Histopathology

Histopathology 2024, **85**, 155–170. DOI: 10.1111/his.15187

Histological interpretation of spitzoid tumours: an extensive machine learning-based concordance analysis for improving decision making

Andrés Mosquera-Zamudio,[1,2] Laëtitia Launet,[3] Adrián Colomer,[3,4] Katharina Wiedemeyer,[5] Juan C López-Takegami,[6] Luis F Palma,[6] Erling Undersrud,[7] Emilius Janssen,[7,8] Thomas Brenn,[5] Valery Naranjo[3] & Carlos Monteagudo[1,2]

[1]Universitat de València, [2]INCLIVA, Instituto de Investigación Sanitaria, [3]Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, HUMAN-tech, Universitat Politècnica de València, [4]valgrAI: Valencian Graduate School and Research Network of Artificial Intelligence, Valencia, Spain, [5]Department of Pathology and Laboratory Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada, [6]Grupo de investigación IMPAC, Fundación Universitaria Sanitas, Bogotá, Colombia, [7]Department of Pathology, Stavanger University Hospital and [8]Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway

Date of submission 10 November 2023
Accepted for publication 16 March 2024

Mosquera-Zamudio A, Launet L, Colomer A, Wiedemeyer K, López-Takegami J C, Palma L F, Undersrud E, Janssen E, Brenn T, Naranjo V & Monteagudo C

(2024) *Histopathology* **85**, 155–170. https://doi.org/10.1111/his.15187

Histological interpretation of spitzoid tumours: an extensive machine learning-based concordance analysis for improving decision making

The histopathological classification of melanocytic tumours with spitzoid features remains a challenging task. We confront the complexities involved in the histological classification of these tumours by proposing machine learning (ML) algorithms that objectively categorise the most relevant features in order of importance. The data set comprises 122 tumours (39 benign, 44 atypical and 39 malignant) from four different countries. BRAF and NRAS mutation status was evaluated in 51. Analysis of variance score was performed to rank 22 clinicopathological variables. The Gaussian naive Bayes algorithm achieved in distinguishing Spitz naevus from malignant spitzoid tumours with an accuracy of 0.95 and kappa score of 0.87, utilising the 12 most important variables. For benign versus non-benign Spitz tumours, the test reached a kappa score of 0.88 using the 13 highest-scored features. Furthermore, for the atypical Spitz tumours (AST) versus Spitz melanoma comparison, the logistic regression algorithm achieved a kappa value of 0.66 and an accuracy rate of 0.85. When the three categories were compared most AST were classified as melanoma, because of the similarities on histological features between the two groups. Our results show promise in supporting the histological classification of these tumours in clinical practice, and provide valuable insight into the use of ML to improve the accuracy and objectivity of this process while minimising interobserver variability. These proposed algorithms represent a potential solution to the lack of a clear threshold for the Spitz/spitzoid tumour classification, and its high accuracy supports its usefulness as a helpful tool to improve diagnostic decision-making.

Address for correspondence: C Monteagudo, Universitat de Valencia, Valencia, Spain. e-mail: carlos.monteagudo@uv.es

Abbreviations: ANOVA, Analysis of Variance; AS^{oid}T, Atypical Spitzoid Tumor; AST, Atypical Spitz Tumor; BAMS, BRAF-mutated and morphologically Spitzoid; DT, Decision Tree; GNB, Gaussian Naive Bayes; KNN, K-Nearest Neighbor; LR, Logistic Regression; ML, Machine Learning; NGS, Next Generation Sequencing; SM, Spitz Melanoma; SN, Spitz Nevus; S^{oid}M, Spitzoid Melanoma; ST, Spitz Tumor; SVM, Support Vector Machines; TILs, Tumor-infiltrating lymphocytes; WHO, World Health Organization; WSI, Whole Slide Image.

Keywords: computer-aided diagnosis, histopathology, machine learning, melanocytic tumours, spitzoid tumours

Introduction

Sophie Spitz described a series of 13 children with melanocytic tumours with histopathological features of malignancy but favourable outcomes, except for one case with a fatal result. In this study she identified four questions: two related to clinical factors and treatment and the other two related to histological features, as useful tools to outline differences from conventional melanoma and as a marker of clinical behaviour. Initially, she coined the term 'juvenile melanomas' for these tumours, although nowadays we know that they also appear in older populations.[1] Currently, they are known as Spitz tumours (ST) when, as well as a typical morphology (large epithelioid and/or spindle melanocytic cells with variable nuclear atypia), they harbour HRAS mutations or kinase gene fusions but no BRAF or NRAS mutations. Conversely, tumours with the same morphology with no knowledge of the genetic alterations and/or the presence of BRAF or NRAS mutation can be categorised as spitzoid tumours.[2–4]

Spitz tumours must be categorised as benign, malignant and a third category of high diagnostic challenge lying in between: Spitz naevus (SN), Spitz melanoma (SM) and Spitz melanocytoma/atypical Spitz tumour (AST). The same subclassification is applied to the spitzoid tumours SN, $\mathrm{S^{oid}M}$ and $\mathrm{AS^{oid}T}$.[2,5]

One of the most challenging categories to identify is the AST.[5,6] This group makes up only 2% of all ST, and while most have a positive outcome, there is a small percentage that can result in fatal consequences and distant metastasis.[7,8] However, it remains unclear how to distinguish those with malignant clinical behaviour from the others.

Despite all the molecular discoveries that help to understand ST as a distinct group with specific mutagenic driver alterations, histological appearance continues to be the first and often the only useful tool for diagnostic interpretation worldwide. However, it has important shortcomings that are both a major concern and a major challenge in dermatopathology. One of the most important weaknesses of histological examination is the high interobserver variability.[9,10]

In addition, the implementation of a complete 'all-inclusive' study is hampered by limited access to perform complex and expensive molecular studies, especially in pathology laboratories with limited economic resources.

In this regard, the questions related to histopathology that Sophie Spitz mentioned are still relevant and have not yet been completely clarified. In the scientific literature, attempts to subclassify this group of tumours on the basis of histological features create flexible boundaries, where the same features appear with variable oscillation or with more prominence in one of the three ST classes mentioned above, and therefore subjectivity remains an essential part of this diagnostic interpretation.[2,5,11] Essentially, the challenge lies in the fact that the more than 20 histological features used for diagnosing STs have not yet been objectively prioritised or systematically evaluated to ascertain their impact on the histopathological diagnosis.

Conversely, artificial intelligence (AI) applied to histopathology (known as computational pathology) has shown significant benefits in increasing the efficiency and accuracy of pathologists' diagnosis, providing quantitative measurements of biomarkers to classify diseases into subtypes and predict outcomes, reducing the interobserver variability in differentiating benign from malignant tumours and their grading.[12] Although the results that AI shows are promising, only a small fraction of all these studies are approved for clinical purposes. This is due mainly to the lack of generalisability of their methodologies as one of the most common problems.[13]

In this study, we propose a machine learning (ML) model based on an analysis of variance (ANOVA) according to different clinicopathological variables commonly used for the diagnosis of STs to objectively characterise, in order of relevance, the most important features according to the algorithm tested. Therefore, we attempt to demonstrate that with the interpretation shown by the model, pathologists could improve the certainty of using some particular features with more diagnostic significance and reduce the relevance of other characteristics for a proper classification of STs, and evaluate the utility of these algorithms in predicting low or high grade within the AST category.

Materials and Methods

ETHICAL ISSUES

The Ethics Committees of the University Clinic Hospital (Valencia, Spain), Stavanger University Hospital (Stavanger, Norway), Cumming School of Medicine (Calgary, Canada) and Fundación Universitaria Sanitas (Bogotá, Colombia) approved the study (2020/114, 2019/747/RekVest, 2117-21) as part of the Clarify Project (clarify-project.eu) from the European Commission's Horizon 2020 Program for Research and Innovation, under the Marie Skłodowska-Curie grant agreement no. 860627, which was conducted in conformity with the principles of the Declaration of Helsinki.

SAMPLE DATA

This study was conducted in the Pathology Department at the University Clinic Hospital of Valencia, Spain. We collected ST cases from four institutions of different countries (Spain, Norway, Canada and Colombia) diagnosed by expert dermatopathologists (C.M., E.U., T.B. and J.C.L-T., respectively). All specimens had been formalin-fixed, paraffin-embedded and processed in each pathology laboratory according to standardised institutional protocols. Clinical and histopathological variables were obtained from each institution and gathered together with the whole slide image (WSI) of each case on which to corroborate the diagnosis.

Cases were first divided into three categories based solely upon their histopathological and immunohistochemical features (HMB-45, Ki67 and p16), when available: SN, AS^{oid}T, and S^{oid}M (see Figure 1, Tables 1 and 2), according to the World Health Organisation (WHO) classification of skin tumours.[2] In addition, we divided the AS^{oid}T category according to its histological appearance: low-grade, when it was closer to SN and high-grade if closer to melanoma. When classifying these tumours as low- or high-grade, it is important to note that some subjectivity may be unavoidable due to differences in how clinical and histological characteristics are interpreted. Nevertheless, we emphasised the presence of four to six mitoses/mm^2, atypical mitosis, expansile nests and ulceration in order to subclassify them as high-grade.[2,14] Specifically, we considered high-grade AS^{oid}T when at least two of the following criteria were present: four to six mitoses/mm^2, atypical mitosis, expansile nests, marked nuclear pleomorphism and ulceration (Table 1).

CDKN2A status was assessed by fluorescence *in-situ* hybridisation (FISH) using a 9p21 probe (Vysis LSI 9p21; Abbot Molecular Inc., Des Plaines, IL, USA), labelled with spectrum-red 2824 fluorophore, and a Chr9 centromeric probe (Vysis LSI CEP; Abbott) labelled with spectrum-green fluorophore and with Salsa® methylation-specific multiplex ligation-dependent probe amplification (MLPA) probemix ME024-B2 9p21 CDKN2A/2B region (MRC-Holland, Amsterdam, the Netherlands).[15] We further obtained information on BRAF and NRAS mutation status by immunohistochemistry (anti-BRAF V600E, clone: VE1; Ventana®; anti-RAS Q61R, clone: SP174 Abcam®, respectively) and/or next-generation sequencing (NGS) for atypical and malignant lesions. Cases without BRAF or NRAS mutation were considered as bona fide AST and SM.

MACHINE LEARNING ASSESSMENT AND STATISTICAL ANALYSIS

Model building and statistical computations were performed using Python (https://www.python.org/) in the Jupyter Notebook platform (https://jupyter.org/) with seaborn, NumPy, scikit-learn and pandas libraries.

The primary objective of this study is to evaluate the effectiveness of an ML model in classifying spitzoid tumours to assist pathologists in the objective use of clinical and histological information. The first step aimed to distinguish between benign and malignant spitzoid tumours, as well as between both categories (low- and high-grade) of the AS^{oid}T. In addition, we evaluated the performance of a multiclass ML model in distinguishing between SN, AS^{oid}T and S^{oid}M. Secondly, as the Spitz lineage is currently genetically defined and simulating the histopathological diagnostic process, the study focused upon the comparison between benign (SN) and non-benign (AST + SM) tumours, followed by the comparison between both non-benign categories (AST versus SM) (Figure 2).

ML algorithms

To carry out these objectives, we selected a range of ML algorithms; namely, logistic regression (LR), Gaussian naive bayes (GNB), support vector machines (SVM), decision tree (DT) and K-nearest neighbour (KNN), to evaluate the ML model's capability to predict the subclassification of these tumours based on the tabulated clinicopathological variables as input. Each selected algorithm provides different advantages in handling the complexity of the task and the nature

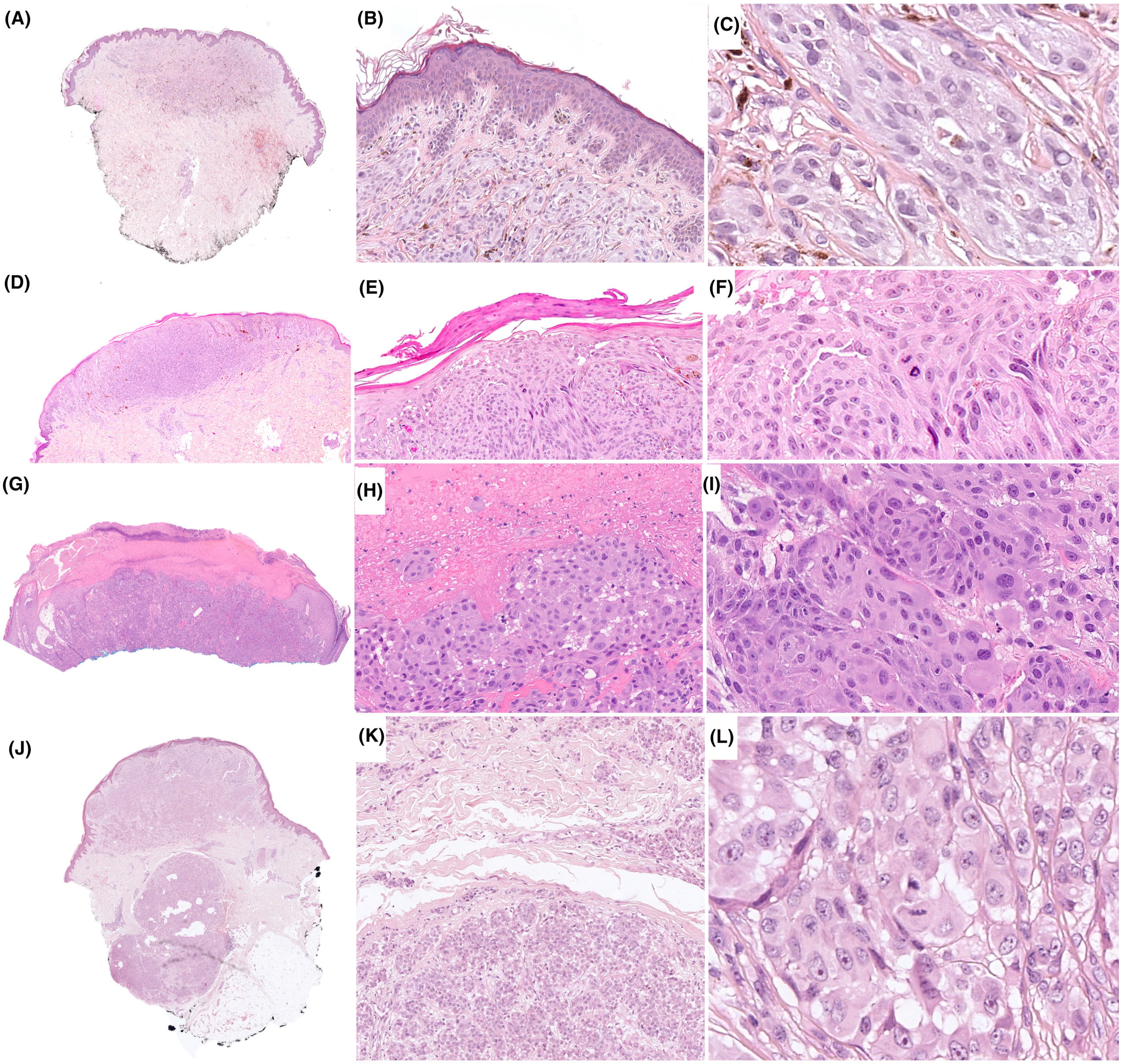

Figure 1. Spitzoid tumours: **A–C**. SN. **D–F**, Low-grade AS^{oid}T with epidermal consumption and mitosis. **G–I**, High-grade AS^{oid}T with ulceration and more prominent cellular atypia. **J–L**, (S^{oid}M) with penetrating expansile mass composed of atypical spitzoid melanocytes and deep mitosis.

of the data set and enables the development of efficient models, even with limited data.

Algorithm evaluation

To evaluate our respective models, multiple evaluation metrics were selected, including kappa score, accuracy, F1-score, precision, sensitivity and specificity. The kappa score was evaluated to determine concordance with pathologists' diagnoses. It was chosen as the main reference metric for optimising and fine-tuning the models because of its particular relevance in the medical field. This metric can be interpreted as follows: a kappa score of less than 0.20 indicates poor agreement, 0.21–0.40 suggests fair agreement, 0.41–0.60 denotes moderate agreement, 0.61–0.80 signifies substantial agreement and higher than 0.80 represents almost perfect agreement. Finally, confusion matrices were also computed to facilitate visualisation of the model's performance.

Table 1. Histological classification of spitzoid tumours

Feature	SN	Low-grade AS^{oid}T	High-grade AS^{oid}T	S^{oid}M
Dimensions	< 5–6 mm	> 5–10 mm	> 5–10 mm	> 5 mm often > 10 mm
Symmetry	Present	May be present	May be present	Uncommon
Circumscription	Well demarcated	Well or poorly demarcated	Well or poorly demarcated	Poorly demarcated
Ulceration	**Uncommon**	**Uncommon**	**May be present**	**May be present**
Expansile nests	**Absent**	**Absent**	**May be present**	**Common**
Pagetoid pattern	If any, central and focal	Greater than SN	Greater than SN	Extensive
Epidermal consumption	Extremely Uncommon	May be present	May be present	Common
Maturation	Present	Partial or absent	Partial or absent	Uncommon
Mitotic rate	< **2**/mm^2	**2–3**/mm^2	**4–6**/mm^2	> **6**/mm^2
Atypical mitosis	**Absent**	**Absent**	**May be present**	**Common**
Deep mitosis	Extremely uncommon	May be present	May be present	Yes
Necrosis	Absent	Uncommon	Uncommon	May be present
Cell type	Enlarged epithelioid/ spindle	Enlarged epithelioid/ spindle	Enlarged epithelioid/ spindle	Enlarged epithelioid/ spindle
High-grade nuclear atypia	Absent	May be present	May be present	Common
Peritumoural TILs	May be present	May be present	May be present	May be present
Intratumoural TILs	May be present	May be present	May be present	May be present
Pigmentation	May be present	May be present	May be present	May be present
Prominent solar elastosis	Uncommon	May be present	May be present	More common than AST
Pulverocytes	Extremely uncommon	Extremely uncommon	May be present	May be present
Kamino bodies	May be present	May be present	May be present	Extremely uncommon
Sclerosis/desmoplasia	May be present	May be present	May be present	May be present

The features in bold type were considered more relevant for distinguishing low-grade AS^{oid}T from high-grade AS^{oid}T.

Model's optimisation

To determine the base models for each of the specific approaches, we conducted a thorough comparison of the results obtained from the five selected ML algorithms. Prior to the comparison, each of the specific algorithms underwent a fine-tuning process to select the best parameters based on the average performance across cohorts in the cross-validation process. This optimisation allowed us to identify the models that reached the best performance in distinguishing three different comparisons: the first being SN versus S^{oid}M as illustrated in Figure 2A; the second, SN versus AST + SM (Figure 2B); and the third, AST versus SM (Figure 2C).

Training data

The training and validation process of the first experiment compared 39 SN versus 39 S^{oid}M from Spain; the second included 18 AST versus 18 SM cases; and the third included 38 SN versus 38 AST + SM from all three countries.

Analysis of clinicopathological variables

The ANOVA[16] was performed to rank the variables based on their importance and assess the differences between the three comparisons mentioned above and understand the variations among the clinicopathological variables when comparing two classes. Subsequently, we refined the model selection process by

Table 2. Immunohistochemical findings and CDKN2A deletion status (FISH/MLPA).

	p16			CDKN2A status			HMB-45 maturation			Ki67			
BRAF/NRAS status	No loss	Loss	NA	No/Del	HET/Del	HOM/del	NA	Present	Absent	NA	< 20%	≥ 20%	NA
Low-grade AS^{oid}T													
Non-mutated (*n* = 13)	8	1	4	4	1	0	8	3	0	10	5	0	8
Mutated (*n* = 0)	0	0	0	0	0	0	0	0	0	0	0	0	0
NA (*n* = 9)	4	3	2	0	0	0	9	6	1	2	7	0	2
Total (*n* = 22)	12	4	6	4	1	0	17	9	1	12	12	0	10
High-grade ASoidT													
Non mutated (*n* = 6)	2	1	3	1	0	0	5	4	0	2	4	0	2
Mutated (*n* = 2)	1	1	0	1	1	0	0	1	1	0	2	0	0
NA (*n* = 14)	6	2	6	0	0	0	14	10	1	3	11	0	3
Total (*n* = 22)	9	4	9	2	1	0	19	15	2	5	17	0	5
Total AS^{oid}T													
n = 44	21	8	15	6	2	0	36	24	3	17	29	0	15
S^{oid}M													
Non-mutated (*n* = 19)	2	8	9	1	1	1	16	0	3	16	6	2	11
Mutated (*n* = 11)	4	2	5	1	2	1	7	1	2	8	2	1	8
NA (*n* = 9)	3	3	3	0	4	0	5	1	1	7	2	2	5
Total (*n* = 39)	9	13	17	2	7	2	28	2	6	31	10	5	24

No/Del, no deletion; HET/Del, heterozygous deletion; HOM/Del, homozygous deletion; NA, not applicable.

iteratively reducing the feature set from the initial 22 variables, assessing the impact on model performance with each reduction. This systematic reduction of the variables aimed to determine the minimal feature subset necessary while still achieving a kappa score higher than 0.60. This approach allowed us to identify the threshold at which the model maintained its predictive accuracy, revealing the most impactful features for our analysis.

Results

The clinicopathological features are summarised in Table 3. A total of 122 cases of ST, including (a) SN (*n* = 39), which histologically corresponded to 10 junctional or compound conventional cases, 10 desmoplastic, 10 pigmented (but not Reed's naevi), three plexiform, two pagetoid, two halo Spitz, one hyalinising and one angiomatous; (b) AS^{oid}T (*n* = 44); and (c) S^{oid}M (*n* = 39). There is a higher prevalence of ST in females compared to males among all three categories. The mean age of patients in the S^{oid}M category is higher (45.64 years) compared to those in the SN (23.82 years) and AS^{oid}T (26.41 years) categories. SN is typically diagnosed in younger people, while the age range for AS^{oid}T falls between that of SN and S^{oid}M. The average clinical follow-up of the study was 73 months. For the 22 cases of high-grade AS^{oid}T,

Figure 2. Overall ML and statistical analysis pipeline: **A–C**, Histological variables were concatenated to clinical variables of the patient (age at diagnosis, sex, tumour location), and each case was labelled according to the pathologists' diagnosis. Using ANOVA, the clinicopathological features were stratified. Subsequently, a ML model was implemented to see the performance according to the ANOVA's score in predicting (**A**) SN versus S^{oid}M; high-grade AS^{oid}T versus low-grade (AS^{oid}T); (**B**) AST versus SM and (**C**) SN versus AST + SM. **D**, Multiclass ML model to predict SN, AS^{oid}T and S^{oid}M.

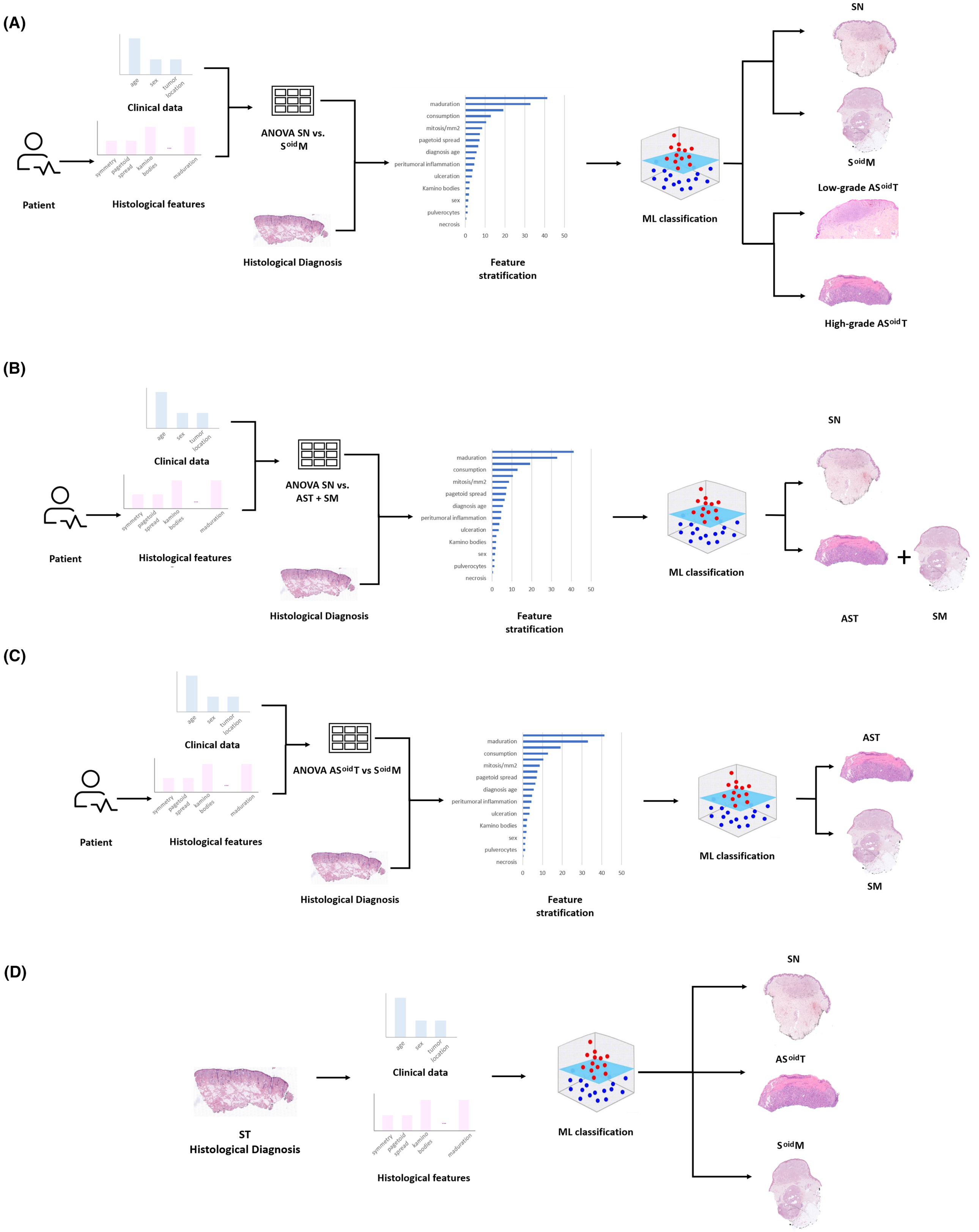
(A)
age
sex
tumor location
Clinical data
symmetry
pagetoid spread
kamino bodies
...
maduration
Histological features
Patient
ANOVA SN vs. S^{oid}M
Histological Diagnosis
maduration
consumption
mitosis/mm2
pagetoid spread
diagnosis age
peritumoral inflammation
ulceration
Kamino bodies
sex
pulvercoytes
necrosis
0
10
20
30
40
50
Feature stratification
ML classification
SN
S^{oid}M
Low-grade AS^{oid}T
High-grade AS^{oid}T
(B)
age
sex
tumor location
Clinical data
symmetry
pagetoid spread
kamino bodies
...
maduration
Histological features
Patient
ANOVA SN vs. AST + SM
Histological Diagnosis
maduration
consumption
mitosis/mm2
pagetoid spread
diagnosis age
peritumoral inflammation
ulceration
Kamino bodies
sex
pulvercoytes
necrosis
0
10
20
30
40
50
Feature stratification
ML classification
SN
+
AST
SM
(C)
age
sex
tumor location
Clinical data
symmetry
pagetoid spread
kamino bodies
...
maduration
Histological features
Patient
ANOVA AS^{oid}T vs S^{oid}M
Histological Diagnosis
maduration
consumption
mitosis/mm2
pagetoid spread
diagnosis age
peritumoral inflammation
ulceration
Kamino bodies
sex
pulvercoytes
necrosis
0
10
20
30
40
50
Feature stratification
ML classification
AST
SM
(D)
ST
Histological Diagnosis
age
sex
tumor location
Clinical data
symmetry
pagetoid spread
kamino bodies
...
maduration
Histological features
ML classification
SN
AS^{oid}T
S^{oid}M

Table 3. Clinicopathological data

Diagnosis	SN	Low-grade $AS^{oid}T$	High-grade $AS^{oid}T$	Total $AS^{oid}T$	$S^{oid}M$
Number of cases	39	22	22	44	39
Age					
Mean	23.82	24.31	30.45	27.38	45.64
Min	3	1	4	5	10
Max	85	78	70	148	91
Sex					
Female	26 (67%)	14 (64%)	13 (59%)	27 (61%)	22 (56%)
Male	13 (33%)	8 (36%)	7 (32%)	15 (34%)	17 (44%)
NA	0 (0%)	0 (0%)	2 (9%)	2 (5%)	0 (0%)
Tumour location					
Head and neck	5 (13%)	6 (27%)	6 (27%)	12 (27%)	3 (8%)
Trunk	11 (28%)	1 (5%)	1 (5%)	2 (5%)	1 (3%)
Upper limb	8 (21%)	6 (27%)	1 (5%)	7 (16%)	15 (38%)
Lower limb	8 (21%)	5 (23%)	7 (32%)	12 (27%)	11 (28%)
NA	7 (18%)	4 (18%)	7 (32%)	11 (25%)	9 (23%)
Symmetry	38 (97%)	20 (91%)	21 (95%)	41 (93%)	26 (67%)
Ulceration	0 (0%)	0 (0%)	1 (5%)	1 (2%)	6 (15%)
Pagetoid spread	6 (15%)	5 (23%)	7 (32%)	12 (27%)	19 (49%)
Expansile nests	1 (3%)	2 (9%)	3 (14%)	5 (11%)	18 (46%)
Number of cases with mitosis	2 (5%)	11 (50%)	11 (50%)	22 (50%)	29 (74%)
Mitosis/mm^2					
Mean	0.05	2.09	2.45	2.27	3.74
Min	0	1	0	1	0
Max	2	3	6	6	30
Atypical mitosis	0 (0%)	0 (0%)	4 (18%)	4 (9%)	23 (59%)
Deep mitosis	0 (0%)	3 (14%)	1 (5%)	4 (9%)	13 (33%)
Atypia/pleomorphism	4 (10%)	15 (68%)	19 (86%)	34 (77%)	35 (90%)
Maturation	39 (100%)	14 (64%)	12 (55%)	26 (59%)	8 (21%)
Kamino bodies					
0	33 (85%)	19 (86%)	21 (95%)	40 (91%)	35 (90%)
1	2 (5%)	0 (0%)	1 (5%)	1 (2%)	4 (10%)
2–5	2 (5%)	1 (5%)	0 (0%)	1 (2%)	0 (0%)
6–10	1 (3%)	1 (5%)	0 (0%)	1 (2%)	0 (0%)
> 10	1 (3%)	1 (5%)	0 (0%)	1 (2%)	0 (0%)

Table 3. (*Continued*)

Diagnosis	SN	Low-grade AS^{oid}T	High-grade AS^{oid}T	Total AS^{oid}T	S^{oid}M
Peritumoural TILs					
Absent	14 (36%)	5 (23%)	10 (45%)	15 (34%)	7 (18%)
Discontinuous	24 (62%)	14 (64%)	10 (45%)	24 (55%)	27 (69%)
Dense	1 (3%)	3 (14%)	2 (9%)	5 (11%)	5 (13%)
Intratumoural TILs					
Absent	22 (56%)	5 (23%)	9 (41%)	14 (32%)	12 (31%)
Weak	17 (44%)	17 (77%)	11 (50%)	28 (64%)	23 (59%)
Intense	0 (0%)	0 (0%)	2 (9%)	2 (5%)	4 (10%)
Epidermal consumption	1 (3%)	8 (36%)	5 (23%)	13 (30%)	19 (49%)
Prominent elastosis	1 (3%)	1 (5%)	1 (5%)	2 (5%)	2 (5%)
Pigmentation	14 (36%)	9 (41%)	8 (36%)	17 (39%)	18 (46%)
Sclerosis	6 (15%)	2 (9%)	1 (5%)	3 (7%)	4 (10%)
Type of cell					
Epithelioid	18 (46%)	17 (77%)	20 (91%)	37 (84%)	28 (72%)
Spindle	15 (38%)	1 (5%)	1 (5%)	2 (5%)	6 (15%)
epi/spin	6 (15%)	4 (18%)	1 (5%)	5 (11%)	5 (13%)
Pulverocytes	0 (0%)	0 (0%)	2 (9%)	2 (5%)	2 (5%)
Necrosis	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (3%)

TILs, tumour-infiltrating lymphocytes; epi/spin, epithelioid and spindle; NA, not applicable.

there was a 5% incidence of both local recurrence and regional relapse, with no distant metastasis or deaths. The 22 cases of low-grade AS^{oid}T showed no events in any category. Finally, among the 39 S^{oid}M cases, there was a 3% local recurrence rate, 18% regional relapse, 8% distant metastasis and a 5% death rate (see Table 4).

We were able to obtain information on BRAF and NRAS mutation status from 51 cases of AS^{oid}T and S^{oid}M (21 and 30, respectively). With regard to AS^{oid}T, we found that only one case (histologically high-grade) was BRAF mutated (BAMS)[3] and one case (also histologically high-grade) was NRAS mutated. Therefore, most (90%) of our AS^{oid}T can

Table 4. Clinical follow-up

Diagnosis	SN	Low-grade AS^{oid}T	High-grade AS^{oid}T	Total AS^{oid}T	S^{oid}M
Number of cases	39	22	22	44	39
Local recurrence	0 (0%)	0 (0%)	1 (5%)	1 (2%)	1 (3%)
Regional relapse	0 (0%)	0 (0%)	1 (5%)	1 (2%)	7 (18%)
Distant metastasis	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (8%)
Death	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (5%)

Table 5. BRAF V600E and NRAS Q61R mutation status

	Low-grade AS^{oid}T (n = 22)	High-grade AS^{oid}T (n = 22)	Total AS^{oid}T (n = 44)	S^{oid}M (n = 39)
BRAF V600E				
Mutated	0	1	1	9
Non-mutated	13	7	20	21
NA	9	14	23	9
NRAS Q61R				
Mutated	0	1	1	2
Non-mutated	11	4	15	27
NA	11	17	28	10

NA, not applicable.

be considered to be bona fide Spitz melanocytoma/AST. In contrast, nine of 30 S^{oid}M were BRAF mutated and two were NRAS mutated. Thus, the majority (63%) can be considered as true Spitz melanomas (SM) (Tables 2 and 5). The lack of information on the specific genetic drivers of the different subtypes of ST is a limitation of our study.

We collected data on p16 immunostaining from 51 cases of AS^{oid}T and S^{oid}M (29 of 44 and 22 of 39, respectively) and CDKN2A deletion status from 19 cases. Eight AS^{oid}T and S^{oid}M cases had loss of p16 immunostaining. Eight cases had no CDKN2A copy number alteration. Heterozygous deletion was found in nine cases and homozygous deletion in two cases (both SM). The low-grade AS^{oid}T cases with p16 loss were not upgraded because there was no CDKN2A homozygous deletion. These results, as well as those of HMB-45 and Ki67, are shown in Table 2. Although relevant, we finally decided not to include the immunohistochemical biomarkers because the limited availability of samples with this information hindered the ability to achieve an effective ML performance.

ML MODEL PERFORMANCE USING ANOVA

The best-performing ML model in classifying SN versus S^{oid}M using the 22 clinicopathological variables was the GNB model, with an accuracy of 0.95 and a kappa score of 0.87 (Table 6), thus making it the backbone model for further experiments for this prediction task. During the ANOVA analysis of the clinicopathological features, cellular atypia and pleomorphism reported the highest score (41.30), followed by maturation (32.88) and atypical mitosis (19.13). Conversely, necrosis, sclerosis and solar elastosis had the worst scores (0, 0.48 and 1.16, respectively), thus suggesting less relevance in the possible diagnosis, as depicted in Figure 3A. In the validation phase, considering the 12 features with the highest ANOVA significance scores, the GNB yielded identical results. When we narrowed it down to the top six features according to ANOVA scores, we noticed just a small difference in the kappa score (0.84), maintaining the same accuracy. This ML model was tested on an international cohort with four cases from Norway and four from Colombia (each with two SN and two S^{oid}M), together with four cases from Spain. The model reached a kappa score of 0.67 and an accuracy of 0.83 on the test set, as shown in Table 7.

In the comparison of SN versus AST + SM, the GNB model outperformed the other four machine learning models, achieving a kappa score of 0.72 and an accuracy of 0.87 (Table 6). Cellular atypia and pleomorphism was identified as the most important feature, similar to the previous comparison, with a score of 13.16. The next three most significant features included tumour location, with a score of 9.92, the presence of atypical mitoses at 8.93 and mitotic count/mm^2 at 7.61. During the validation phase, the model sustained a substantial level of agreement, evidenced by a kappa score of 0.65 and an accuracy of 0.83, using the 13 features with the highest scores from ANOVA analysis (Figure 3B). The model further improved in the test phase, achieving a kappa score of 0.75 and an accuracy of 0.88 (Table 7).

Furthermore, for the comparison between AST and SM, the LR model stood out as the most effective ML strategy. It attained a kappa score of 0.66 and an accuracy rate of 0.85 (Table 6). The kappa score remained within the range for substantial agreement (0.61) when the last six of the 22 features according to the ANOVA ranking for this group were discarded. This was achieved with a sensitivity of 0.82 and a specificity of 0.75. However, the performance significantly decreased after reducing the feature set to 15 characteristics (Table 7). The highest scores in this comparison were atypical mitosis (9.02) followed by maturation (7.89), age (7.20) and expansile nest (6.66) (Figure 3C).

MULTICLASS ML MODEL PERFORMANCE

The confusion matrix obtained regarding the classification of the three-class GNB model distinguishing

Table 6. Validation results of the models trained with the binary classifier

Model	Kappa	ACC	F1-score	Precision	Sensitivity	Specificity
SN versus S^{oid}M						
GNB	**0.87** ± **0.11**	**0.95** ± **0.04**	**0.95** ± **0.04**	**0.95** ± **0.04**	**0.93** ± **0.13**	**0.92** ± **0.10**
LR	0.83 ± 0.17	0.93 ± 0.07	0.93 ± 0.07	0.93 ± 0.07	0.86 ± 0.13	0.97 ± 0.05
DT	0.76 ± 0.23	0.91 ± 0.08	0.90 ± 0.09	0.92 ± 0.07	0.77 ± 0.29	0.90 ± 0.09
SVM	0.63 ± 0.31	0.84 ± 0.12	0.84 ± 0.12	0.84 ± 0.12	0.77 ± 0.23	0.87 ± 0.11
KNN	0.66 ± 0.23	0.86 ± 0.07	0.86 ± 0.08	0.86 ± 0.09	0.73 ± 0.21	0.93 ± 0.10
SN versus AST + SM						
GNB	**0.72** ± **0.08**	**0.87** ± **0.04**	**0.86** ± **0.04**	**0.89** ± **0.04**	**0.76** ± **0.10**	**0.97** ± **0.07**
LR	0.66 ± 0.10	0.83 ± 0.05	0.83 ± 0.05	0.87 ± 0.03	0.71 ± 0.09	0.97 ± 0.07
DT	0.43 ± 0.26	0.75 ± 0.12	0.68 ± 0.14	0.77 ± 0.14	0.67 ± 0.17	0.79 ± 0.12
SVM	0.52 ± 0.27	0.77 ± 0.13	0.77 ± 0.13	0.78 ± 0.14	0.70 ± 0.13	0.83 ± 0.14
KNN	0.55 ± 0.21	0.77 ± 0.12	0.76 ± 0.13	0.83 ± 0.08	0.58 ± 0.15	0.97 ± 0.07
AST versus SM						
GNB	0.25 ± 0.27	0.65 ± 0.14	0.59 ± 0.18	0.67 ± 0.24	1.00 ± 0.00	0.28 ± 0.28
LR	**0.66** ± **0.23**	**0.85** ± **0.09**	**0.84** ± **0.10**	**0.88** ± **0.08**	**0.92** ± **0.11**	**0.76** ± **0.22**
DT	0.37 ± 0.25	0.69 ± 0.14	0.71 ± 0.14	0.72 ± 0.14	0.63 ± 0.11	0.85 ± 0.30
SVM	0.61 ± 0.24	0.82 ± 0.11	0.82 ± 0.11	0.86 ± 0.09	0.68 ± 0.19	**0.96** ± 0.08
KNN	0.62 ± 0.37	0.82 ± 0.19	0.82 ± 0.18	0.85 ± 0.18	0.73 ± 0.23	0.90 ± 0.20

Note that a grid search for parameters' optimisation based on cross-validation results with the Cohen kappa's score was performed for each of the presented models. GNB, Gaussian naive Bayes; LR, logistic regression; DT, decision tree; SVM, support vector machines; KNN, K-nearest neighbour. The best model performance for each experiment is highlighted in bold.

between SN, AS^{oid}T and S^{oid}M revealed, first, that a high rate of correct predictions was observed for SN cases, with a total of 27 of 29 cases correctly classified. However, two SN instances were erroneously classified as AST. Conversely, the classification of AS^{oid}T cases showed more challenging performance with nine cases correctly classified, but also 11 cases misclassified as SN and nine cases as S^{oid}M. For S^{oid}M cases, a high rate of correct classification was observed, with 21 cases accurately identified. However, there were eight cases that were misclassified as AS^{oid}T (see Figure 4).

AS^{OID}T

ML model performance with binary ML model used in SN versus S^{oid}M

We evaluated the model's performance of the binary ML model used previously for SN versus S^{oid}M (see Figure 2D) by analysing the confusion matrix, which compared the predicted classifications of low-grade AS^{oid}T and high-grade AS^{oid}T lesions to their classifications according to the pathologists. Note that the AS^{oid}T data set used in these steps encompassed cases from four different institutions (Canada, Colombia, Spain and Norway). O the 22 cases of low-grade AS^{oid}T lesions, 11 were correctly classified as low-grade AS^{oid}T, while 11 were misclassified as high-grade AS^{oid}T. The majority of AS^{oid}T lesions (62%) were classified as malignant, indicating a potential risk for malignancy. However, the cases classified as benign highlight the difficulty in accurately distinguishing between malignant and benign AS^{oid}T lesions (Figure 4B).

Discussion

ML models have proved feasible in distinguishing between benign and malignant melanocytic

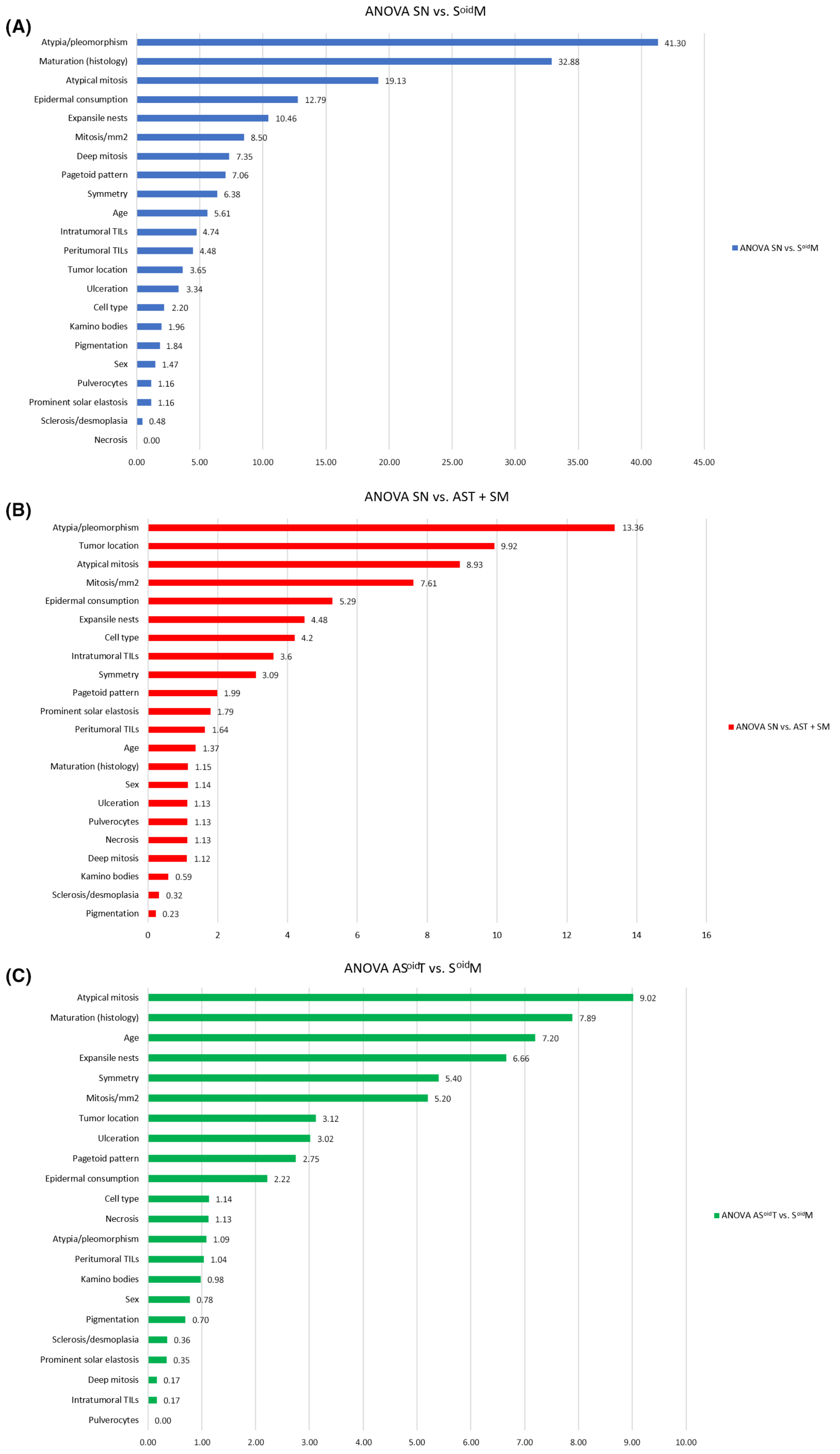

Figure 3. ANOVA score results for each predictor variable of our data set. TILs, tumour-infiltrating lymphocytes. [Colour figure can be viewed at wileyonlinelibrary.com]

Table 7. Validation of the models trained with the best binary classifier selected in the previous step, selecting subsets of the features depending on their ANOVA score (see Figure 3)

Features	Kappa	ACC	F1-score	Precision	Sensitivity	Specificity
SN versus $S^{oid}M$						
Validation						
22 (all)	0.87 ± 0.11	0.95 ± 0.04	0.95 ± 0.04	0.95 ± 0.04	0.93 ± 0.13	0.92 ± 0.10
12	**0.87 ± 0.11**	**0.95 ± 0.04**	**0.95 ± 0.04**	**0.95 ± 0.04**	**0.93 ± 0.13**	**0.92 ± 0.10**
6	0.84 ± 0.23	0.95 ± 0.07	0.94 ± 0.09	0.96 ± 0.06	0.87 ± 0.27	0.95 ± 0.10
Test (multinational)						
12	0.67	0.83	0.86	1.0	0.75	1.0
SN versus AST + SM						
Validation						
22	0.72 ± 0.08	0.87 ± 0.04	0.86 ± 0.04	0.89 ± 0.04	0.76 ± 0.10	0.97 ± 0.07
13	**0.65 ± 0.10**	**0.83 ± 0.05**	**0.83 ± 0.05**	**0.84 ± 0.05**	**0.80 ± 0.09**	**0.85 ± 0.08**
Test (multinational)						
14	0.88	0.88	0.88	0.88	0.88	0.88
AST versus SM						
Validation						
22	0.66 ± 0.23	0.85 ± 0.09	0.84 ± 0.10	0.88 ± 0.08	0.92 ± 0.11	0.76 ± 0.22
19	0.61 ± 0.28	0.82 ± 0.13	0.82 ± 0.13	0.84 ± 0.13	0.87 ± 0.11	0.75 ± 0.22
16	**0.61 ± 0.28**	**0.82 ± 0.13**	**0.82 ± 0.13**	**0.84 ± 0.13**	**0.87 ± 0.11**	**0.75 ± 0.22**
15	0.17 ± 0.26	0.53 ± 0.18	0.49 ± 0.22	0.56 ± 0.28	0.72 ± 0.27	0.45 ± 0.33

The SN versus $S^{oid}M$ and the SN versus AST + SM models were tested with external cases from Colombia and Norway, in addition to institutional cases from Spain. The best performance for each experiment is highlighted in bold.

tumours.[17] Additionally, the ANOVA analysis has proved to be a valuable tool for identifying and selecting the most significant features, enhancing the performance of ML models to achieve greater accuracy in various applications related to tumour diagnosis.[18–21]

As spitzoid tumours continue to pose significant challenges in dermatopathology, this study aimed to evaluate the effectiveness of ML models in distinguishing benign from malignant tumours, as well as predicting the subclassification of the atypical intermediate category, based on 22 clinicopathological features in different cohorts diagnosed by dermatopathologists from four different countries. The primary goal was to rank these features objectively according to their relevance, providing valuable insights to pathologists in identifying the most significant factors for accurate diagnosis.

It is well known that in many pathology laboratories worldwide no genetic information is available, but still the subclassification and especially the distinction between benign and malignant is mandatory. For this reason, we also implemented our ML models in order to categorise and rank the clinical and histological features that are available in all centres to select the most important variables in order to improving and saving time in the diagnostic decision-making process.

We demonstrate that our GNB ML algorithm, through simulation of the histopathology diagnostic workflow, can discriminate between benign and non-benign Spitz tumours based solely upon the 14

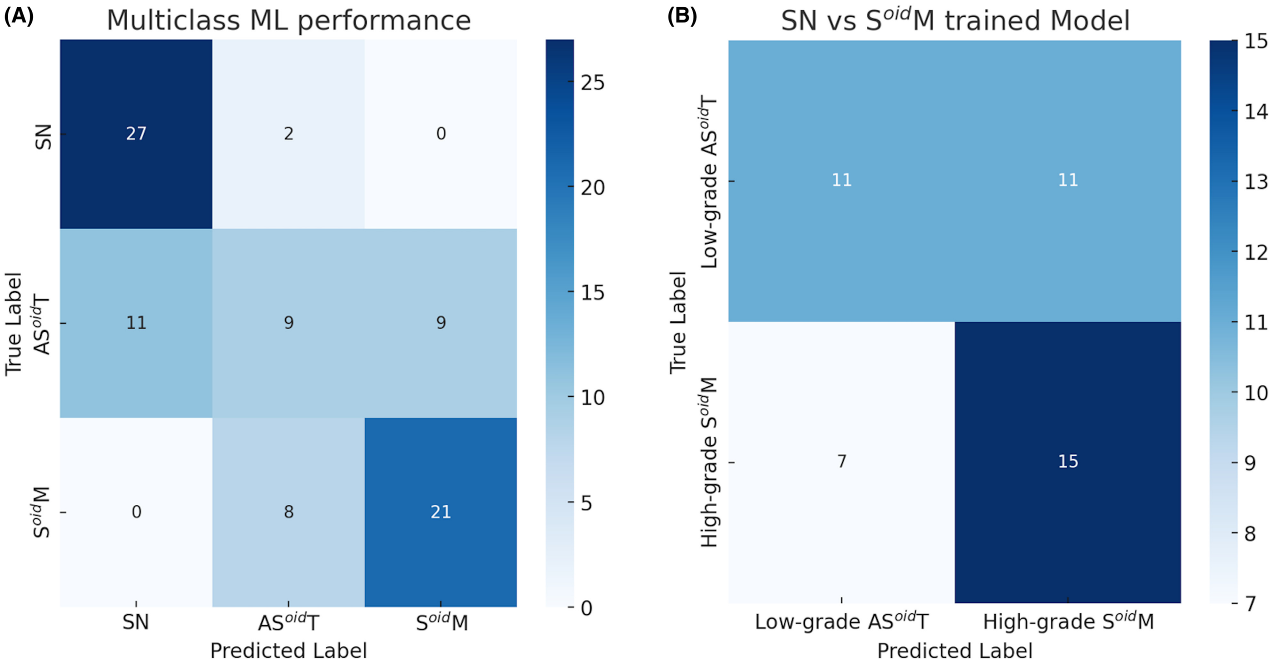

Figure 4. Confusion matrix of the ML models experiments including AS^{oid}T cases. **A**, Validation results of a multi-class model trained to differentiate between benign, malignant and AS^{oid}T. **B**, Validation of the AS^{oid}T model predictions using the binary model applied for SN versus S^{oid}M. [Colour figure can be viewed at wileyonlinelibrary.com]

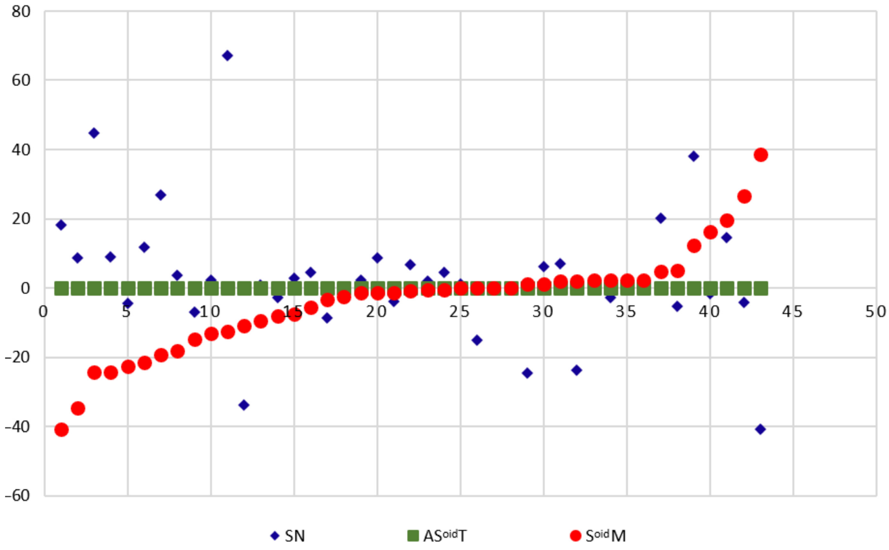

Figure 5. Dispersion map comparing the histological features between ASoidT, SN and S^{oid}M. [Colour figure can be viewed at wileyonlinelibrary.com]

most important clinical and histological features according to the ANOVA categorisation, achieving a test kappa score of 0.88 with high accuracy, sensitivity and specificity. Furthermore, our LR ML model is also effective in distinguishing AST versus SM with an acceptable substantial agreement maintaining very high performance in the other metrics. Therefore, with these two ML algorithms, we have shown that an objective evaluation and reduction of the clinicopathological variables can be conducted in order to make an accurate diagnosis.

When testing the ML model on cases from other countries between SN and S^{oid}M we observed a reduction in accuracy to 0.83, with the kappa score also decreasing to 0.67 (Table 7). However, these results still demonstrate an acceptable level of agreement and accuracy. This finding also aligns with existing literature that highlights the high interobserver variability observed in ambiguous tumours. Previous studies, such as Bhoyrul *et al.*[10] and Colloby *et al.*,[22] have reported similar difficulties in achieving consensus among pathologists. The impact of interobserver variability has also been highlighted in studies by Elmore *et al.*[23] and Lodha *et al.*[24]

During the validation of the AS^{oid}T using the binary model for the SN versus S^{oid}M comparison, we observed that a significant number of cases were classified as S^{oi}M. This implies that certain features present in these cases tend to resemble those of melanoma (S^{oid}M) rather than naevi (SN), leading to a diagnosis of a malignancy. The dispersion map in Figure 5, based on the results presented in Table 3, further supports this finding. Specifically, characteristics such as pagetoid spread, expansile nests, atypical mitosis and deep mitosis show a closer resemblance between AS^{oid}T and S^{oid}M than between AS^{oid}T and SN, indicating a higher similarity between AS^{oid}T and S^{oid}M. This convergence of characteristics between AS^{oid}T and S^{oid}M reinforces the rationale behind the experimental observations.

Our findings with the ANOVA categorisation, analysing among these three different comparisons, revealed some consistent trends in the most important features. Among the 22 features we looked at, atypical mitosis, atypia/pleomorphism and mitosis/mm^2 always ranked in the top six, highlighting their significant role in our predictive models (Figure 3).

In summary, this approach shows promising potential to facilitate the diagnosis of challenging melanocytic tumours with spitzoid features by utilising ML models to objectively categorise clinicopathological parameters. This approach not only simplifies the diagnostic process by minimising the number of variables needed, but also enhances diagnostic accuracy for these tumours with known or unknown BRAF/NRAS mutation status. Implementation of this method may potentially reduce interobserver variability between pathologists improving the morphological categorisation of these tumours, in order to facilitate their histopathological diagnostic interpretation. Despite our lack of information regarding the specific genetic drivers of the different subtypes of ST, our algorithms distinguish successfully between SN and S^{oid}M, SN and non-benign ST (AST + SM), as well as between AST and SM. Overall, this model offers promising implications for improving the clinical workflow and diagnostic practices in the field of dermatopathology.

Acknowledgements

This work has received funding from Horizon 2020, the European Commission's Framework Programme for Research and Innovation, under the grant agreement no. 860627 (CLARIFY), PI20/00094, Instituto de Salud Carlos III, FEDER European Funds and INNEST/2021/321 (SAMUEL). The work of A.C. has been supported by the ValgrAI—Valencian Graduate School and Research Network for Artificial Intelligence and Generalitat Valenciana and Universitat Politècnica de València (PAID-PD-22).The material from Norway has been collected as part of the project 'Pathology services in the Western Norway Health Region—a centre for applied digitisation' which is financed through a strategic investment from the Western Norway Health Authority.

Conflicts of interest

The authors declare no conflicts of interest. The founders had no role in the study's design; in the collection, analysis or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Spitz S. Melanomas of childhood. *Am. J. Pathol.* 1948; **24**; 591–609.
2. Gerami P, Bahrami A, Busam KJ, de la Fouchardière A, Kazakov DV, Massi D, et al. In: Elder D, Barnhill R, editors. Spitz Melanoma, Vol. **12** *Of WHO classification of tumours*, 5 ed. Lyon, France: International Agency for Research on Cancer; 2023. https://tumourclassification.iarc.who.int/chapters/64, [Internet; beta version ahead of print].
3. Zhao J, Benton S, Zhang B *et al.* Benign and intermediate-grade melanocytic tumors with BRAF mutations and spitzoid morphology: a subset of melanocytic neoplasms distinct From melanoma. *Am. J. Surg. Pathol.* 2022; **46**; 476–485.
4. Raghavan S, Peternel S, Mully T *et al.* Spitz melanoma is a distinct subset of spitzoid melanoma. *Mod. Pathol.* 2020; **33**; 1122–1134.
5. Yeh I, Busam KJ. Spitz melanocytic tumours - a review. *Histopathology* 2022; **80**; 122–134.

6. Barnhill RL, Argenyi ZB, From L *et al.* Atypical Spitz nevi/tumors: lack of consensus for diagnosis, discrimination from melanoma, and prediction of outcome. *Hum. Pathol.* 1999; **30**; 513–520.
7. Ruijter CGH, Ouwerkerk W, Jaspars EH *et al.* Incidence and outcome of Spitzoid tumour of unknown malignant potential (STUMP): an analysis of cases in The Netherlands from 1999 to 2014. *Br. J. Dermatol.* 2020; **183**; 1121–1123.
8. Lallas A, Kyrgidis A, Ferrara G *et al.* Atypical spitz tumours and sentinel lymph node biopsy: a systematic review. *Lancet Oncol.* 2014; **15**; e178–e183.
9. Cazzato G, Massaro A, Colagrande A *et al.* Dermatopathology of malignant melanoma in the era of artificial intelligence: a single institutional experience. *Diagnostics (Basel)* 2022; **12**; 1972.
10. Bhoyrul B, Brent G, Elliott F *et al.* Pathological review of primary cutaneous malignant melanoma by a specialist skin cancer multidisciplinary team improves patient care in the UK. *J. Clin. Pathol.* 2019; **72**; 482–486.
11. Massi D, De Giorgi V, Mandalà M. The complex management of atypical Spitz tumours. *Pathology* 2016; **48**; 132–141.
12. Berbís MA, McClintock DS, Bychkov A *et al.* Computational pathology in 2030: a Delphi study forecasting the role of AI in pathology within the next decade. *EBioMedicine* 2023; **88**; 104427.
13. Mosquera-Zamudio A, Launet L, Tabatabaei Z *et al.* Deep learning for skin melanocytic tumors in whole-slide images: a systematic review. *Cancers (Basel)* 2022; **15**; 42.
14. Spatz A, Calonje E, Handfield-Jones S, Barnhill RL. Spitz tumors in children: a grading system for risk stratification. *Arch. Dermatol.* 1999; **135**; 282–285.
15. Martinez Ciarpaglini C, Gonzalez J, Sanchez B *et al.* The amount of melanin influences p16 loss in Spitzoid melanocytic lesions: correlation with CDKN2A status by FISH and MLPA. *Appl. Immunohistochem. Mol. Morphol.* 2019; **27**; 423–429.
16. St»hle L. Analysis of variance (ANOVA). *Chemom. Intell. Lab. Syst.* 1989; **6**; 259–272. https://www.sciencedirect.com/science/article/pii/0169743989800954.
17. Li S, Chu Y, Wang Y *et al.* Distinguish the value of the benign nevus and melanomas using machine learning: a meta-analysis and systematic review. *Mediat. Inflamm.* 2022; **2022**; 1734327.
18. Burti S, Zotti A, Bonsembiante F, Contiero B, Banzato T. A machine learning-based approach for classification of focal splenic lesions based on their CT features. *Front. Vet. Sci.* 2022; **9**; 872618.
19. do Nascimento MZ, Martins AS, Azevedo Tosta TA, Neves LA. Lymphoma images analysis using morphological and non-morphological descriptors for classification. *Comput. Methods Prog. Biomed.* 2018; **163**; 65–77.
20. Shao S, Mao N, Liu W *et al.* Epithelial salivary gland tumors: utility of radiomics analysis based on diffusion-weighted imaging for differentiation of benign from malignant tumors. *J. Xray Sci. Technol.* 2020; **28**; 799–808.
21. Vijithananda SM, Jayatilake ML, Hewavithana B *et al.* Feature extraction from MRI ADC images for brain tumor classification using machine learning techniques. *Biomed. Eng. Online* 2022; **21**; 52.
22. Colloby PS, West KP, Fletcher A. Observer variation in the measurement of Breslow depth and Clark's level in thin cutaneous malignant melanoma. *J. Pathol.* 1991; **163**; 245–250.
23. Elmore JG, Barnhill RL, Elder DE *et al.* Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *BMJ* 2017; **357**; j2813.
24. Lodha S, Saggar S, Celebi JT, Silvers DN. Discordance in the histopathologic diagnosis of difficult melanocytic neoplasms in the clinical setting. *J. Cutan. Pathol.* 2008; **35**; 349–352.