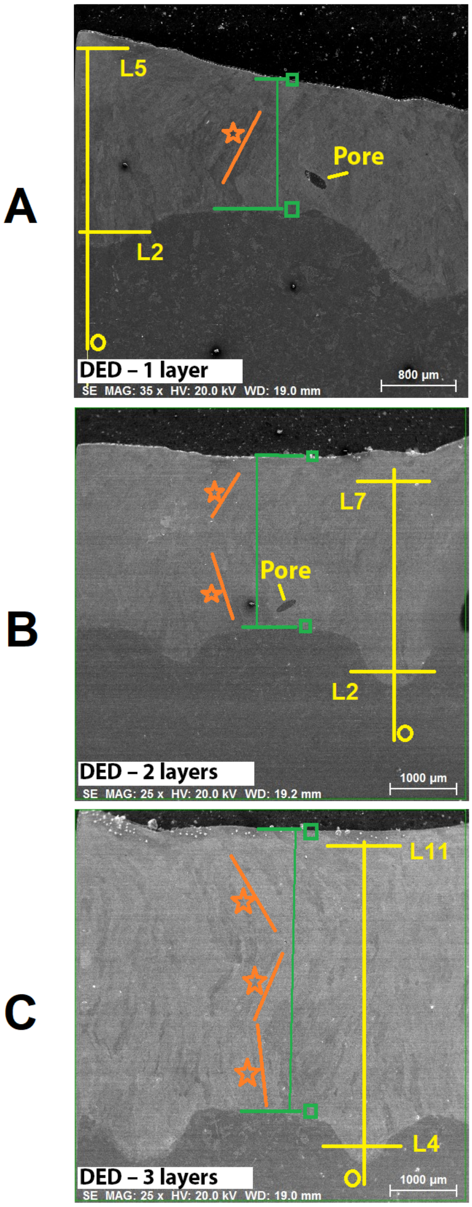
Figure 7 - Micrographs obtained via OM, dark field. Deposited layers: A, one; B, two; and C, three.



Source: The author.

Another aspect visible in Figure 8 is the existence of regions of greater penetration of the deposit on the substrate forming the already mentioned beads, regions related to the central portion of the laser beam during scanning, whose thermal action gradually decreases in positions that distance themselves from this central line.

Figure 8 - Micrographs obtained via SEM, low magnification. Deposited layers: A, one; B, two; and C, three.



Source: The author.

Figure 8 also shows, by marking a small circle at the base, vertical lines cut by horizontal lines with indications of numbers and letters (L2 and L5 in 8A; L2 and L7 in 8B; and L4 and L11 in 8C). They serve as a reference for a study of Vickers microhardness along the vertical profile of each sample under study, which will be presented below, highlighting that measurements were made on horizontal levels 0.5 mm apart from each other, starting at line 1, in position totally belonging to the substrate (L1, not shown). The lines represented in Figure 8 refer to the horizontal profiles in which the indentations began to occur in regions of the deposits (L2 in A and B and L4 in C), as well as to the final positions of greater height evaluated in the region of the deposits (L5 in A; L7 in B; and L11 in C).

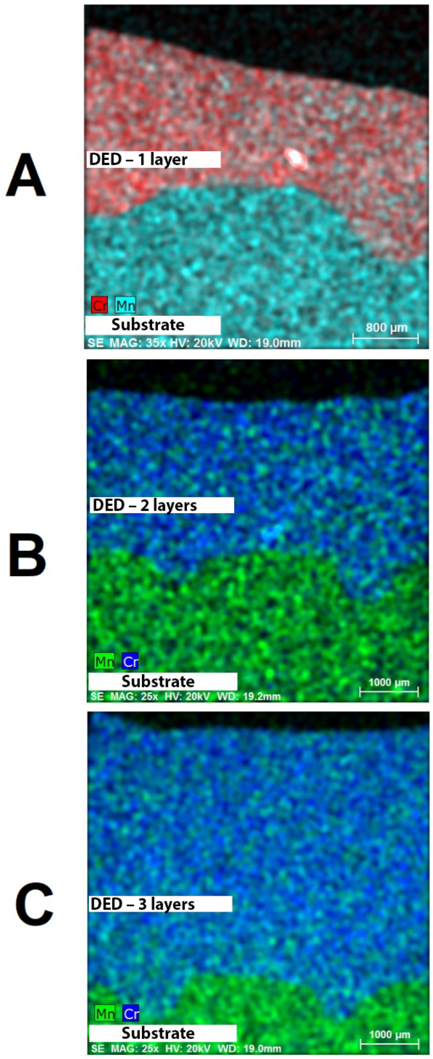
Moreover, the micrographs presented in Figure 8 were treated using the ImageJ software, allowing us to estimate the growth of the layers as having 1.51 mm in the sample with one layer, 2.55 mm in the sample with two, and 4.22 mm in the deposition with three layers (respectively in Figures 8A, 8B, and 8C). Such measurements were made from the region between two beads to the upper surface of the deposit, which are identified by horizontal lines, accompanied by squares, represented in the images of Figure 8. We highlight that we aimed to evaluate the estimated growth of the deposit and show that the increase in the number of layers deposited, with the parameters used, resulted in an effective growth of the deposit. The definition of a more representative value of the total deposit thickness would require a greater number of measurements, with images of different regions, given the possibility of variation in deposit thickness in different regions, aspects that were not the subject of this study.

An elemental chemical mapping in a microarea by EDS/SEM was conducted to distinguish the regions of the substrate and the deposit, as well as to identify the extent of the dilution zone, a region in which the mixture of alloying elements of the substrate and the deposit is expected. Given these objectives, although there are other alloying elements in these two portions under study, the alloying elements Cr and Mn were chosen for observation, because their compositions stand out, in percentage terms, in the deposit

(Cr, approximately 5%) and in the substrate (Mn, approximately 1.6%), with a small participation in the other portion (Cr 0.02% in the substrate; expected Mn of about 0.35% in the deposit), as can be seen in Tables 1 and 2, a contrast that allowed a better distinction between deposit and substrate.

The results of the chemical mapping by EDS are presented in Figure 9, indicating the existence of well--defined domains for the two portions of the material (substrate and deposit) and showing that the dilution region was not extensive, as well as confirming an adequate penetration of the deposit on the substrate.

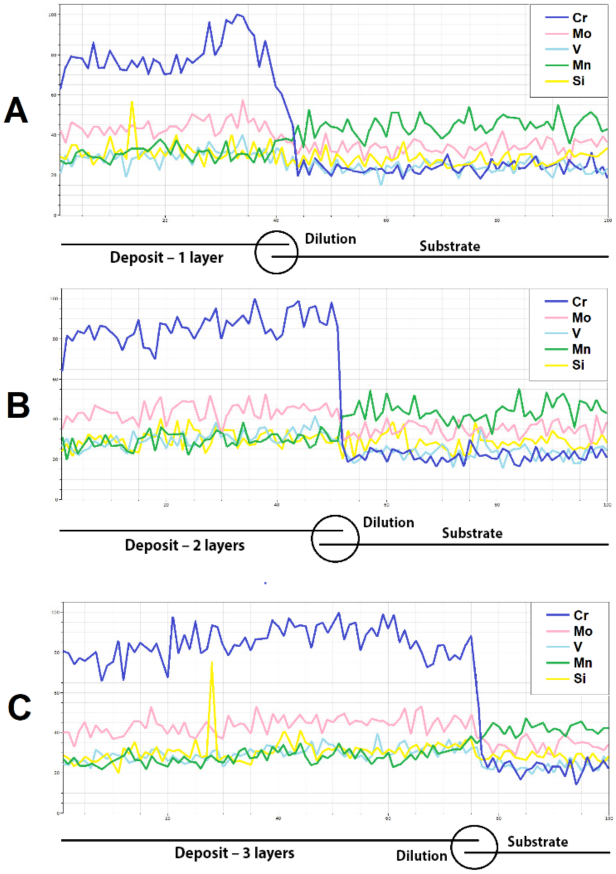
Figure 9 - Mapping of areas by EDS. Deposited layers: A, one; B, two; and C, three.



Source: The author.

An elemental chemical mapping by EDS/SEM was also performed (Figure 10), starting from the deposit towards the substrate, which confirmed the existence of two domains with relatively homogeneous chemical composition (deposit and substrate), separated by a transition portion in which the chemical composition changed smoothly. This revealed a good level of chemical dilution, in addition to confirming that such chemical mixing occurred to a limited extent [15], and this pattern was observed in all three conditions under study. Furthermore, the elemental mappings (Figure 10) revealed the evolution of the other alloying elements not selected in the microareas mappings of Figure 9, in addition to corroborating the already mentioned idea that Cr and Mn were able to indicate the regions of each of the domains (deposit and substrate) and the transition zone between them.

Figure 10 - Elemental mapping by EDS. Deposited layers: A, one; B, two; and C, three.



Source: The author.

Finally, the analysis with Vickers microhardness (HV) test, following the vertical profiles already mentioned, identified in Figure 8 by the vertical lines with circle marking at their base, can be seen in Figure 11, in which the mean values of HV and the standard deviations (SD) for each of the horizontal lines along the vertical profiles already mentioned are presented. An increase in HV is observed in the direction from the substrate towards the upper portions of the H13 deposit, despite a slight reduction in HV values in some measurement lines, and the SD is a tool to observe how the properties differ in adjacent regions due to the thermal influence of the sequential deposition of tracks [2, 3], above or laterally, as well as the regions influenced by dilution. In addition, large SD values marked the line where the deposits began, as they referred to the lines with indentations in the region of penetration of the deposit into the substrate (beads), already mentioned), as well as in the substrate portions adjacent to the beads (L2, L2, and L4, respectively, in Figures 11A, 11B, and 11C, also illustrated in Figure 8).

4. Conclusions

H13 deposits containing one, two, and three layers showed a good level of anchoring and interaction with the substrate, as well as spatially limited dilution regions with smooth transition of chemical composition at the substrate-deposit interface.

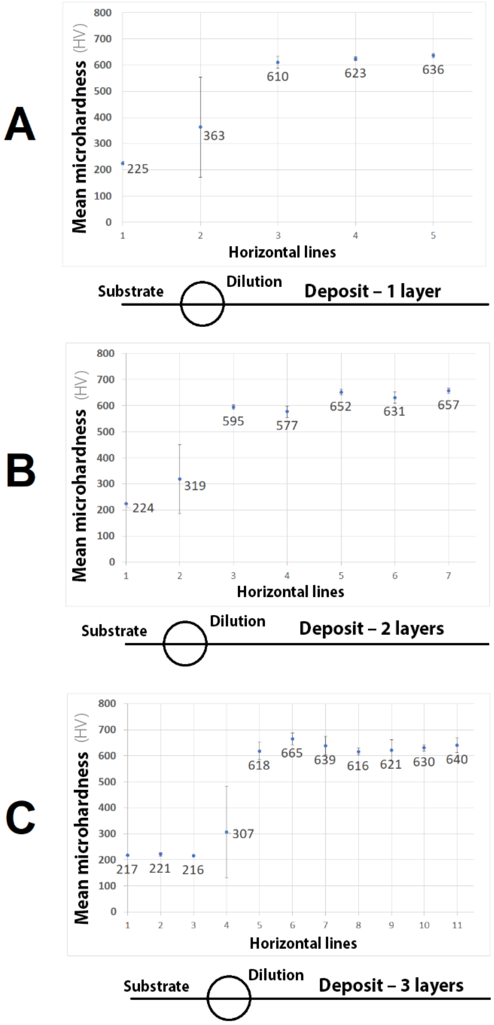
In addition, there was a directional development of the microstructure after solidification, related to the epitaxial growth of the columnar grains and the heat flow throughout the solidification of the layers. Moreover, the thermal influence of laterally adjacent or overlapped tracks was portrayed in the variation of microhardness values along the vertical profile and also in the same horizontal line, as well as in the changes in microstructure morphology, confirming a heterogeneous pattern of microstructure and properties in samples "as manufactured" by AM.

This study opens space for a continuation by similar analysis, but with samples submitted to heat treatment routes, seeking to homogenize microstructure and properties throughout the deposited material.

Acknowledgements

The authors thank: Villares Metals, for offering conventional H13 steel, which was atomized; IEAv and IME, for the production structure and characterization of materials; the Brazilian Center for Physical Research (CBPF), for support in cutting and preparation of substrates; the Brazilian Funding Agency for Studies and Projects (FINEP) and the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) (Pró-Defesa IV and PROEX), for the financial support.

Figure 11 - Mean results of Vickers microhardness (HV) and its standard deviations (SD). Deposited layers: A, one; B, two; and C, three.



Source: The author.

References

[1] ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR NM 122-1**: Aços ferramentas – parte 1: Classificação, designação e composição química. Rio de Janeiro: ABNT, 2005.

[2] ZHAO, X. *et al*. The effect of thermal cycling on direct laser-deposited gradient H13 tool steel: Microstructure evolution, nanoprecipitation behavior, and mechanical properties. **Materials Today Communications**, [*s. l.*], v. 25, 101390, 2020.

[3] JOSHI, S. *et al*. Solidification and microstructure evolution in additively manufactured H13 steel via directed energy deposition: Integrated experimental and computational approach. **Journal of Manufacturing Processes**, Amsterdam, v. 68, p. 852-866, 2021.

[4] KANG, M. *et al*. The effects of annealing temperature and cooling rate on carbide precipitation behavior in H13 hot-work tool steel. **Journal of Alloys and Compounds**, Amsterdam, v. 627, p. 359-366, 2015.

[5] BOHLEN, A. *et al*. Additive manufacturing of tool steel by laser metal deposition. **Procedia Cirp**, Amsterdam, v. 74, p. 192-195, 2018.

[6] DEBROY, T. *et al*. Additive manufacturing of metallic components–process, structure and properties. **Progress in Materials Science**, Amsterdam, v. 92, p. 112-224, 2018.

[7] LIMA, M. S. F. Processo de AM por adição de lâminas, por deposição com energia direcionada e híbridos. *In*: VOLPATO, N. (org.). **Manufatura Aditiva**: Tecnologias e Aplicações da Impressão 3D. São Paulo: Blücher, 2017. p. 271-287.

[8] AMIRABDOLLAHIAN, S. *et al*. Tempering behavior of a direct laser deposited hot work tool steel: Influence of quenching on secondary hardening and microstructure. **Materials Science and Engineering: A**, Amsterdam, v. 814, 141126, 2021.

[9] PELLIZZARI, M. *et al*. Optimizing direct laser metal deposition of H13 cladding on CuBe alloy substrate. **Surface & Coatings Technology**, Amsterdam, v. 432, 128084, 2022.

[10] RAMOS, R. *et al*. Parametrização de equipamento de deposição por energia direcionada utilizando pó de aço ferramenta H13 atomizado em água. *In*: CONGRESSO ANUAL DA ABM, 76., 2023, São Paulo. **Anais** […]. [*S. l.*]: ABM, 2023. p. 2473-2483. DOI: 10.5151/2594-5327-39934.

[11] BRAGA, V. *et al*. Microstructural and mechanical aspects of laser metal deposited H13 powder for die repair. **Materials Today Communications**, Amsterdam, v. 29, 102945, 2021.

[12] SOUZA, A. C. A. **Produtos de transformações de fases obtidos sobre distintas condições de resfriamento em aço baixo carbono com microadição de molibdênio laminado a quente**. 2011. Projeto de Conclusão de Curso (Graduação) – Universidade Federal Fluminense, Volta Redonda, 2011.

[13] GE, J. *et al*. Investigation on H13 buildups produced with wire arc additive manufacturing: Deposition strategies-induced microstructural evolution and mechanical performances. **Journal of Alloys and Compounds**, Amsterdam, v. 860, 157893, 2021.

[14] YANG, X. *et al*. Microstructure evolution and mechanical properties of H13 steel produced by Selective Electron Beam Melting. **Materials Characterization**, Amsterdam, v. 203, 113053, 2023.

[15] DASS, A.; MORIDI, A. State of the art in directed energy deposition: From additive manufacturing to materials design. **Coatings**, Basel, v. 9, n. 7, n. 418, 2019.

VOL.40 Nº4 2023
https://doi.org/10.22491/rmct.v40i4.12481.en

Identification and classification of roadway distress using an integrated system with convolutional neural network

Kaio Gefferson de Almeida Mesquita*[1], Yasmin Pereira de Brito Barroso[1], Thiago de Sousa Tostes[1], Luan Pablo de Holanda Barros[1], Thiago Prezotte Reis[1], Frederico Rodrigues[1]
[1]Imtraff Engenharia e Mobilidade
Av. Cristiano Machado, 640 - sala 1106 - Sagrada Família, Belo Horizonte (MG), Brasil
***e-mail: kaio@det.ufc.br**

ABSTRACT: Road transportation is essential to Brazil's economy, dominating goods and passenger movements. However, the country faces challenges in road quality, with only a fraction being paved and many exhibiting structural issues. Accurate recognition of these problems is crucial for effective road interventions. Manual approaches prove costly, slow, and error-prone, leading to the emergence of convolutional neural networks (CNNs) as a solution. This study introduces a multi-label CNN method for real-time distress detection on roads, aiming to reduce errors between distress classes in both aggregated and disaggregated categories. The results show clear advantages for the model with aggregated pathology categories, demonstrating improved accuracy, recall, and average precision. Comparative analysis highlights enhanced accuracy with category aggregation and exclusion of poorly represented categories, resulting in an average error of just 7%. In conclusion, the operational model significantly contributes to optimizing the definition of road interventions.

*KEYWORDS:***Computer vision. Highways. Distress Detection. Convolutional Network. Flexible Pavement.*

RESUMO: No Brasil, o transporte rodoviário desempenha um papel crucial na economia, responsável por movimentações de mercadorias e passageiros. No entanto, as estradas enfrentam desafios significativos, com grande parte não pavimentada e muitas apresentando problemas estruturais, como fissuras, trincas e buracos. A detecção precisa dessas patologias é essencial para a intervenção adequada nas rodovias. Este estudo propõe um método inovador utilizando redes neurais convolucionais (CNNs), um subconjunto do aprendizado profundo, para a detecção de patologias nas estradas. O foco é reduzir erros entre classes considerando categorias agregadas e desagregadas. O método mostrou vantagens claras, com melhorias na acurácia, recall e média de precisão ao usar categorias agregadas. A análise comparativa revelou que a agregação de categorias e a exclusão de categorias pouco representadas aprimoraram a precisão do modelo. Com um erro médio de 7%, o modelo já está operacional, contribuindo de maneira prática para otimizar processos de intervenções em estradas.

PALAVRAS-CHAVE: Visão Computacional. Rodovias. Detecção de Patologias. Rede Convolucional. Pavimentos Flexíveis.

1. Introduction

Transportation infrastructure far exceeds most other in-demand assets in most countries. Developed and well-maintained road networks are crucial to develop the economy of a nation, especially for those with a road-based transportation matrix, such as Brazil. Brazilian roads transport about 61% of its goods and 95% of its passengers. Although Brazilian highways span 1,720,700 km, only 12.4% of these roads are paved — of which 59% were considered inadequate in at least one criterion (geometry, signage, or pavement) — and 52.4% of the evaluated highways had some defect in their pavement [6].

Studies indicate that visible pavement distresses show signs of structural wear [3] — openings with varying directions and shapes — that can cause accidents. Detecting and properly diagnosing distresses are of paramount importance. Repairing them or monitoring their behavior in cases in which repair is impossible minimizes damage or completely solves the problem. This necessitates regular monitoring and inspections for maintenance and failure prevention. Traditional manual visual inspections by engineers or specialized technicians have limitations, such as subjective evaluations, high cost and execution time, and difficulties addressing the scale and frequency necessary to adequately monitor extensive road networks. The manual approach is falling out of favor due to its subjectivity, time, and high financial and human resource costs. Moreover, manual inspection is becoming impractical due to the required scale and frequency. Technological advances have increased the trend for automated surveys to evaluate pavement with equipment ranging from cameras to technologies such as sensors and computer vision [16].

Over the past eight years, convolutional neural networks (CNNs) — a type of neural network widely known as deep learning — has gained remarkable popularity in many industries, especially in computer vision [11]. Convolutional neural networks lie within the deep learning universe. They have been developed to classify images as they can constantly analyze data and recognize patterns. A CNN can automate processes, but it also offers a particularly suitable tool to detect pavement distresses. Its singular effectiveness in detecting edges in images by applying convolution operations outperforms other deep learning models. This specific capability makes identifying paving problems quick and simple and improves result accuracy and reliability, making CNNs essential in applications in which accurate detection is crucial, such as road infrastructure maintenance. In addition to automating processes, CNNs are particularly suitable for detecting pavement distresses, outperforming other deep learning models due to their unique efficiency in detecting edges in images stemming from convolution operations. However, the application of deep learning techniques in pavement engineering is constantly evolving, with advances in hardware and software. A review of the literature on the application of deep learning to pavement maintenance found that most studies focused on road distresses, showing that attention has predominantly revolved on classifying, detecting, and segmenting cracks [10][14], rather than on furthering the detection of other types of less frequent distresses and on reducing detection errors, the focus of this study. Thus, manually assessing distresses loses accuracy due to human errors, requires more time than field collection itself, and presents methodological difficulties. Therefore, this study hypothesizes that models to identify distresses may lose accuracy during training due to their large number of types, being unable to generalize novel cases well and causing overfitting [9].

In this context, this research aims to propose a method to train and apply a convolutional neural network with multiple labels (creating boxes during training) to identify distresses on highways in real time, reducing aggregate errors between classes and generalizing new cases. The process seeks to automate the registration and diagnosis of highway pavement, expanding the pavement management system by comparing two models: one trained with an aggregated specific distress set and a generic one with various types of disaggregated distresses, according to the specifications of the DNIT standard (2003). This study also compared these models with manual counting and classification, using sample data from several stretches of a Brazilian highway with three lanes per route and comparing these results with performance indicators to evaluate the adequacy of the models to distress types.

2. Literature review

A road infrastructure is essential for any society. Structural defects due to aging and environmental variations can negatively affect its durability, which highlights the importance of monitoring and management strategies. Such infrastructure management generally has the preservation and extension of the useful life of long-term structures as its fundamental objective [4]. To maintain the integrity of structures, it is important to detect the beginning of any defect. Fissures, holes, and cracks constitute the most common distresses in pavements and the surface of structures; such structural degradation may threaten safety and reduce pavement life [23].

The first and most common procedure to assess the health of a structure is visual inspection, which is expensive and laborious. Moreover, manual inspection requires highly trained experts and is subject to their judgment. These limitations have motivated research in industry and academia on automated distress detection approaches [5]. Image-based approaches improve cost effectiveness due to the widespread availability of cameras and smartphones [13]. Computer vision techniques have effectively automated the identification on distresses by images. Their application has grown in recent decades [23].

Image-based distress detection studies usually fall into (i) manual methods and (ii) automatic approaches based on the extraction of characteristics. Early studies focused on methods such as edge detectors [1], morphological operations, and thresholding [7]. Manually extracted characteristics are fed into classic machine learning algorithms, such as support vector machines, random forests, and neural networks. Several distress detection studies have been proposed based on deep learning. They are categorized by image classification, object recognition, and semantic segmentation [12].

Deep learning methods focus on learning hierarchies of features based on higher levels that combine lower-level features. Learning features at various abstraction levels can train complex functions that directly map inputs to outputs from the data [10][14].

Convolutional neural networks (CNNs) were inspired by the functioning of the visual cortex for image classification [16]. CNNs can receive an input image and assign filters that can be learned to the objects in the image, differentiating one from the other [2]. The pre-processing CNNs require is much lower than in other classification algorithms, making processing computationally manageable since the network reduces the received images without losing information.

A CNN typically requires many tagged images to obtain high prediction accuracy. However, it is often difficult to collect thousands of images and label them manually. Transfer learning can build models with fewer inputs. Instead of starting the learning process from scratch, transfer learning chooses a model that has been trained on a larger dataset (such as ImageNet) to address a similar issue. This model uses pre-trained weights and a learning rate that adapts as needed [22].

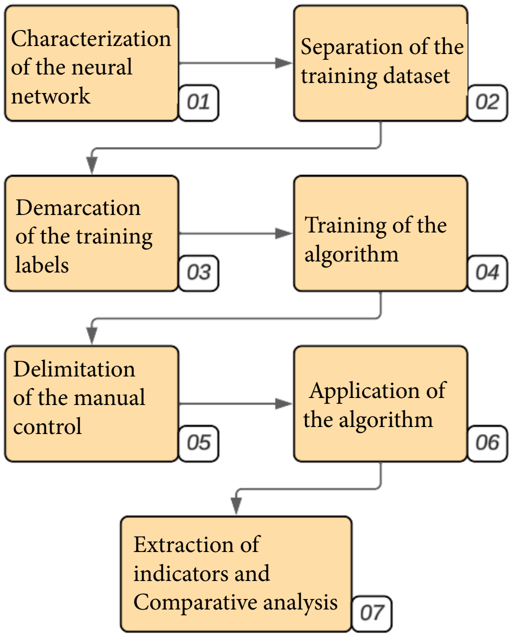
Learning rates consist of a tuning parameter in an optimization algorithm that determines step size in each iteration, moving it toward minimal loss function [17]. [20] stresses that it should vary within a range of values, rather than operating under a fixed or exponentially decreasing value. Its minimum and maximum limits are fixed, and the learning rate varies cyclically between these limits. That study proposed training for some epochs to estimate the minimum and maximum limit values.

3. Methods

A seven-stage development was proposed: (i) characterization of the neural network; (ii) separation of the training dataset; (iii) demarcation of the training labels; (iv) training of the algorithm; (v) delimitation of the manual control; (vi) application of the algorithm; and (vii) extraction of indicators and comparative analysis (Figure 1). The entire process was carried out in Python 3.10 and developed on Yolo v7 (You Only Look Once); a framework implemented with deep learning from a darknet. Based on this algorithm, implementations were carried out to optimize the speed and detection of several distresses in the same image, as will be described.

In total, three hypotheses were tested in this study: (i) training with disaggregated distress categories can negatively impact detection accuracy; (ii) distresses with more training samples tend to obtain more detections in new datasets, whereas those with fewer samples may decrease model accuracy; and (iii) the models to identify distresses may lose accuracy during training due to the large number of types, being unable to generalize new cases well.

Figure 1. - Methodology to detect distresses with a personalized neural network.

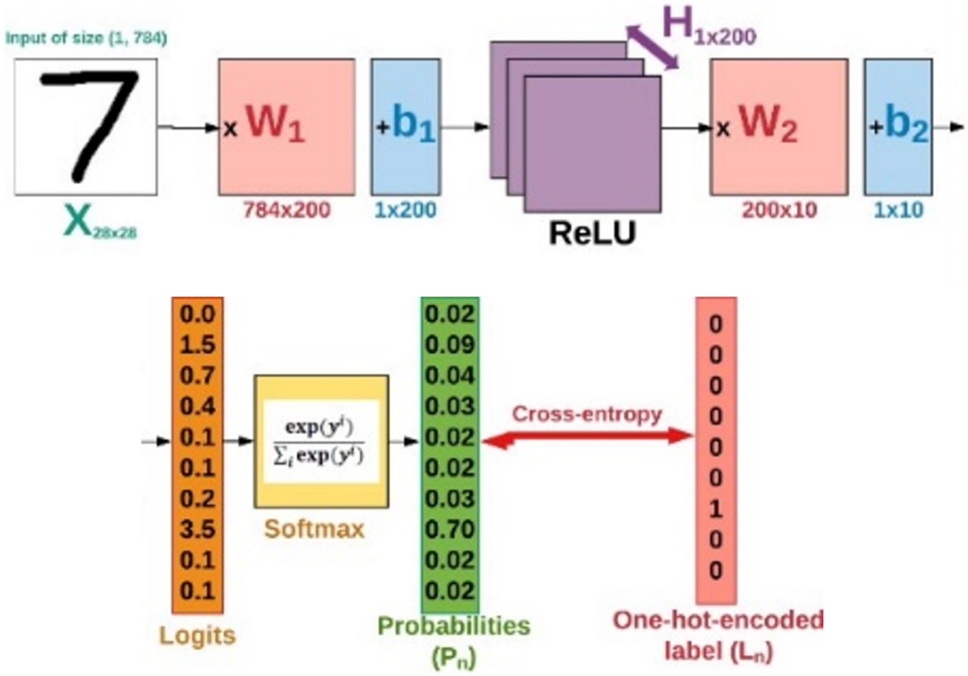


Source: Authors.

The first stage consists of characterizing and describing the convolutional neural network, as per the topic below in this section. Then, it is necessary to separate sample images for each road distress category under analysis. The highway analyzed in this study has three lanes per route, and information was collected from all lanes. Due to the Brazilian General Data Protection Law [15], the location of the highway in this case study will remain undisclosed as this section will only focus on the method to identify distresses. Then, the labels for each category were first manually then optimally demarcated according to the identification standard. The used tool was developed by the authors on Python and the pyqt5 framework to optimize this process.

The algorithm was trained on a 16-core 32-GB RAM GPU, enabling parallel processing and indicator visualization during training. Training calibrates the weights of each variable (or category) for each activation function in each layer of the neural network. The process is evaluated by a loss function comparing the original probability distribution with the trained distribution and decreasing the error with each sample picture epoch. The weights are updated by backpropagation, in which the derivative of each activation function is calculated in relation to each weight of the function to try to reduce the error in the next epoch since the derivative tends to zero when a local or global optimal point is found (Figure 2).

Figure 2 - Structure of the process of identifying objects in an image by CNN.



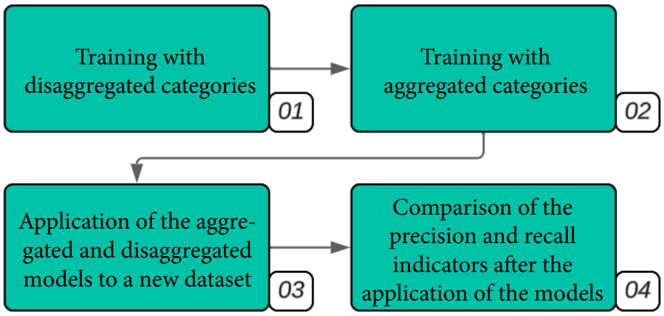
Source: [9].

A sample manual identification control was chosen for each applied model to compare human perception and artificial intelligence results according to error variation. Finally, after training the algorithms, the recall, accuracy, and mean precision indicators and the variation of these parameters between the aggregated (10 classes) and segregated distress (25 categories) models were compared. This variation stems from variations in some categories that are often difficult to identify (even manually). Thus, some categories were added, and the results were compared to test hypothesis I.

The darknet structure was chosen as it can process images in up to 0.05 seconds under a 30 frames-per-second rate. A contribution of this study involves its use of pre-trained initial weights to identify distresses, even before the new training process. The non-random initialization of the weights of each layer optimizes the training process and can ensure it avoids stagnating in an optimal place and reach the global optimal point of the required activation function.

Finally, Figure 3 objectively describes how this work intends to test the suggested hypotheses. First, the disaggregated and aggregated models (same database) were separately trained and applied to a new dataset (Step 3). The first hypothesis is tested in the training stage and the second one, in the application stage. In the fourth stage, in which the indicators are extracted, model overfitting is evaluated, thus testing the third and final hypothesis.

Figure 3 - Methodology for testing hypotheses.



Source: Authors.

3.1. Dataset and creation of labels to train distress identification models

The image database consists of photographic records of videos of a Brazilian three-lane highway. The recordings were carried out vertically with a GoPro Hero 10 Black 1,920-x-1,080-pixel resolution camera attached to the rear of a vehicle, enabling the visualization of the pavement due to its perpendicular relation to it. Thus, these images only show sections of the pavement and, at a fixed height and angle, can improve distress visualization and dimensioning. Figure 4 exemplifies the process of drawing the labels and the manual definition of the categories.

The groups demarcated for training in each model and the percentage of images per category are shown in Table 1. Only the acronyms of the distresses are shown in the tables to improve visualization. Their description is given below: W – Wear; FI – Fissures; A – Alligator Cracking without Erosion; AE – Alligator Cracking with Erosion; PH – Pot Holes; P – Patching; STC – Short Transverse Cracks; LTC – Long Transverse Cracks; SLC – Short Longitudinal Cracks; LLC – Long Longitudinal Cracks; BC – Block Cracking without Erosion; BCE – Block Cracking with Erosion; U – Upheaval; S – Shoving; Ra – Raveling; SC – Slippage Cracks; Ru – Rutting. An important observation for understanding defect types is that "fissures" configure capillary distresses that are perceptible to the naked eye at a distance equal to or below 1.50 m from the pavement. Cracks, on the other hand, show greater thicknesses than those of fissures and are thus, easily visible to the naked eye. Cracks fall into three classifications according to their thickness and erosion: FC-1, FC-2, and FC-3: those with openings greater than those of fissures (as mentioned above) but smaller than 1 mm; those with openings greater than 1 mm but without erosion at their edges; and cracks with openings greater than 1 mm with erosion at their edges, respectively. Moreover, the classification of FC-1 cracks applies only to isolated cracks, whether longitudinal, transverse, or shrinkage [8]. It is important to stress that the aggregate model compiled some categories — as explained in Table 1 — and removed those showing low occurrence to test the influence of these categories on class separation accuracy.

Figure 4 - Demarcation of the labels on the dataset.

Source: Authors.

Table 1 - Summary of training data on the aggregated (a) and disaggregated (b) models

Aggregate model	Images	Proportion
W	120	3.1%
FI	341	8.8%
A	496	12.8%
AE	61	1.6%
PH	304	7.8%
P	619	15.9%
STC	790	20.3%
LTC	751	19.3%
SLC	369	9.5%
LLC	35	0.9%

Disaggregated Model	Images	Proportion
W	120	3.1%
FI	338	8.7%
A	496	12.7%
AE	61	1.6%
PH	304	7.8%
P	619	15.8%
STC-FC1	416	10.6%
STC-FC2	348	8.9%
STC-FC3	26	0.7%
LTC-FC1	399	10.2%
LTC-FC2	312	8.0%
LTC-FC3	40	1.0%
TLC-FC1	120	3.1%
SLC-FC2	239	6.1%
SLC-FC3	10	0.3%
LLC-FC1	4	0.1%
LLC-FC2	24	0.6%
LLC-FC3	7	0.6%
BC	2	0.1%
BCE	-	-
U	1	0.0%
S	11	0.3%
Ra	5	0.1%
SC	1	0.0%
Ru	4	0.1%

continue

Source: Authors.

3.2. Neural network training

Convolutional neural network training includes mathematical optimization by a cost function (loss function) to automatically adjust for neural network biases and weights. In this procedure, the training dataset with expected labels is introduced to the optimization procedure to find a global minimum loss function.

Training the CNN consumed a lot of computational time, and the use of non-random initial weights is of paramount importance to speed up this process. An alternative to GPUs is field-programmable gate arrays or application-specific integrated circuits. As for the hyperparameters of the models, this study decided to retain them in both in order to evaluate the influence of the training data without external factors. Thus, a 4 lot and 350 epochs were defined.

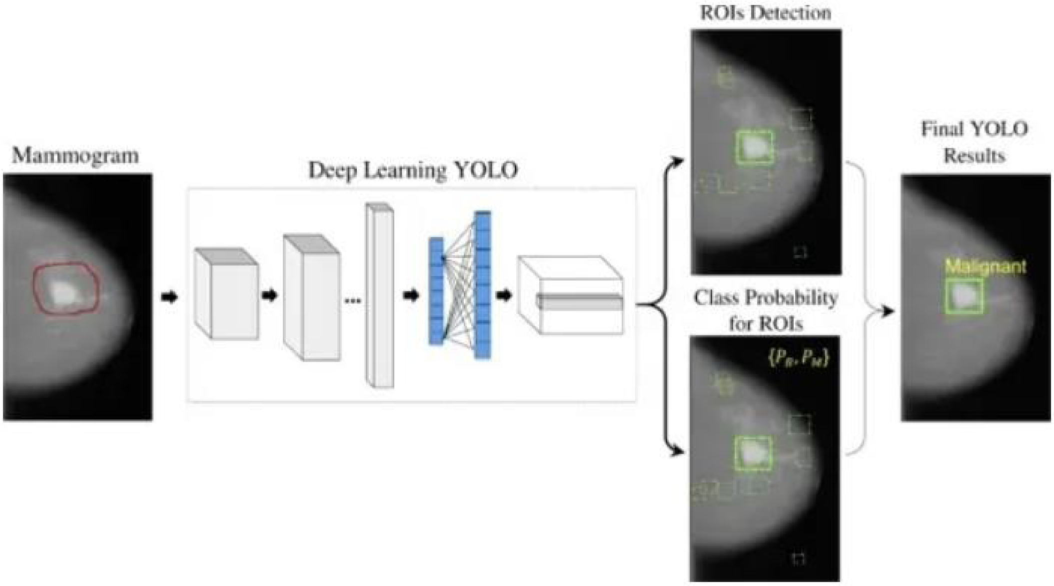
3.3. The algorithm in detail

First, the algorithm converts the images into an S-x-S bounded grid. A 19-x-19 grid was used in this study, as in [9]. Each cell is responsible for predicting five bounding boxes in case that cell has more than one object. The confidence score — which indicates how accurate the algorithm is that a given bounding box has an object — is also retrieved.

Cells also predict a class for each box, thus functioning as a classifier: a probability value is provided for each possible class. The confidence value for the bounding box and the prediction of the classes are combined into a final score, which informs the likelihood of that box containing a specific object. The 19-x-19 grid in this study resulted in 361 cells. In total, five bounding boxes are detected by each cell, totaling 1,805 boxes. Since most boxes will have an extremely low confidence value, only those with a score equal to or greater than 50% were considered. This threshold classification can be manipulated for model testing, but only a single threshold was considered in this study.

The use of Yolo stems from its effective prediction capacity in a single network epoch (Figure 5). Detection by dividing an image into several parts and executing a classifier in each such part would be necessary before this version, which would necessitate running the same classifier thousands of times over the same image [19].

Figure 5 - Example of single-epoch detection with Yolo.



Source: [19].

Building a pre-trained CNN could adequately identify pavement distresses [18]. Thus, the upper layers of the network often require huge changes in parameters, whereas the deeper layers — which have been well trained in detecting basic features (such as edges and contours) — need small modifications. The CNN is adjusted in two stages since transfer learning is used in the network architecture. The first step freezes the first layer group to avoid updating any of its parameters. The first layer group is thawed in the second step. Training resumes from the previous training state.

3.4. Indicators

The models were compared after the neural network was applied to evaluate the predictive capacity of the classification models based on test data and evaluate whether the models really "learned," i.e., if they could satisfactorily extend their predictions to unknown data sets. Precision, recall, mean average precision (MAP), and accuracy were used as indicators to evaluate the prediction level of the classification models. Recall measures the ability of a model to find all positive examples. It is calculated as the ratio of correct positive predictions to the total number of positive examples in the dataset. This metric aims to minimize false negatives. Accuracy measures the ability of a model to avoid labeling a negative example as a positive one, which is calculated as the ratio of correct positive predictions to the total number of positive predictions. Since different classes of objects can have different levels of difficulty in detection, MAP calculates the average precision for all classes in the dataset, offering a general view of the performance of a model across all classes. In addition to the indicators above, model accuracy in relation to manual identification for each category was also tested [21].

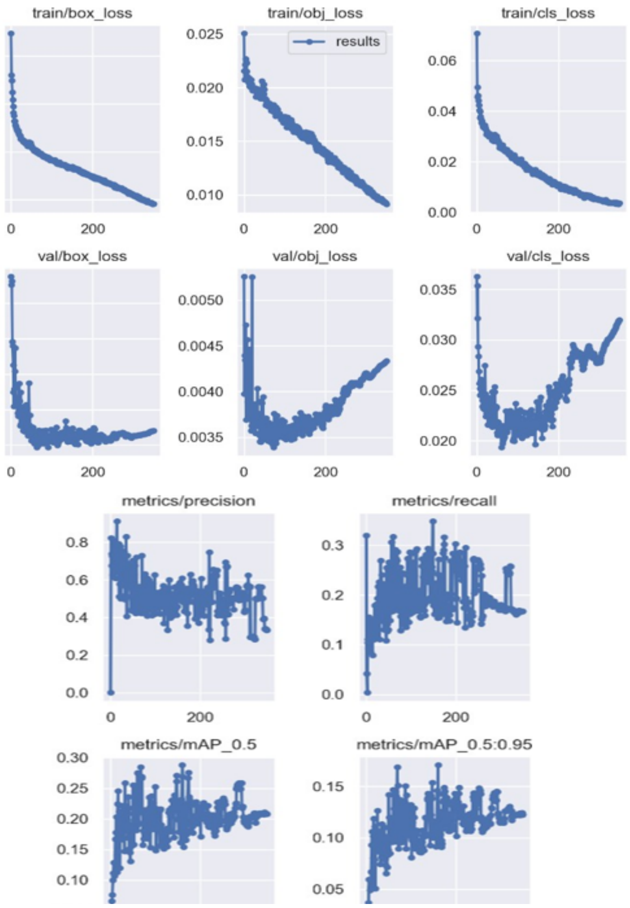
4. Results

Table 1 shows the data used to train each model and Table 2, the results of the model with the aggregated classes, that with the disaggregated classes, and the manual count. It is worth noting that the process limited itself to an 80.5-km highway stretch. Finally, this study compared the extracted results with each other, validating them with the control and analyzing the measurement error in distress detection.

4.1. Training results

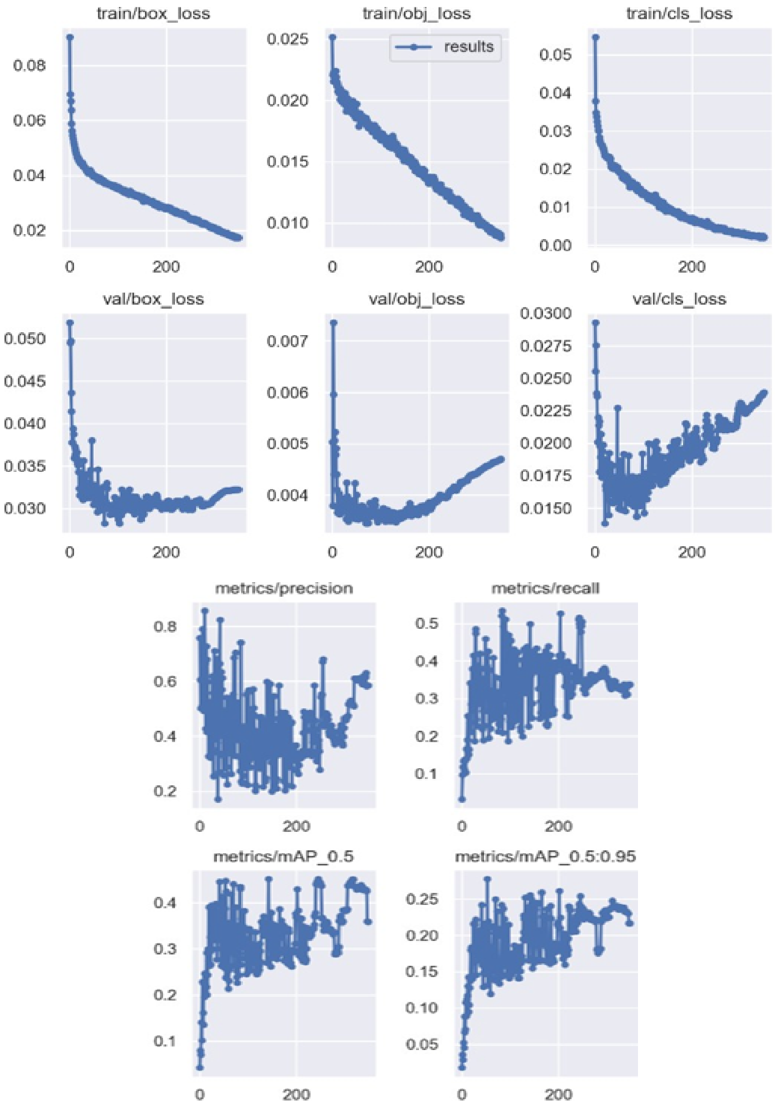
The following figures refer to the disaggregated and aggregated models, respectively. This study generated them during training and express the metrics to assess distress identification quality. Note that, reinforcing the hypothesis of improvement of the model with aggregated classes, the metrics proved to be generally superior in this model than in that with disaggregated classes. After 300 epochs, recall begins to stabilize at around 0.18 for the disaggregated model and at 0.34 for the aggregate one, indicating an 88% improvement. MAP consolidates itself around 0.12 for the former model and around 0.23 for the latter, indicating a 91% improvement, i.e., the capacity for positive events increases under a smaller possibility of the same class occurring due to its subclasses. Applying this model in practice requires the manual separation of the appropriate categories focusing on a certain class at a time, thus reducing the perception error due to wear. Regarding accuracy, the number of epochs failed to stabilize this metric in any case. Even so, the graph of the aggregate model indicates a trend of improvement in its final portion.

Figure 6 - Training result – disaggregated model.



Source: Authors.

Figure 7 - Training result – aggregated model.



Source: Authors.

4.2. Model Applications

Figure 8 exemplifies the practical results of distress identification as explained in the method section. Note that the identification for aggregate categories showed higher accuracy in all classes, except for alligator cracking with erosion. In these cases, the disaggregated model showed a 38% lower error rate than the control. The transverse cracks and longitudinal cracks categories obtained a 57 and 62% success rates, respectively.

Table 2 reflects the differences in metrics due to the practical application of the chosen models, comparing their results with the control group. Note the evidence validating the hypothesis that aggregating categories improves model accuracy since the mean error in the aggregated model only totaled 7%, totaling 407% in the disaggregated model, making its application unfeasible. Regarding hypothesis 2 (on the relation of sample data and model improvement), this study found no noticeable increase in indicators and identifications, requiring new analyses.

Figure 8 - Application of the (a) aggregated and (b) disaggregated models.

Source: Authors.

Table 2 - Comparative prediction results

Distresses	Model Aggregated	Manual
W	1	2
FI	2	2
A	230	404
AE	58	10
PH	26	4
P	116	81
STC	9	3
LTC	3	4
SLC	66	105
LLC	125	67

Distresses	Disaggregated Model	Manual
W	415	2
FI	13	2
A	343	404
AE	36	10
PH	43	4
P	1447	81
STC-FC1	4	1
STC-FC2	10	1
STC-FC3	1	1
LTC-FC1	13	2
LTC-FC2	0	1
LTC-FC3	4	1
SLC-FC1	218	94
SLC-FC2	48	10
SLC-FC3	1	1
LLC-FC1	100	55
LLC-FC2	65	11
LLC-FC3	1	1
BC	0	1
BCE	0	1

continue

Distresses	Model Aggregated	Manual
U	0	1
S	1	0
Ra	6	1
SC	0	0
Ru	0	0

Source: Authors.

Thus, the model with aggregated classes showed a significant improvement in relation to that with disaggregated ones according to the total number of distresses and excluding those that occurred only in the disaggregated model up to 682 manually counted instances. This study highlights the deviation between instances of model identification and manual counts: 384 aggregated and 2,213 disaggregated counts, respectively.

5. Final considerations

In summary, the aggregate model showed an average error of only 7%, a recall of 0.34, and a MAP of 0.23. On the other hand, the disaggregated model obtained an alarming average error of 407%, a recall of 0.18, and a MAP of 0.12, thus validating the hypothesis that aggregating classes and removing poorly represented ones could improve the accuracy of a model since, only with these methods, the results of the aggregate model trained with the same photos of the disaggregated model proved to be superior.

The influence of the network architecture and hyperparameters may influence training as much as the amount of data since variations on activation functions, loss functions, and learning rates may directly impact the identification of edges, silhouettes, and shapes. However, this hypothesis is yet to be evaluated.

Our future studies will use a more robust database encompassing the main Brazilian highways, with better distribution of distress classes and in portions with varying pavement since distresses can have completely different appearances in different scenarios. It is of great importance for academia and the market to offer the different models of identification by computer vision to Brazilian highways since those near the North and Northeast tend to show greater wear and tear.

References

[1] ABDEL-QADER, I.; ABUDAYYEH, O.; KELLY, M. E. Analysis of Edge-Detection Techniques for Crack Identification in Bridges. Journal of Computing in Civil Engineering, v. 17, p. 255-263, 2003.
[2] Academy, D. S. Deep learning book. [S. l.]: [S. n.], 2019.
[3] ALI, L.; ALNAJJAR, F.; JASSMI, H. A.; GOCHO, M.; KHAN, W.; SERHANI, M. A. Performance evaluation of deep cnn-based crack detection and localization techniques for concrete structures. Sensors, v. 21, n. 5, 2021.
[4] CAGLE, R. F. Infrastructure Asset Management: An Emerging Direction; AACE International Transactions: Morgantown, WV, USA, 2003.
[5] CHA, Y. J.; CHOI, W.; SUH, G.; MAHMOUDKHANI, S.; BUYUKOZTUK, O. Autonomous Structural Visual Inspection Using Region-Based Deep Learning for Detecting Multiple Damage Types. Computer-Aided Civil and Infrastructure Engineering, v. 33, p. 731-747, 2017.
[6] CNT - CONFEDERAÇÃO NACIONAL DO TRANSPORTE. Pesquisa CNT de Rodovias 2019. Brasília, DF: CNT, 2019.
[7] CORD, A.; CHAMBON, S. Automatic Road Defect Detection by Textural Pattern Recognition Based on AdaBoost. Computer-Aided Civil and Infrastructure Engineering, v. 27, p. 244-259, 2011.
[8] DNIT - DEPARTAMENTO NACIONAL DE INFRAESTRUTURA DE TRANSPORTES. Norma DNIT 005/2003 - TER. Defeitos nos pavimentos flexíveis e semirrígidos Terminologia. Rio de Janeiro: DNIT, 2003.
[9] GÉRON, A. Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, v. 1. p. 856, 2019.

[10] GOPALAKRISHANAN, K.; KHAITAN, S. K.; CHOUDHARY, A.; AGRAWAL A. Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection. Constr-Build Mater, v. 157, p. 322-330, 2017.

[11] GOPALAKRISHANAN, K. Deep learning in data-driven pavement image analysis and automated distress detection: A review. Data, v. 3, n. 3, 2018.

[12] HSIEH, Y. A.; TSAI, Y. J. Machine Learning for Crack Detection: Review and Model Performance Comparison. Journal of Computing in Civil Engineering, v. 34, 04020038, 2020.

[13] JAHANSHAHI, M. R.; KELLY, J. S.; MASRI, S. F.; SUKHATME, G. S. A survey and evaluation of promising approaches for automatic image-based defect detection of bridge structures. Structure and Infrastructure Engineering, v. 5, p. 455-486, 2009.

[14] KHANDELWAL, R. Computer vision: Instance segmentation with maskr-cnn. [S. l.]: [S. n.], 2019. Disponível em: https://towardsdatascience.com/computer-visioninstance-segmentation-with-mask-r-cnn 7983502fcad1. Acesso em: 14 jul. 2023.

[15] BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2020. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/lei/l14020.htm. Acesso em: 11 ago. 2023.

[16] MASSUCATTO, J. D. P. Aplicação de conceitos de redes neurais convolucionais na classificação de imagens de folhas. Dissertação (Mestrado) - Universidade Tecnologica Federal do Paraná, Curitiba, 2019.

[17] MURPHY, K. P. Machine Learning: A Probabilistic Perspective. [S. l.]: [S. n.], 2012.

[18] NIE, K.; WANG, Pavement Distress Detection Based on Transfer Learning, 2018 5th Int Conf Syst Informatics, ICSAI 2018, no. Icsai, p. 435-439, 2018.

[19] REDMON, J. You Only Look Once Unified Real Time Object Detection. Washington, D.C.: University of Washington, 2016.

[20] SMITH, L. N. Cyclical learning rates for training neural networks. Proc - 2017 IEEE Winter Conf Appl Comput Vision, WACV 2017, no. April, p. 464-472, 2017.

[21] SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. Information Processing & Management, v. 45, n. 4, p. 427-437, 2009.

[22] TAJBAKHSH N.; SHINA, J. Y.; GURUDU, S. R.; HURST, R. T.; KENDALL, C. B.; GOTWAY, M. B.; JIANMIN, L. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? Computer Vision and Pattern Recognition, v. 35, n. 5, p. 1299-1312, 2016.

[23] YANG, F.; ZHANG, L.; YU, S.; PROKHOROV, D.; MEI, X.; LING, H. Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. IEEE Transactions on Intelligent Transportation Systems, v. 21, p. 1525-1535, 2020.

VOL.40 Nº4 2023
https://doi.org/10.22491/rmct.v40i4.12510.en

A statistical methodology for reducing wear and tear on railway tracks and comparing it with real wear and tear

Larissa Fernandes Nunes[1], **larissa.nunes@mrs.com.br, Orcid 0000-0001-9337-9408**
Luiz Antônio Silveira Lopes[1], **laslopes@ime.eb.br laslopes@ime.eb.br, Orcid 0000-0002-2163-9204**
Luiz Paulo Brandao[1], **brandao@ime.eb.br, Orcid 0000-0002-2623-648X**
Wellington Luiz Pereira[2], **wellington.pereira@mrs.com.br, Orcid 0009-0000-6867-6822**
João Carlos De Serpa[2], **joao.serpa@mrs.com.br, Orcid 0009-0004-0110-698X**
[1] **Instituto Militar de Engenharia**
[2] **MRS Logística**

ABSTRACT: Railway tracks represent one of the most expensive assets for railways, and constant studies aim to extend their useful life. This work proposed equations for W1, W2, and W3 (variation of rail billet dimensions vertically, horizontally, and with an inclination of 45°) to understand wear and tear. Using real data, the equations were developed and applied to an eight-year-old rail, revealing significant differences between measured and expected values. This discrepancy can be attributed to variations between railway sections and the statistical model used. It is suggested that a more detailed approach consider individual characteristics of each section and additional variables. This highlights the need for personalized analysis of track wear, considering the diversity of conditions along the railway.

KEYWORDS:*Railway track. Useful life projection. Simplee linear regression. Service life. Wear and tear positions.*

RESUMO: Os trilhos ferroviários representam um dos ativos mais dispendiosos para as ferrovias, e constantes estudos visam prolongar sua vida útil. Este trabalho propôs equações para W1, W2 e W3 (variação de dimensões do boleto do trilho na vertical, horizontal e com inclinação de 45°) para entender o desgaste. Usando dados reais, as equações foram desenvolvidas e aplicadas a um trilho com oito anos de uso, revelando diferenças significativas entre os valores medidos e esperados. Essa discrepância pode ser atribuída às variações entre os trechos ferroviários e ao modelo estatístico utilizado. Sugere-se que uma abordagem mais detalhada considere características individuais de cada trecho e variáveis adicionais. Isso destaca a necessidade de uma análise personalizada do desgaste dos trilhos, levando em conta a diversidade das condições ao longo da ferrovia.

PALAVRAS-CHAVE: Trilho ferroviário. Projeção de vida útil;. Regressão linear simples. Vida útil em serviço. Posições de desgaste.

1. Introduction

Rails represent one of the most valuable components in the Brazilian railway sector, mainly attributed to the nature of their material, the complex manufacturing process and the lack of national production. Consequently, extending the useful life of this asset is a priority for railway companies, aiming to minimize and postpone additional costs. To achieve this goal, preventive maintenance practices are employed to reduce wear and tear and prevent the appearance of Rolling Contact Fatigue (RCF).

Rail wear and tear is the reduction of the cross-sectional area over its period of use, which creates a direct effect on the wheel-rail contact. This process results in the alteration of the mechanical properties of the material, which can adversely affect the dynamics of the railway vehicle, compromising reliability and safety of the permanent track [1-3].

Wear and tear of the railway is influenced by many factors, including:

1. Axle load: Studies have shown the relevance of axle load in the wear and tear rate of the rails [4; 5; 6];
2. Rail location: The position of the rail, such as in tunnels and curves, impacts its wear and tear [6];
3. Radius of curvature of the road: The radius of curvature directly influences the wear and tear of the rails [7];

4. Track geometry: The configuration of the railway track is a critical factor [8];
5. Speed of the compositions: The speed at which trains operate affects the wear and tear of the tracks [3];
6. Metallurgical rail properties: The characteristics of the rail material are decisive [6];
7. Dynamics of wheel-rail contact: The interaction between the wheel and the rail is crucial to understanding wear and tear [9];
8. Coefficient of friction: This factor is essential to understand the forces at stake [9];
9. Wheel and rail profile: The interaction of the rail and wheel profile directly influences wear and tear due to the contact ellipse [10].

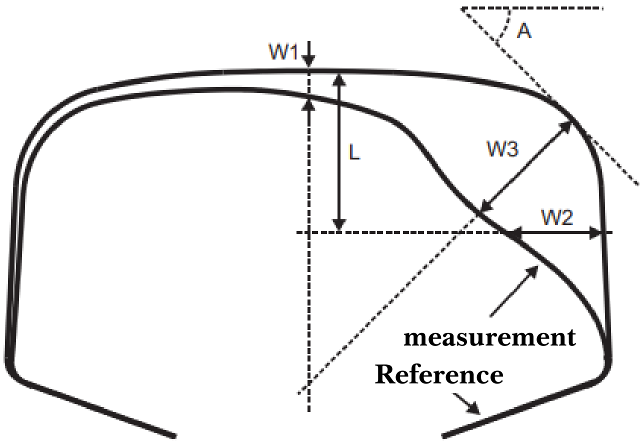
These studies highlight the complexity and multidimensionality of the phenomenon of wear and tear on railway tracks, underlining the importance of a holistic approach to the maintenance and layout of railways to minimize such effects.

Railways wear and tear is usually manifested in two main ways: vertical and horizontal/lateral. The vertical one is a widely observed phenomenon, which results from friction, abrasion, and the load imposed by the wheels of wagons and locomotives. This type of wear and tear leads to a progressive reduction in the height of the rail, potentially compromising the stability of the railway [11].

On the other hand, horizontal, or lateral, wear and tear affect the side of the rails, particularly at the top and base. This wear and tear is mainly induced by horizontal forces generated, e.g., when trains run on curves [11].

For the detailed analysis of wear and tear on rails, a study proposed a measurement method that identifies three critical areas, called **W1**, **W2,** and **W3**. **W1** corresponds to the change in the vertical profile at the top of the rail. **W2** refers to horizontal variation, whereas **W3** indicates the change measured at an angle of 45° [12]. Figure 1 shows the illustration of a rail, which clarifies the exact location of each measurement parameter, providing a clear view of the critical points for evaluation of wear and tear.

Figure 1 - Wear and tear analysis parameters.



Source: Adapted [13].

RCF defects arise from cyclic interactions between wheels of the train and rails during transport. These interactions result in shear forces that create surface defects in both the rails and wheels. The surface layer affected by these stresses undergoes a plastic deformation, resulting in an increase in hardness and changes in the microstructure of the material, due to hardening. This process can lead to the appearance of superficial cracks that can evolve and cause component failure, if not properly treated.

RCF is characterized by specific patterns of failures that manifest as nonconformities on the surface of the rails. Typical failures associated with contact fatigue include:

- Head checks: These are small cracks that appear at the corner of the rail gauge, arranged approximately 45° in relation to the rolling direction, resulting from cyclic stress on the surface [14];
- Shelling: As described by Rice [*apud* 15], this phenomenon occurs when head check type cracks deepen and expand inside the rail, reaching significant dimensions. This can lead to scaling of the rail material, which is a process known as shelling;
- Squats: They are the detachment of material from the surface of the rail, forming cavities. This type of defect is often associated with localized stress and accelerated wear and tear [14].

These failures highlight the importance of preventive maintenance practices and continuous monitoring of the railway infrastructure, aiming to mitigate the harmful effects of the RCF and ensure the safety and longevity of the railway system.

The adoption of strategies such as lubrication at the point of contact between wheel and rail and grinding represent effective methods in preventing and attenuating the effects of wear and tear and RCF [16].

The application of lubricants plays a crucial role in reducing and optimizing friction at wheel-rail contact, contributing significantly to decreasing wear and tear on these critical components. There are mainly two types of lubricants used for this purpose, namely:

- Friction modifiers: They are applied to the top of the rail to adjust the level of friction between the wheel and the rail, thus minimizing vertical wear and tear. The use of friction modifiers helps to balance the interaction between the wheels and the rails, extending their useful life [17];
- Grease: It is applied to the sides of rails, particularly in tight radius curves. The function is to mitigate horizontal or lateral wear and tear by protecting the rails against the abrasive effects of lateral forces exerted during the passage of trains [18].

Studies and practices in the field demonstrate the effectiveness of these approaches in maintaining the integrity of rail systems [19; 20]. The appropriate choice and careful application of lubricants are key to ensuring the efficiency of these preventive measures, standing out as vital components in the management of railway maintenance.

Rail grinding is a crucial preventive maintenance technique aimed at removing surface layers that have been deformed due to constant train traffic. This procedure is essential to restore the optimal geometry of the rail and improve the interaction between the wheel and the rail, eliminating defects such as RCF that can compromise the stability of the trains during circulation [21; 22].

Besides grinding and lubrication, designing and estimating the proper time for in-service rail replacement emerges as a significant proactive strategy. This approach aims to predict the remaining life of the tracks, enabling planning in advance for the replacements needed to maintain the safety and operational efficiency of the track. Such an estimate is critical for rail asset management, ensuring that operations are conducted safely and cost-effectively, minimizing disruptions and maximizing the longevity of the infrastructure.

This study aimed to compare the durability of the TR-68 rail in terms of wear and tear, using both projections based on statistical methods—specifically Pearson's linear correlation and simple linear regression—based on hard data, and the analysis of the cross-section of a rail withdrawn from operation after eight years due to a fracture. The results revealed that it is not feasible to define a single equation for each wear and tear position, capable of determining the useful life of the rail along the entire length of a railroad. This is due to significant variables, such as the characteristics of the terrain, the volume and type of cargo transported—and the speed of the railway compositions that are specific to each stretch and significantly impact the forecast of the asset's durability. This study also suggests that the complex reality of rails requires the consideration of a broader set of variables, indicating that simple linear regression may not be the most accurate methodology for this purpose.

2. Statistical methodology for predicting the useful life of rails

Initially, this study focused on the analysis of the correlation level between rail wear and tear and the volume of cargo transported, expressed in Millions of Gross Tons Transported (MGTT). To this end, Equation 1 determined the Pearson's correlation coefficient (**r**) used, aiming to quantify the degree of association between these two variables.

$$r = \frac{n \,.\, \sum xiyi - (\sum xi).\,(\sum yi)}{\sqrt{[n(\sum xi^2) - (\sum xi)^2]\,.\,[n(\sum yi^2) - (\sum yi)^2]}} \qquad (1)$$

In this context:

n represents the total number of data pairs analyzed;

xi and **yi** correspond, respectively, to the variables under study.

The aim of examining Pearson's correlation coefficient (**r**) is to determine the level of linear relationship between the two variables in question, **xi** and **yi** [23].

Table 1 shows in detail several inferences that can be drawn by interpreting the value of **r**, which was calculated by Equation 1.

Table 1 - Interpretation of the value obtained for Pearson's coefficient (r).

Coefficient of Pearson's correlation (r)	Linear correlation
r = 1	Perfect positive
0.8 ≤ r < 1	Strong positive
0.5 ≤ r < 0.8	Moderate positive
0.1 ≤ r < 0.5	Moderate positive
0 ≤ r < 0.1	Negligible positive
0	Zero
-0.1 ≤ r < 0	Negligible negative
-0.5 ≤ r < -0,1	Weak negative
-0.8 ≤ r < -0.5	Moderate negative
-1 ≤ r < -0.8	Strong negative
r = -1	Perfect negative

Source: adapted [14].

When Pearson's correlation coefficient (**r**) is equal to 0, there is an absence of linear correlation between the variables analyzed. On the other hand, **r** values close to 1 or -1 suggest a strong linear correlation between the variables, whether positive or negative, respectively. A positive correlation implies that as one variable increases, the other also increases, whereas a negative correlation indicates that the increase in one variable results in the decrease in the other. An **r** value equal to 1 or -1 denotes a perfect linear correlation, whether fully direct or inverse.

In addition to numerical analysis, the relationship between variables can be visualized graphically by means of a scatter plot. An arrangement of the variables that closely aligns with a straight line on the graph indicates a strong linear correlation (with Pearson's coefficient close to 1). In contrast, a random distribution of the variables in the graph signals a weak linear correlation (Pearson's coefficient close to 0) [23].

Based on Pearson's coefficient and the observation of the scatter plot, the simple linear regression model is employed to elucidate the relationship between the two variables in focus. This model was also used to make projections about the useful life of the rails, seeking to provide a simple tool for future forecasts based on historical data.

Simple linear regression is a statistical method that aims to establish a mathematical equation to describe the relationship between two specific variables, which are classified as [25; 26]:

- Dependent variable (**Y**), also known as the response variable, of which the behavior is to be predicted or better understood;
- Independent variable (**X**), also called the explanatory variable, which is used to make predictions or explain the variations in the dependent variable.

Equation 2 outlines the theoretical model to construct this simple linear equation [27]. This model provides the basis for calculating the linear relationship between variables, enabling educated predictions and analysis on how one variable may affect the other.

$$Y = \alpha X + \beta + \varepsilon i \qquad (2)$$

In which:

X, **Y**: variables under analysis;

α and β: unknown parameters;

εi: random error.

The values of α and β can be obtained by Equations 3 and 4 [27].

$$\beta = \frac{n.\sum xi\ .\ yi - (\sum xi).(\sum yi)}{n.(\sum xi^2) - (\sum yi^2)} \qquad (3)$$

$$\alpha = \overline{y} - \beta . \overline{x} \qquad (4)$$

In which, $\overline{x}$ is the mean of the values of **xi** and $\overline{y}$ is the mean of the values of **yi**.

3. Data collection

The data used in this statistical analysis, which covered linear correlation and simple linear regression, were collected between 2019 and 2022. The research focused on the Vignole TR-68 profile rail, using specific wear and tear measurements (**W1**, **W2** and **W3**). Figure 2 shows that the MiniProf equipment was used to measure wear and tear.

Figure 2 - Field measurement with the mini prof.

Source: Author.

The data compiled for this research reflect specific characteristics of the railway system studied, including:

- Estimated annual MTBT: ~ 150;
- Axle for transported load: 32 tons;
- Flow: Heavy Haul loaded;
- Installation position: External rail in medium radius curve.

In addition, the investigation included the analysis of a rail sample that was replaced due to a fracture after eight years of continuous use. Figure 3 illustrates the cross-section of the affected rail, showing initial signs of cracking by shelling and subsequent progression due to fatigue.

Figure 3 - Cross-section of the rail sample used in this research.

Source: Author.

The following data are from this sample:

- Estimated annual MTBT: ~ 50;
- Axle for transported load: 32 tons;
-Flow: Heavy Haul loaded;
- Installation position: Tight radius curved outer rail.

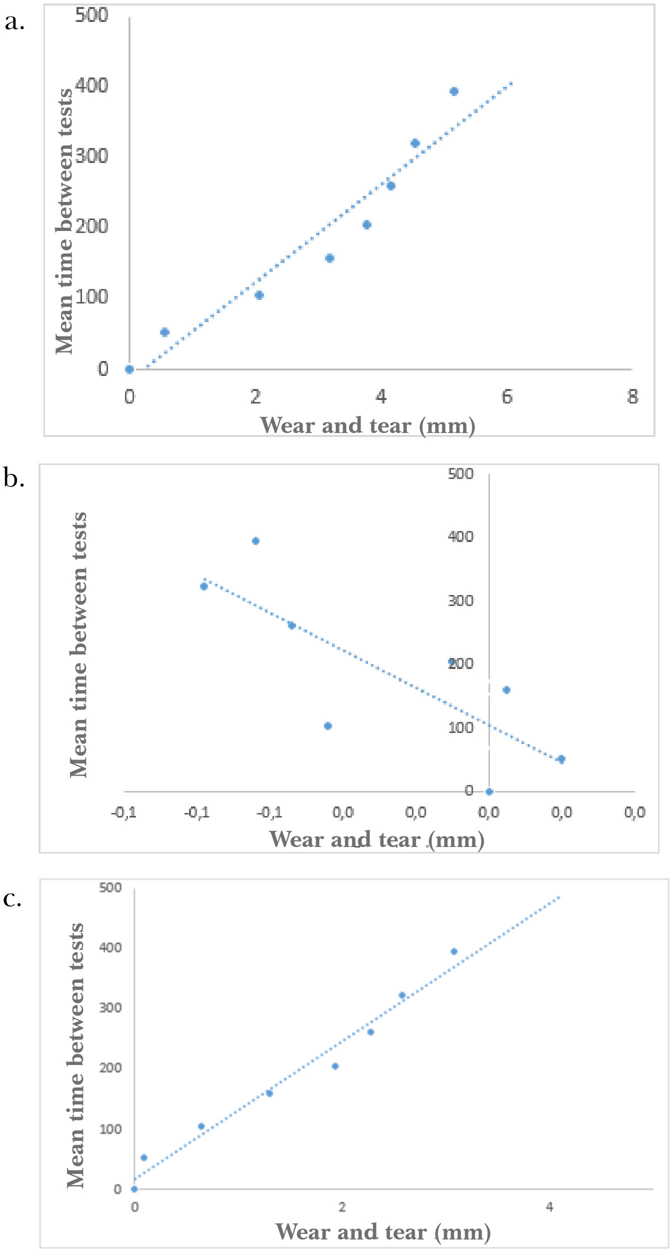
4. Results and discussion

Figure 4 illustrates the scatter plots and trend lines that were plotted using Excel to examine possible visual relationships between MTBT and wear and tear on railways, considering each assessed wear and tear position.

Note that, in the **W1** and **W3** (Figures 4a and 4c) positions, there is a linear relationship between the two variables. In contrast, for the position **W2** (Figure 4b), according to the data collected in the field, the observed wear and tear was almost non-existent, resulting in negative values in the MiniProf. Studies have shown that the application of the third body in the wheel rail contact can eliminate the lateral wear generated in curves with radio between 437 and 159 m [16].

Table 2 shows the degree of association between the variables calculated, which determined the value of Pearson's correlation coefficient (**r**), the coefficient of determination (**r²**), and the adjusted coefficient of determination.

Figure 4 - Scatter plots with trend lines for positions (a) w1, (b) w2 and (c) w3.



Source: Author.

When analyzing Pearson's correlation coefficients (**r**), determination (**r²**), and the adjusted determination shown in Table 2, note that the values for **W1** and **W3** are very close to 1, suggesting a strong and positive linear correlation between the variables (Table 1). On the other hand, for **W2**, a positive linear correlation of moderate intensity is observed (Table 1), which is in line with the pattern in the scatter plot for this position (Figure 4b). This means that for **W2** position, approximately 79% of the variation in the dependent variable can be explained by the independent variables included in the model. To the positions **W1** and **W3**, this explainability is higher, exceeding 90%.

Table 2 - Values of the Person coefficients (r), determination (r2) and adjusted determination for the three wear and tear positions.

Position	r2	r	Adjusted r2
W1	0.9287035	0.9636926	0.916820753
W2	0.624447	0.7902196	0.561854795
W3	0.9746352	0.9872361	0.970407679

Source: Author.

Table 3 described the estimated values for the parameters α and λ obtained with the support of Excel, derived from the linear regression model, as well as the associated standard error.

Table 3 - Linear regression coefficients.

	Model	Coefficients	Standard error
W1	**MTBT**	15.485	26.736
	Wear and tear	69.279	7.8365
W2	**MTBT**	104.32	41.068
	Wear and tear	2941.5	931.29
W3	**MTBT**	16.545	41.068
	Wear and tear	114.675	931.29

Source: Author.

The equations shown in Table 3 associated with the adjusted regression model, obtained with the support of Excel, are as follows:

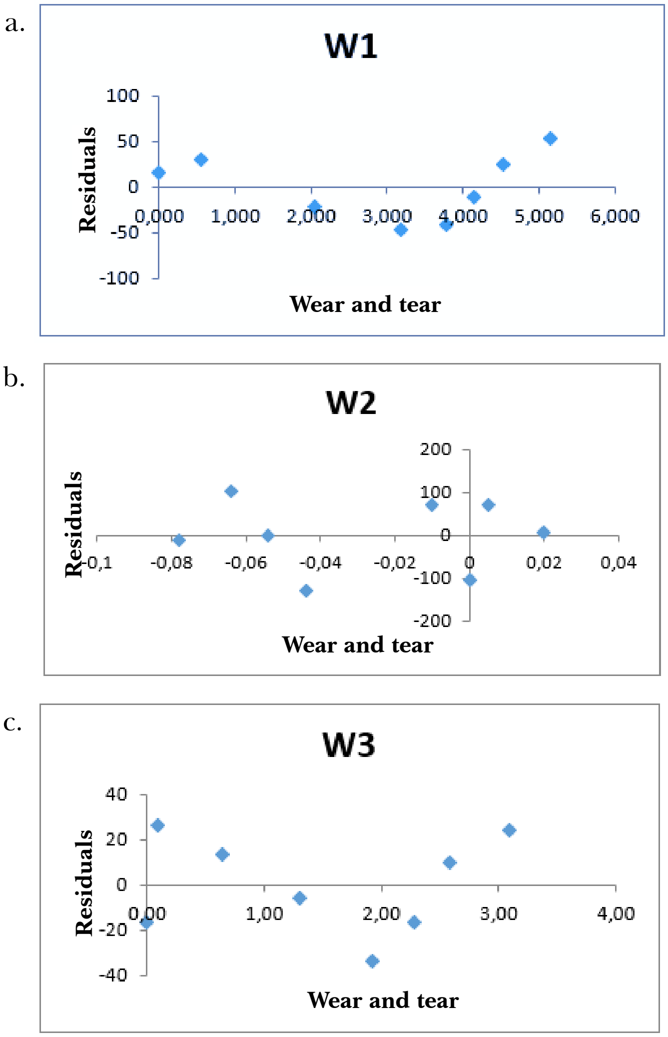
$$W1,\ MTBT = 69.279 Wear\ and\ tear - 15.485 \quad (5)$$

$$W2, MTBT = -2941.5 Wear\ and\ tear + 104.32 \quad (6)$$

$$W3, MTBT = 114.67 Wear\ and\ tear + 16.545 \quad (7)$$

Figure 5 shows graphs of residuals in relation to rail wear and tear. Residuals are the differences between the observed values and the values predicted by the regression model. Analysis of these results can provide valuable insights into the model's suitability for the actual data [28].

Figure 5 - Graphs of residuals of the three wear and tear positions under analysis, in which (a) refers to W1, (b) refers to W2, and (c) refers to W3.



Source: Author.

Analyzing the graphs of the residuals, it is verified that they have an asymmetric distribution. The graphs referring to the **W1** and **W3** positions (Figures 5a and 5c, respectively) show a positive asymmetry, whereas the graph for the **W2** position (Figure 5b) has a negative asymmetry.

Asymmetry in the residuals may indicate that the linear regression model is not ideal to describe the relationship between the variables or that there is no linearity or heteroscedasticity not captured by the model. This may point to the need for adjustments, such as the inclusion of nonlinear terms, variable transformations, or consideration of other forms of modeling [28].

Table 4 shows the ANOVA results (p-value) for each of the positions that were verified after the analysis of the residuals.

Table 4 - ANOVA results.

Variable	p-value
W1	0.0001
W2	0.0196
W3	5.15E-06

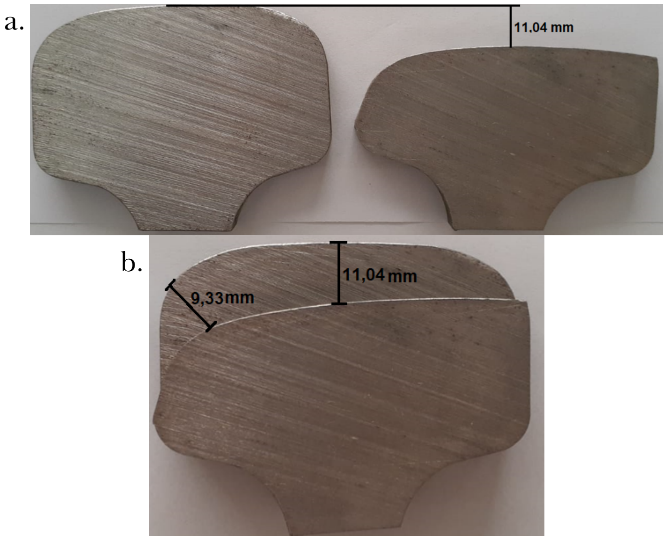
Source: Author.

For p-values above 0.050, we do not have evidence to support the assertion that the linear regression model is significant to explain the variable of interest, indicating that independent variables do not influence the dependent variable. However, if the p-value is ≤ 0.050, we can infer that at least one of the variables is relevant to explain the dependent variable, i.e., there is an influence of at least one independent variable on the dependent variable [28]. So, when examining Table 4, all p-values are less than 0.05. Therefore, we conclude that regression is statistically significant, indicating the importance of the linear regression model to explain the dependent variable.

Using Equations 5 and 7 and excluding the data referring to the **W2** position due to the low wear rate observed during the analysis period—as evidenced in Figure 4b—the estimated calculation for the wear and tear profile of the railway that was in operation for eight years was performed (Figure 3). The results showed an estimated wear and tear of 5.82 mm in the **W1** position and 3.24 mm in the **W3** position.

Figure 6 shows that the actual wear and tear experienced by this rail was measured by a caliper and a protractor to obtain the values of **W1** and **W3**.

Figure 6 - (a) Profile of the new and used rail side by side and (b) profile of the new and used rail superimposed with the appropriate measurements of W1 and W3.



Source: Author.

When evaluating the calculated values for **W1** (5.82 mm) and **W3** (3.24 mm) using Equations 5 and 7 and comparing them with the actual values recorded for the rail being used (**W1** = 11.04 mm and **W3** = 9.33 mm), there is a considerable discrepancy. This disparity can be attributed to the specific location of each rail on the railway. The section used to derive the equation has a curve with a mean radius of more than 900 m and approximately 12 MTBT per month. In contrast, the sample that provided the actual values of **W1** and **W3** (Figure 3) was installed in a section with a tighter radius curve, less than 250 m and approximately 4 MTBT per month. In addition to the divergences in the geometric design of the road, there is a significant disparity in the volume of cargo transported.

However, despite the discrepancy in the absolute values found, the relative values remained consistent with the actual data, i.e., **W1** > **W3**.

Thus, note that, in this case, a single equation based on simple linear regression for each position of analysis (**W1**, **W2**, and **W3**) is inadequate to describe the behavior of the entire railroad under study. The most appropriate would be to obtain equations using the multivariable regression methodology that would consider several parameters, such as radius of curvature, MTBT, ramp, and lubrication. Additionally, due to the particularities in each section of the railway, it would be advisable to show specific equations for each region, considering their individual characteristics.

It is relevant to highlight that methodologies related to modifications in wheel-rail contact, such as changes in wheel and rail profiles, the application of friction modifiers, and the use of grease were not included in the statistical calculations.

5. Conclusion

- The measured values of wear and tear of the rail that showed a fracture were higher than those calculated for the same rail using the equations developed in this study. However, the correlation indicating that **W1** is greater than **W3** was consistently observed;
- The simple linear regression approach adopted failed to fully capture all factors in the rail infrastructure that affect rail wear and tear. Therefore, there is a need to refine the model by including other variables that are significant to explain the observed phenomenon;
- Wear and tear is intrinsically linked to the conditions of the environment in which the asset is inserted. Therefore, the discrepancy between the values measured on the fractured rail and those obtained by the equation can be attributed, in part, to the location of the rails. The fractured rail was in an environment with conditions completely different from those of the rail used as a reference for the formulation of the equations. This may indicate that the same equation, for each of the wear and tear positions (**W1**, **W2**, and **W3**), is not adequate to describe an entire railway in detail.

References

[1] Butini, E. *et al*. Development and validation of wear models by using innovative three-dimensional laser scanners. **Advances in Mechanical Engineering**, Thousand Oaks, v. 11, n. 8, p. 1687814019870402, 2019.

[2] Jin X. *et al*. A numerical method for prediction of curved rail wear. **Multibody Syst Dyn**, Berlin, v. 18, p. 531-557, 2007.

[3] Meghoe A.; Loendersloot R.; Tinga T. Rail wear and remaining life prediction using meta-models. **International Journal of Rail Transportation**, Abingdon, v. 8, n. 1, p. 1-26, 2019.

[4] Westgeest F. P.; Dekker R.; Fischer R. H. Predicting rail geometry deterioration by regression models. *In*: Berenguer C.; Grall A.; Soares, C. G. (Eds.). **Advances in safety, reliability and risk management**. London: CRC Press. p. 926-933.

[5] Wang J. *et al*. Study on rail profile optimization based on the nonlinear relationship between profile and wear rate. **Mathematical Problems in Engineering**, [*s. l.*], 2017.

[6] AL-Juboori, A. *et al*. Evolution of rail surface degradation in the tunnel: the role of water on squat growth under service conditions. **Engineering fracture mechanics**, Amsterdam, v. 209, p. 32-47, 2019.

[7] Ahac M.; Lakusˇic´ S. Track gauge degradation modelling on small urban rail networks: zagreb tram system case study. *In*: Yaghoubi, H. (Ed.). **Urban transport systems**. Rijeka: InTech, 2017. p. 1-20.

[8] Soleimani H.; Moavenian M. Tribological aspects of wheel–rail contact: a review of wear mechanisms and effective factors on rolling contact fatigue. **Urban Rail Transit**, Berlin, v. 3, n. 4, p. 227-237, 2017.

[9] Li, Z. *et al*. An investigation into the causes of squats – correlation analysis and numerical modeling. **Wear**, Amsterdam, v. 265, n. 9-10, p. 1349-1355, 2008.

[10] Zoeteman, A.; Dollevoet, R.; Li, Z. Dutch research results on wheel/rail interface management: 2001–2013 and beyond. **Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit**, Thousand Oaks, v. 228, n. 6, p. 642-651, 2014.

[11] Sadeghi J.; Akbari B. Field investigation on effects of railway track geometric parameters on rail wear. **Journal of Zhejiang University-SCIENCE A**, Berlin, v. 7, n. 11, p. 1846-1855, 2006.

[12] Olofsson, U.; Telliskivi, T. Wear, plastic deformation and friction of two rail steels – a full-scale test and a laboratory study. **Wear**, Amsterdam, v. 254, n. 1-2, p. 80-93, 2003.

[13] Stock, R.; Pippan, R. RCF and wear in theory and practice—The influence of rail grade on wear and RCF. **Wear**, Amsterdam, v. 271, n. 1-2, p. 125-133, 2011.

[14] Grassie, S. Fatigue failure is often the hidden hazard. **International Railway Journal**, Falmouth, v. 37, n. 2, 1997.

[15] Porcaro, R. R. Efeito da soldagem elétrica por centelhamento na microestrutura, propriedades mecânicas e resistência à fadiga de um aço para trilhos ferroviários - análise experimental e numérica. 2019. Tese (Doutorado em Engenharia de Materiais) – REDEMAT, Ouro Preto, 2019.

[16] Sroba, P. *et al*. The evolution of rail grinding on Canadian Pacific Railway to address deep seated shells in 100% effective lubrication territories. *In*: World Congress on Railway Research, Montreal, Quebec, Canada. 2006, Paris. **Anais** […]. Paris: International Union of Railways, 2006.

[17] Jakob Oertli, S. B. B. **Combating Curve Squeal Phase II**. Paris: International Union of Railways, 2005.

[18] Hardwick, C.; Lewis, R.; Eadie, D. T. Wheel and rail wear—Understanding the effects of water and grease. **Wear**, Amsterdam, v. 314, n. 1-2, p. 198-204, 2014.

[19] Ferrer, B. P. Avaliação em laboratório do efeito da formulação e das propriedades de graxas lubrificantes no desempenho tribológico do contato roda-trilho. 2020. Dissertação (Mestrado em Engenharia Mecânica e de Materiais) – Universidade Tecnológica Federal do Paraná, Curitiba, 2020.

[20] Hutchings, I.; Shipway, P. **Tribology**: friction and wear of engineering materials. Cambridge: Butterworth--heinemann, 2017.

[21] Stock, R.; Schwarzenberger, R.; Kubin, W. Rail milling as an integrated strategy for modern rail asset management. *In*: CORE 2018: Rail: Smart, Automated, Sustainable, Conference on Railway Excellence, 30 April-2 May 2018, Sydney, NSW, Australia. [*s.n.],* 2018.

[22] Steyn, E. **Rail surface treatment–grinding and milling–competing or complementary technologies**. 2019. Dissertação (Mestrado em Engenharia de Materiais) – Chalmers University Of Technology, Gothenburg, 2019.

[23] Rodrigues, S. C. A. **Modelo de regressão linear e suas aplicações**. 2012. Relatório de Estágio – Universidade da Beira Interior, Covilhã, 2012.

[24] Santos, C. M. A. **Estatística Descritiva** – Manual de auto-aprendizagem. Lisboa: Edições Sílabo, 2007.

[25] Curral, J. **Statistics Packages**: A General Overview. Glasgow: Universidade de Glasgow, 1994.
[26] Barbetta, P. A. **Estatística aplicada às Ciências Sociais**. 5. ed. Florianópolis: UFSC, 2003.
[27] Yamauti, M. M. **Regressão linear simples nos livros de estatística para cursos de Administração**: um estudo didático. 2013. Dissertação (Mestrado em Educação Matemática) – Pontifícia Universidade Católica de São Paulo, São Paulo, 2013.
[28] Moreira, M. S. *et al*. **Regressão linear simples e múltipla**. Bioestatística quantitativa aplicada. Porto Alegre: UFRGS, 2020.

PÓS-GRADUAÇÃO NO IME

Bolsas: CAPES, CNPq e FAPERJ

Mestrado

- Sistemas e Computação
- Engenharia Mecânica

Mestrado e Doutorado

- Química
- Ciência dos Materiais
- Engenharia de Defesa
- Engenharia Nuclear
- Engenharia Elétrica
- Engenharia de Transportes

Matrículas: Semestral

Tel.: (21) 2546-7114 – Fax: (21) 2546-7089 – www.ime.eb.br – sd1@ime.eb.br