ORIGINAL ARTICLE

Neuropathology and
Applied Neurobiology
JOURNAL OF THE BRITISH NEUROPATHOLOGICAL SOCIETY

WILEY

Analysing cerebrospinal fluid with explainable deep learning: From diagnostics to insights

Leonille Schweizer[1,2] | Philipp Seegerer[3,4] | Hee-yeong Kim[5] |
René Saitenmacher[3] | Amos Muench[1,2] | Liane Barnick[1] | Anja Osterloh[1] |
Carsten Dittmayer[1] | Ruben Jödicke[1,2] | Debora Pehl[6] | Annekathrin Reinhardt[7] |
Klemens Ruprecht[8] | Werner Stenzel[1] | Annika K. Wefers[9] |
Patrick N. Harter[10,11,12] | Ulrich Schüller[9,13,14] | Frank L. Heppner[1,2,15,16] |
Maximilian Alber[4,17] | Klaus-Robert Müller[3,18,19,20] | Frederick Klauschen[19,21,22]

Correspondence
Leonille Schweizer, Institute of Neurology (Edinger Institute), University Hospital Frankfurt, Goethe University, Frankfurt am Main, Germany.
Email: leonille.schweizer@kgu.de

Abstract

Aim: Analysis of cerebrospinal fluid (CSF) is essential for diagnostic workup of patients with neurological diseases and includes differential cell typing. The current gold standard is based on microscopic examination by specialised technicians and neuropathologists, which is time-consuming, labour-intensive and subjective.

Methods: We, therefore, developed an image analysis approach based on expert annotations of 123,181 digitised CSF objects from 78 patients corresponding to 15 clinically relevant categories and trained a multiclass convolutional neural network (CNN).

Results: The CNN classified the 15 categories with high accuracy (mean AUC 97.3%). By using explainable artificial intelligence (XAI), we demonstrate that the CNN identified meaningful cellular substructures in CSF cells recapitulating human pattern recognition. Based on the evaluation of 511 cells selected from 12 different CSF samples, we validated the CNN by comparing it with seven board-certified neuropathologists blinded for clinical information. Inter-rater agreement between the CNN and the ground truth was non-inferior (Krippendorff's alpha 0.79) compared with the agreement of seven human raters and the ground truth (mean Krippendorff's alpha 0.72, range 0.56–0.81). The CNN assigned the correct diagnostic label (inflammatory, haemorrhagic or neoplastic) in 10 out of 11 clinical samples, compared with 7–11 out of 11 by human raters.

Conclusions: Our approach provides the basis to overcome current limitations in automated cell classification for routine diagnostics and demonstrates how a visual explanation framework can connect machine decision-making with cell properties and thus provide a novel versatile and quantitative method for investigating CSF manifestations of various neurological diseases.

Leonille Schweizer and Philipp Seegerer contributed equally to this work.

For affiliations refer to page 14

INTRODUCTION

The analysis of cerebrospinal fluid (CSF) is a key procedure for the diagnosis of central nervous system diseases. An indispensable element of CSF analysis remains the microscopic analysis of cytocentrifuged CSF samples to obtain differential cell counts [1, 2]. Cell type determination and quantification are still crucial for the differential diagnosis of acute and chronic infections (e.g., the prevalence of granulocytes, activated lymphocytes and plasma cells) and subarachnoid haemorrhage invisible on CT and MRI imaging (e.g., detection of erythrophages, haemosiderophages and haematoidin crystals) as well as neoplastic meningitis [1, 3–6].

Gold-standard morphological profiling of CSF cells relies on visual inspection by highly specialised technicians and/or board-certified neuropathologists, and manual counting is only feasible for high-quality samples of intermediate or low cell density [1, 2]. While the results are highly dependent on experience and expertise and prone to observer bias and oversights, procedures are labour-intense and time-consuming with limitations for quality control and economic scalability. Furthermore, specialised personnel is usually not available 24/7 for time-critical decisions in emergency situations [2].

Absolute quantification of CSF cells and erythrocytes by automated cytometry is applied in many laboratories for a rough orientation of the cell density. However, current commercial cytometers developed for blood cell analysis are unable to reliably count low cellularity CSF samples (<30 cells/µl; normal: <5 cells/µl), to differentiate CSF cell types and to provide differential cell counts [2]. Deep learning algorithms have been developed for cell detection and differentiation in blood and bone marrow specimens. Recently, the detection of epithelial tumour cells has been demonstrated in CSF samples by a neural network, but no holistic approach addressing all relevant diagnostic questions in CSF diagnostics is currently available [7–9].

Although some algorithms for specific diagnostic tasks, like the identification of blast cells in acute myeloid leukaemia in bone marrow samples, have reached human-level performance [10], very few have been clinically implemented. Besides the lack of generalisability of the algorithm to external data sets, one essential issue is that training data may not be representative of samples encountered in daily clinical work, which demonstrate higher variability in quality and specimen preservation [11]. Moreover, pathology and medicine, in general, are characterised by "long tail" distributions of diagnoses, that is, only a few disease entities make up the majority of the cases and a plethora of differential diagnoses exist that are very rare. Most current alleged "clinical grade" AI approaches perform well for frequent diseases, but are unable to properly classify rare cases [12]. Furthermore, the clinical value of algorithms is limited by oversimplification and reducing the complexity of the pathologists' tasks by incorporating several layers of information into the decision-making and diagnostic labelling of samples [13].

> **Key Points**
> - We compiled a real-world dataset of 123,181 digitised CSF objects from 78 patients, annotated into 15 diagnostically relevant categories
> - Deep learning can accurately classify different cell types and shows high agreement with human experts
> - Interpretable visualisation allows explaining the machine predictions in an intuitive manner

We, therefore, set out to compile a real-world CSF dataset containing several sources of variation (e.g., different laboratory sample pre-processing, staining protocols, scanners, sample qualities, cell preservation states, object densities and common artefacts) and all diagnostically relevant cell types/objects to train a robust algorithm for cell type differentiation with the potential to solve complex diagnostic tasks. We developed a convolutional neural network (CNN), which reliably recognises 15 categories of diagnostically relevant CSF cell types and objects. To validate our model and assess its realistic usefulness in diagnostic practice, we validated the CNN-based approach by comparing its performance to that of seven board-certified neuropathologists from different academic institutions. By using explainable AI methods, we identified morphological features learned by the model to discriminate between specific cell types, which allowed us to further validate the model and explain its current limitations.

MATERIALS AND METHODS

CSF dataset collection, processing and annotation

We selected 128 CSF specimens from 78 patients, which were diagnostically evaluated at the Department of Neuropathology Charité – Universitätsmedizin Berlin between 2008 and 2020 (Table S1). Diagnoses included pleocytosis ($n = 8$), inflammation ($n = 6$), chronic haemorrhage ($n = 20$) and neoplastic meningitis ($n = 44$; Figure 1A). Slides were stained in two different laboratories according to standard procedures with May–Grünwald–Giemsa. The quality of each slide was assessed prior to digitisation, allowing for a mix of high, intermediate and low quality depending on the extent of autolytic and artificial changes (Table S1). Slides were scanned at a magnification of 40× using two different slide scanners: Hamamatsu NanoZoomer HAT 2.0 and 3D Histech P150.

Whole slide images were tiled in 2000 × 2000 pixel pictures (Figure 1B). Fifteen clinically relevant object categories were defined:

(A) Dataset

(B) Scanning and Tiling

(D) Training of the CNN

(C) Annotation
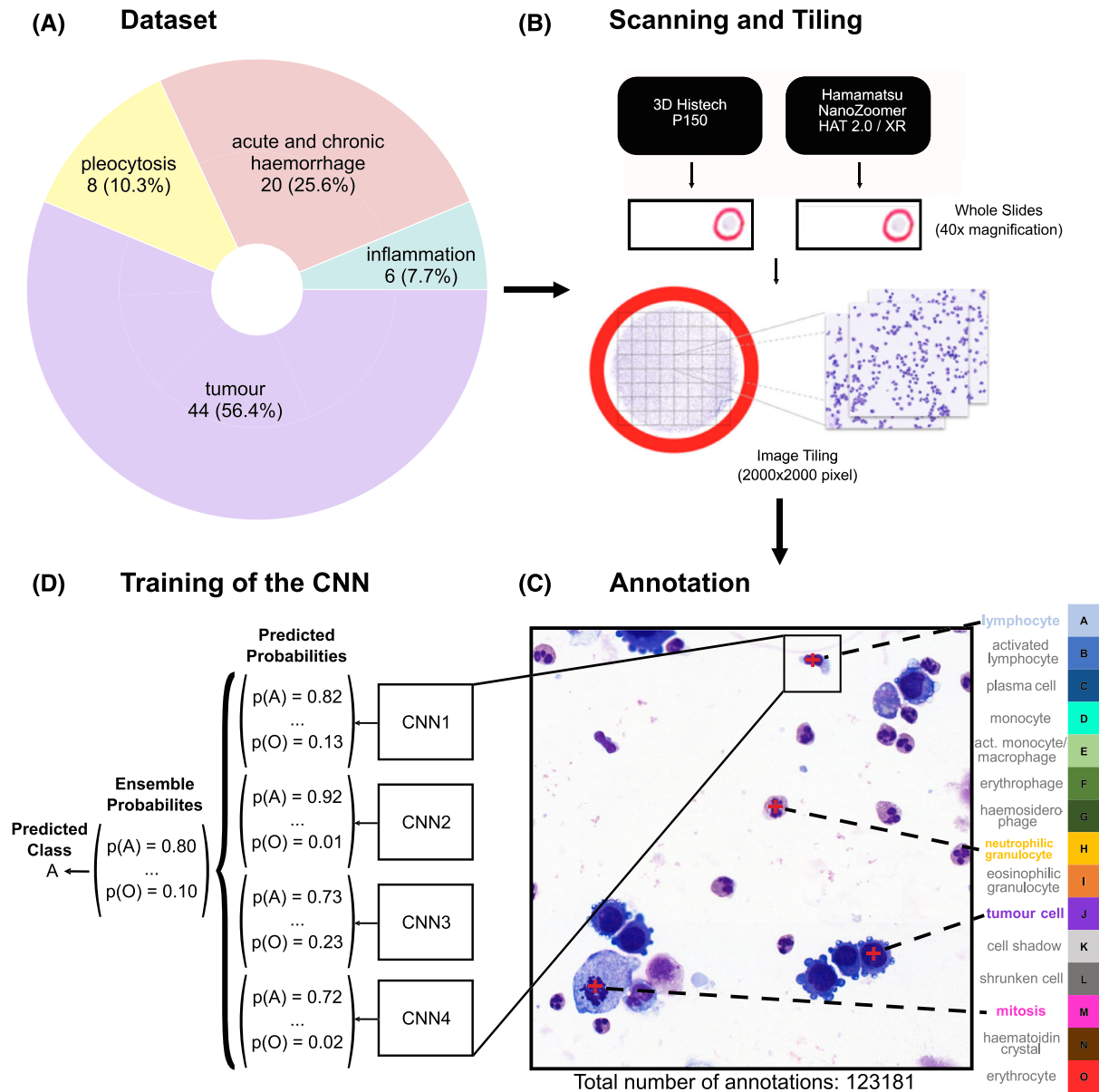
FIGURE 1 Data set and strategy. The pipeline of CSF cell detection. (A) The number of whole-slide images per diagnosis (i.e., tumour, pleocytosis, acute and chronic haemorrhage and inflammation) was influenced by the prevalence of rare cell types. (B) Image processing: scanning and tiling of whole-slide images. (C) Annotation of 15 object categories. (D) Patches are extracted from the image and used to train an ensemble of four convolutional neural networks (CNN) for multiclass prediction. The predicted class is based on the average of the four individual predictions.

lymphocyte (A), activated lymphocyte (B), plasma cell (C), monocyte (D), activated monocyte/macrophage (E; subcategories activated monocyte and macrophage), haemosiderophage (F), erythrophage (G), neutrophilic granulocyte (H), eosinophilic granulocyte (I), tumour cell (J; subcategories: carcinoma, melanoma, lymphoma, leukaemia), mitosis (M), haematoidin crystal (N), erythrocyte (O; Figure 2). A total of 123,181 objects were annotated using the object-based image analysis framework CognitionMaster (https://sourceforge.net/projects/cognitionmaster/, Figure 1C).

Data partitioning for model selection

The dataset was partitioned into three subsets: a training set for learning of the DNN weights, a validation set for model selection and a test set for evaluation of the generalisation performance on unseen data. The partitioning was performed on a case level; that is, samples of one case belong to one partition (either training, validation or test) only, so that generalisation performance across patients can reliably be estimated (Appendix).

FIGURE 2 Cerebrospinal fluid (CSF) object categories and diagnostic labels. Five examples are given per category to illustrate variance in scan quality, staining intensity and intraclass variability. (Categories A–C) Lymphoid lineage with common precursor cell. (Categories D–I) Myeloid lineage with common precursor cell. For diagnostic evaluation, only the categories lymphocyte (category A) and monocyte (category D) are considered normal cell types in CSF samples. Increased prevalence of categories B, C, H and I indicate an inflammatory process. Categories F, G, N and O are observed in acute and chronic haemorrhage. The presence of any cancer cells (category J) qualifies for the diagnosis of neoplastic meningiosis. Category M (mitosis) may be seen in neoplastic CSF as well as in inflammatory samples. Scale bar = 20 μm.

Pre-processing and data augmentation

Around each cell annotation, a 128 × 128 px image was cropped from the tile. For annotations close to the boundary, mirror padding was applied to yield crops of the same size. From these crops, the training images were extracted by cropping 112 × 112 px patches at random locations and applying random blur with a Gaussian kernel, random rotation, mirroring and random colour variation (brightness and contrast). Following Tellez et al. [14], a rather strong augmentation was used. For evaluation purposes, patches were centred around the annotation, and no random transformation was applied. The patches were standardised by subtracting the mean and dividing by the standard deviation of each colour channel in the training set.

Model architecture and training

As a backbone for our model, we used a VGG16 model pre-trained on ImageNet and fine-tuned it on our data [15]. The main training objective was a cross-entropy loss function, where the contribution of each class was weighted by the inverse of its relative frequency in the training set. For early stopping based on the validation set performance, macro-averaged sensitivity (i.e., balanced accuracy) was chosen to account for the imbalanced class distribution. Complementary to the cross-entropy, a consistency term was added to the loss function, similar to Bortsova et al. [16]. Each patch was predicted by the network in four different versions: original, horizontally flipped, vertically flipped and both horizontally and vertically flipped. The consistency term is then the average mean squared error of each of the four predictions to their mean. This incentivises the model to be invariant to rotation, in addition to random data augmentation. For evaluation, the predictions were always averaged across the four transformed versions. Tuning of the training hyperparameters was exclusively done using the validation set. The test set was only used to evaluate the generalisation performance of the final models. We used PyTorch (version 1.10.0) for all our experiments. The models were trained on machines with Tesla P100 and Quadro RTX 6000 GPUs. One training run took an average of 11 h, which equals a carbon footprint of 2.5 kg CO_2 assuming average emission values for Germany [17].

ANALYSING CSF WITH EXPLAINABLE DEEP LEARNING

Neuropathology and
Applied Neurobiology
JOURNAL OF THE BRITISH NEUROPATHOLOGICAL SOCIETY

WILEY

5 of 16

Model evaluation, visualisation and explainability

Four individual models (CNN1–4, Figure 1D) were trained, each using one partitioning of the data. To evaluate the performance of the models on unseen data, a joint confusion table of the predictions of all four models on their respective test sets was computed. The confusion table was then normalised such that each row sums to 1. This yields the sensitivity for each class on the diagonal. The average of these class-wise sensitivities is defined as mean sensitivity (macro-average sensitivity). For the combined categories activated monocyte/macrophage (E) and tumour cell (J), we additionally report the sensitivity on the subgroups respectively (e.g., activated monocyte and macrophage for E; carcinoma, melanoma, leukaemia and lymphoma for J) and report the precision and the F1 score for each class (see Table S2). Note that these metrics are not defined for the subgroups of E and J since the subclasses are not predicted by the model.

In order to verify that the models have learned meaningful representations, we projected the 4096-dimensional feature vectors (output of the penultimate layer of the CNNs) of the test samples into 2D by Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [18]. To complement the evaluation, explanation heatmaps were computed using Layer-wise Relevance Propagation (LRP) applying Zennit [19–21]. The explanation heatmaps were computed on each of the four models (CNN1–4) individually (for details, see Appendix).

Validation set (CSF500) for performance comparison to neuropathologists

To compare the performance of the CNN head-to-head to neuropathologists, we created a completely independent evaluation dataset (i.e., CSF500) consisting of 12 different patients with non-specific, inflammatory, haemorrhagic and neoplastic diagnoses (for original diagnosis, see Table S2). As for the training set samples, the CSF500 samples were scanned with Hamamatsu NanoZoomer HAT 2.0 ($n = 4$) and 3D Histech P150 scanners ($n = 8$). We selected 511 cells out of 12 digitised whole slide images (Supplementary CSF500.pptx). All 511 cells were evaluated by seven board-certified neuropathologists and assigned to one of the 15 annotation categories (A-O; P: not applicable). Raters were blinded for clinical information. A specialised and certified CSF technician provided with clinical information served as ground truth (GT) annotation. For CNN predictions of the CSF500 objects, an ensemble of the four models CNN1–4 (Figure 1D) was used by averaging the predicted probability vectors (ranging between 0 and 1) for each cell. Individual results of each rater, the GT, the ensemble CNN and the four individual CNNs for all 511 objects are summarised in Table S3.

Raters were also asked to assign a diagnostic label to each CSF sample (i.e., inflammatory, haemorrhagic, neoplastic or a combination of two labels). For diagnostic labelling of the CNN, samples were determined as haemorrhagic (>8 h) in case an erythrophage, haemosiderophage or haematoidin crystal was detected and as inflammatory upon the presence of either plasma cells or granulocytes. In case both haemorrhagic and inflammatory cells were present, the most abundant cell types determined the diagnosis, in case of equality, "haemorrhagic" was chosen, because granulocytes may be derived from peripheral blood. The presence of a single tumour cell resulted in the additional label "neoplastic." In cases where tumour cells were the most abundant cell types, the sample was diagnosed as "neoplastic" only.

Performance evaluation and statistics

Statistical analysis was conducted using R (version 3.6.3). For performance evaluation, precision (true positives/(true positives + false positives)) and sensitivity (true positives/(true positives + false negatives)) were calculated. Overall, inter-observer reliability was measured using Krippendorff's alpha coefficient as it is recommended in case of missing observations (label P: not applicable) [22]. Calculation of Fleiss' Kappa coefficients additionally allowed the comparison of inter-rater reliability for individual cell type categories. Both coefficients were calculated applying the R package irr (v0.84.1) and yielded highly similar results (mean difference 3.05×10^{-5}). Coefficient matrices were visualised using corrplot v0.92 [23]. Class-wise receiver operating characteristics (ROC) curves were computed in a one-vs-rest fashion. The average AUC was calculated using the one-vs-one approach by Hand et al. [24]. ROC analysis was performed using scikit-learn (version 1.0.2; http://scikit-learn.sourceforge.net).

Data and code availability

The dataset of the study is available at Zenodo (DOI 10.5281/zenodo.6543147, https://zenodo.org/record/6543147). The code is available at https://github.com/pseegerer/csf_cell_classification.

RESULTS

Multiclass prediction

After training the neural network, we first used the test set to estimate the performance on unseen data. We repeated the training four times on different partitions of the dataset (four-fold cross-validation, that is, CNN1–4, see Figure 1D). Confusion matrices of the four CNNs were summarised and are shown in Figure 3A (for details, see Section 2). The trained neural networks demonstrate accurate prediction performances for most cell types with an average area under the ROC curve (AUC) of 0.973 (range for individual categories 0.90–1.00, Figure S2). The average sensitivity for classes A–O was 76% (range 45%–96%; sensitivity, precision, and F1 scores for individual categories are given in Table S4).

FIGURE 3 Confusion matrix: Influence of cell density, CSF quality and number of annotations on predictions. (A) Confusion matrix for the multiobject classifier identifying 15 CSF categories. Percentages represent (row-wise normalised) sensitivities of the ensemble network evaluated on the test set. Cells of the same derivation (lymphoid vs myeloid lineage) and belonging to a morphological continuum were more often mixed up. (B) A higher sensitivity was achieved in samples with low cell density in which objects were not overlapping. CSF samples with inferior quality and artificial changes due to cell deterioration demonstrated lower sensitivity. (C) The number of training samples per class correlates with sensitivity. For category J (tumour cell), the sensitivity of tumour subtypes is highlighted in red. The combined category activated monocyte/macrophage (category E) highlighted in orange represents a large spectrum from activated monocytes (small to intermediate cells, single intracytoplasmatic vacuoles) to macrophages (highly variable size, few to numerous intracytoplasmatic vacuoles) with low sensitivity.

Very good performance was noted for physiological cell types (sensitivity for lymphocytes 87% and monocytes 82%) as well as key cell classes relevant for diagnostic labelling (e.g., neutrophilic granulocytes 96%, tumour cells 79% and haemosiderophages 78%). The confusion matrix shows that misclassification mainly occurred in related morphological classes of the lymphoid or myeloid lineage representing a challenge also to human evaluators. For example, a substantial overlap was noted between plasma cells and activated lymphocytes belonging to a morphological continuum of B-cells with signs of cellular activation. When applying a more tolerant classification scheme allowing for confusion for these two categories, sensitivity for activated lymphocytes increased to 90% and for plasma cells to 96% (Table S2).

FIGURE 4 UMAP visualisation of the dataset. (A) The last hidden layer representation of the test samples (4096D) of CNN1 is visualised in 2D by UMAP (uniform manifold approximation and projection). (B) Tumour classes: Only the tumour samples are shown and coloured by their tumour type. Overlapping subclusters are formed by carcinoma and melanoma as well as lymphoma and leukaemia. Circled inlet: Several samples were misclassified as cell shadows due to autolytic changes. (C) Examples of the original input images with the colour of the cross representing the prediction and the colour of the frame corresponding to the annotation label.

Overall sensitivity was significantly better in samples of low cellularity (e.g., distinctly localised cells without overlapping or abutting cell borders) and high or intermediate quality (Figure 3B). We also saw a high correlation of sensitivity to the number of training samples available for each category (Figure 3C).

Morphological class features learned by neural networks form distinct clusters

To analyse if the models learned meaningful representations, the high-dimensional feature vector of CNN1 was visualised in 2D by applying UMAP, which shows that morphologically defined cell types were projected in discrete clusters in the 2D representation (Figure 4A, see also Figure S3).

While cell shadows and mitosis formed discrete clusters apart from other object categories, we observed overlap in variable degrees between the remaining classes. The transition between clusters not only corresponded to the observed misclassification of related categories as mentioned above (Figure 3A) but also reflected biologically meaningful difficulties encountered in daily clinical work: the overlap of the categories activated lymphocyte and plasma cell potentially reflects on the difficulties in discriminating these transitioning and consecutively developing cell types derived from a common lymphoid precursor cell. Furthermore, it also draws attention to a possible underlying annotation problem based on the difficulty of defining clear cut-offs for the correct designation of cells in continuous transition. A similar overlap was noted for cells of myeloid derivation. Finally, the proximity of activated lymphocytes to B-cell lymphoma cells in the 2D representations hints towards the well-known diagnostic dilemma of distinguishing between neoplastic and inflammatory lymphocytes. Interestingly, normal lymphocytes (usually T-cells in CSF) did not overlap with activated lymphocytes and plasma cells (B-cell lineage) which may be due to the extremely scarce cytoplasm in (naïve) T cells in comparison to activated lymphocytes.

Within the large tumour cluster, tumour cells of different origins were locally aggregated (Figure 4B): melanoma and epithelial tumour cells showed substantial overlap but were more disjunct compared with haematopoietic tumour cells (e.g., lymphoma and leukaemia cells), which holds great potential for tumour subtype prediction in future. Of note, several tumour cells and other cells fell into the cell shadow group (Figure 4C). Upon re-evaluation, these cells demonstrated highly autolytic features which potentially masked the neoplastic or cell-type specific character of the deteriorating and highly artificially altered cells, and would probably have resulted in the classification as autolytic cells by human raters without clinical information (for further examples of misclassified and projected cells, see Supplementary Sprite Figure).

Cell type recognition by the CNN is based on comprehensible and visualisable cytological features

To further elucidate the most relevant image regions for the CNN predictions, we used layer-wise relevance propagation (LRP) to identify image regions (down to pixel resolution) and features with high informative value for classification. Figure 5 shows the LRP heatmaps for selected samples of the CSF500 set for the four different CNNs. Here, evidence in favour of the predicted class is highlighted in red whereas counter-evidence is highlighted in blue; neutral areas are highlighted in grey. In correctly classified examples, the LRP heatmaps highlighted pathognomonic cytological substructures, such as the segmented nucleus and perinuclear halo. Interestingly, for some categories, discrimination of cell borders contains highly relevant information for the correct classification (e.g., 156, 88 and 171), while for other cell types, intracytoplasmic structures (e.g., hemosiderin deposits in 158 and cytoplasmic vacuoles in 293) are more relevant.

Misclassified cells represent common difficulties with segmentation (e.g., define borders in a complex environment with abutting or overlapping objects in 379, 404 and 293), feature similarity (e.g., segmented vs multilobulated nucleus in 379), rare and blurred objects with misfocus on isolated sharp structures ignoring cell context and borders (404) and methodological/technical issues like small image patches which prevented recognition of the class defining morphological feature (e.g., intracytoplasmic erythrocyte in 293). Note, in example 404 the extracellular structure which led to the misclassification as "Haemosiderophage" in CNN1, 2 and 4 was identified by CNN3 as counter-evidence for the correct prediction "Mitosis".

CNN performance compared with neuropathologists

We evaluated the performance of the ensemble CNN (average of CNN1–4) on an independent dataset of 511 cells and compared it with seven raters blinded for clinical information. Overall inter-observer agreement compared with ground truth (GT) varied among neuropathologists (average Krippendorff's alpha = 0.72, range 0.56–0.81; Figure 6A). One of the raters represented a negative outlier (rater 4 vs GT, Krippendorff's alpha = 0.56), which can be mainly attributed to the omission to diagnose tumour cells in any of the samples (Figures 6A and 7A).

We found that the highest overall inter-rater agreement (0.79, measured by Krippendorff's alpha) as well as individual cell categories (mean 0.79, range 0.44–0.98, measured by Fleiss' kappa) of human raters was observed for the GT annotation. We also noticed a tendency to a higher agreement among raters who belonged to the same diagnostic centre (also same centre as the GT: mean 0.74, range 0.71–0.79) compared with raters of other diagnostic institutions (mean: 0.63, range 0.55–0.73).

Overall, neuropathologists, the GT and the ensemble CNN showed very high agreement for categories neutrophilic granulocyte (H), cell shadow (K), mitosis (M), haematoidin crystal (N) and erythrocyte (O; Fleiss' kappa ≥ 0.93) and high agreement on categories lymphocyte (A) and eosinophilic granulocyte (I; Fleiss' kappa ≥ 0.78; Figure 6B). Lower inter-rater agreement was mainly observed for those cell types belonging to a morphological continuum (e.g., to the lymphoid or myeloid lineage). Strikingly, there was a very low
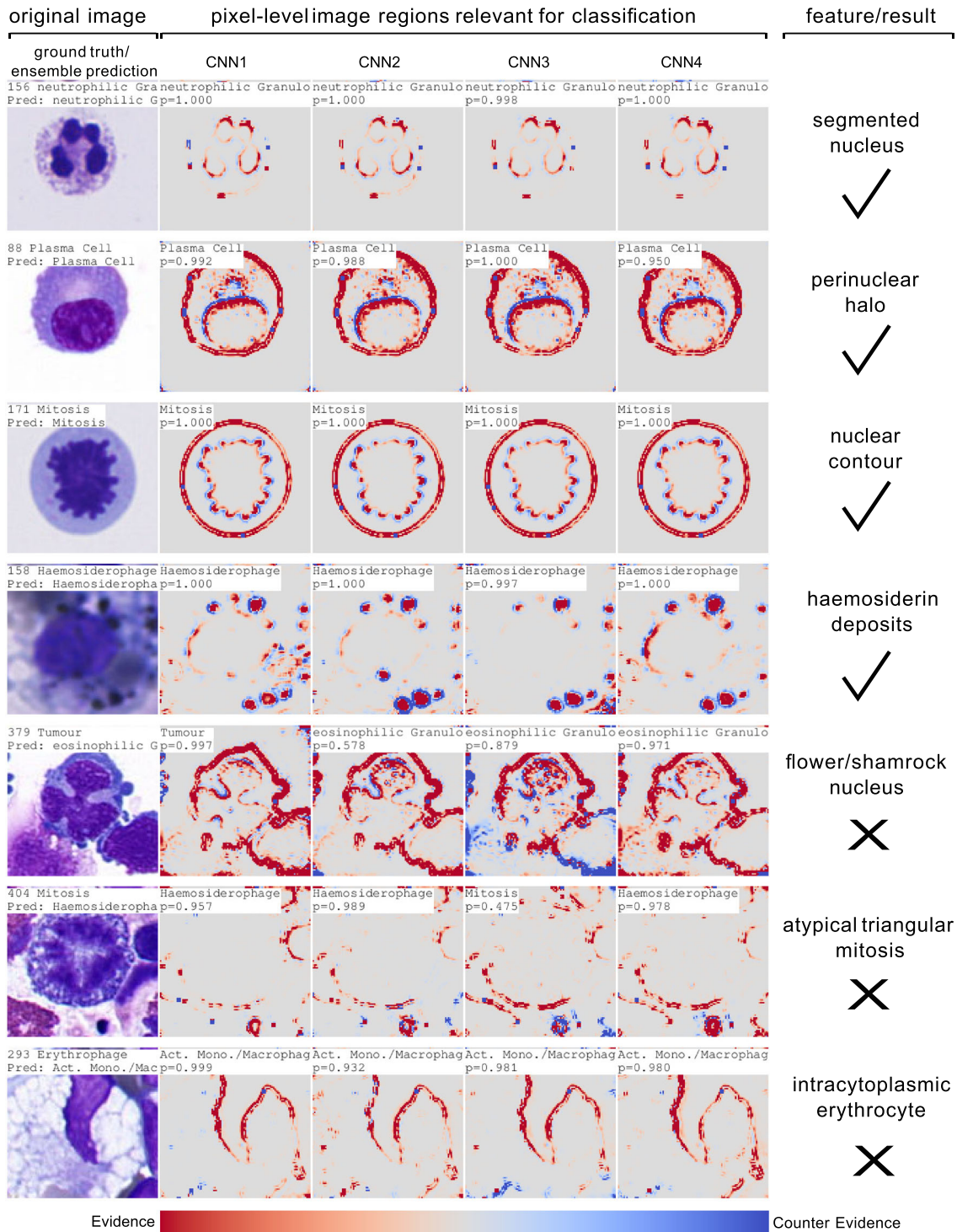
FIGURE 5 Pixel-level image regions relevant for classification. Shown is the original image (with object ID, ground truth label and ensemble prediction) as well as a feature heatmap calculated for each of the four individual CNNs based on layer-wise relevance propagation (LRP). Pixels highlighted in red are evidence in favour of the predicted class whereas blue pixels are evidence against it (e.g., counter-evidence). The degree of confidence is given as probability (p, range 0–1) with high numbers representing high confidence. Examples of correctly (✓) and incorrectly (✗) classified objects are selected and the respective cytological feature relevant to human recognition is specified. Misclassified cells represent common difficulties with segmentation (e.g., define borders in a complex environment with abutting or overlapping objects; 379, 404 and 293), feature similarity (379), rare and blurred objects with misfocus on isolated sharp structures ignoring context (404) and too small image patches (293).
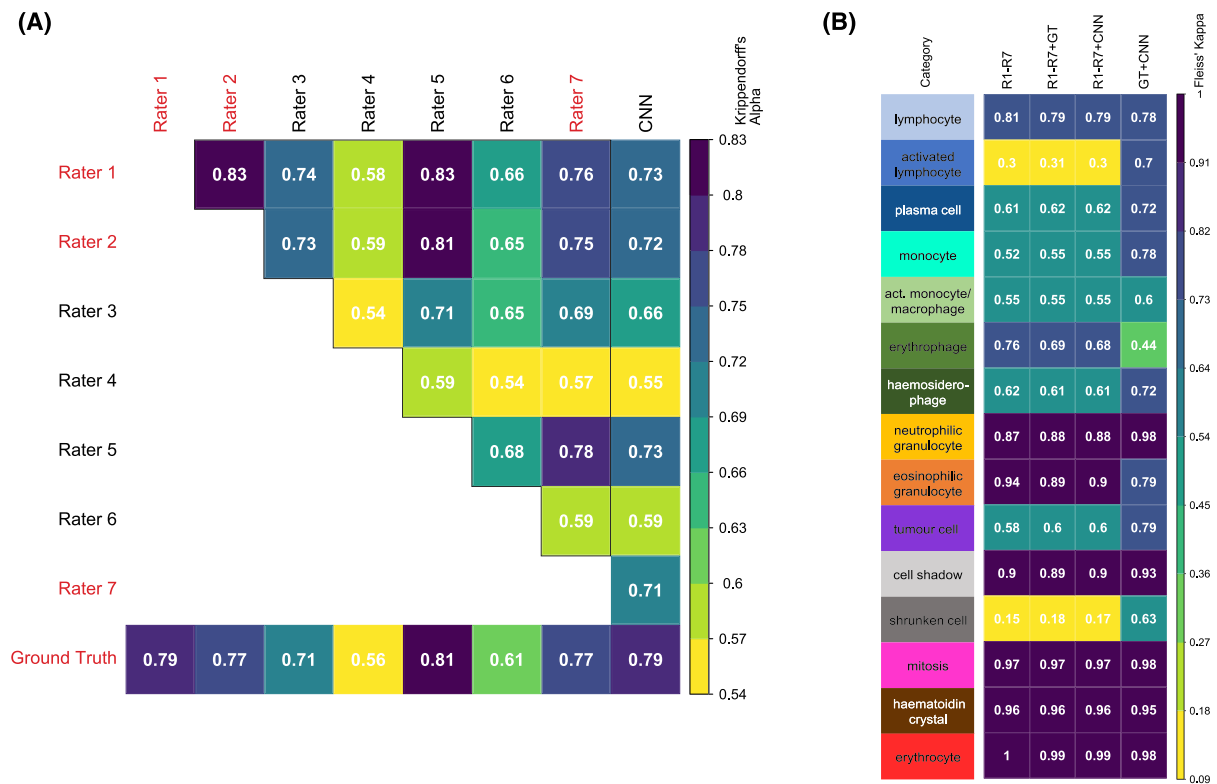
FIGURE 6 Substantial inter-rater agreement between the CNN, seven neuropathologists and the ground truth annotation. (A) Based on the evaluation of 511 cells in 12 different CSF samples, inter-rater reliability between the CNN and the ground truth was non-inferior (Krippendorff's alpha 0.79) compared with the agreement of seven human raters and the ground truth (mean Krippendorff's alpha 0.72, range 0.56–0.81). Raters highlighted in red letters in Figure 6A belong to the same diagnostic centre. (B) Fleiss' kappa values for individual object categories (for details of categories A–O, see Figure 2) were higher for discrete categories (e.g., categories H, K, M, N and O) compared with cells of a common lineage derivation belonging to a morphological continuum (e.g., categories B + C as well as D–G). A substantially lower agreement was achieved for activated lymphocyte (category B) and artificial cell (category L) when incorporating human classifications, which may be influenced by centre-specific category definitions. Abbreviations: GT, ground truth; R, rater

agreement between raters 1–7 and the GT as well as the CNN for the categories activated lymphocyte (B) and shrunken cell (L), which may be attributable to institutionally defined and non-standardised morphological cut-offs and consensus in continuous cellular categories, especially relevant for shrunken cells (i.e., delineation of artificially condensed cells from naked-nucleic lymphocytes).

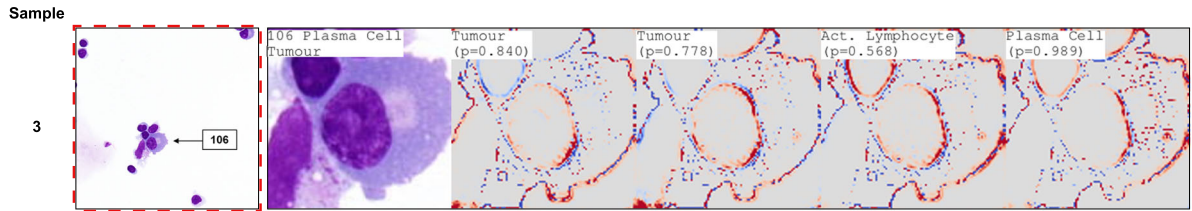## Most CSF samples were correctly diagnosed by the ensemble CNN

CSF samples were correctly labelled based on the CNN predictions in 10/11 cases (Figures 7A and S4). The CNN showed a high capability to identify haemorrhagic samples (3/3; 100%) as well as neoplastic samples (4/4; 100%) with 68–100% of the tumour cells in these samples being correctly identified. The ensemble CNN correctly identified acute inflammatory conditions dominated by neutrophilic and eosinophilic granulocytes (2/2; 100%). The differentiation of inflammatory vs neoplastic lymphocytes

represents a known diagnostic pitfall and dilemma, especially in the absence of clinical information. In CSF sample 3 (diagnosis: neuro-borreliosis), the ensemble CNN wrongly classified a single plasma cell as a tumour cell. The misclassified plasma cell, demonstrating artificial cytoplasmic blebs, a feature often observed in epithelial tumour cells, was classified with low probability scores by three of the four CNNs (Figure 7B). Of note, rater 7 also incorrectly classified three cells as tumour in this sample. In sample 9 (diagnosis: viral meningoencephalitis), three of the seven raters incorrectly classified cells as tumour (10, 14 and 51 cells, respectively), while the CNN only predicted six activated lymphocytes as tumour cells (Figure 7C). Misclassification of CSF sample 1 (diagnosis: low to moderate pleocytosis with iatrogenic blood contamination) by some raters and the CNN as inflammatory was most likely caused by very limited cytological contextual information as well as the design of the CSF500 dataset with an arbitrary selection of four eosinophilic granulocytes attributable to slight haemorrhagic contamination in an otherwise non-specific sample (Supplementary CSF500).

(A)

| Diagnosis | inflammatory | | neoplastic | | haemorrhagic (> 8 h) | | | |

Sample	R1	R2	R3	R4	R5	R6	R7	CNN	Ground Truth
1*									
2	30/30 (100%)	30/30 (100%)	31/30 (100%)	0/30 (0%)	30/30 (100%)	30/30 (100%)	30/30 (100 %)	26/30 (87%)	30/30 (100 %)
3							3/0	1/0	
4	25/25 (100%)	25/25 (100%)	1/25 (4%)	0/25 (0%)	25/25 (100%)		25/25 (100%)	17/25 (68%)	25/25 (100%)
5									
6									
7	21/21 (100%)	22/21 (100%)	21/21 (100%)	0/21 (0%)	21/21 (100%)	21/21 (100%)	20/21 (95%)	21/21 (100%)	21/21 (100%)
8									
9	10/0				14/0	51/0		6/0	
10	25/25 (100%)	25/25 (100%)	25/25 (100%)	0/25 (0%)	25/25 (100%)	25/25 (100%)	0/31 (0%)	17/25 (68%)	25/25 (100%)
11									
12									
Correct Diagnosis	10/11	11/11	11/11	7/11	10/11	8/11	9.5/11	10/11	11/11

(B)

Sample

3 — 106 Plasma Cell Tumour | Tumour (p=0.840) | Tumour (p=0.778) | Act. Lymphocyte (p=0.568) | Plasma Cell (p=0.989)

(C)

Sample

9

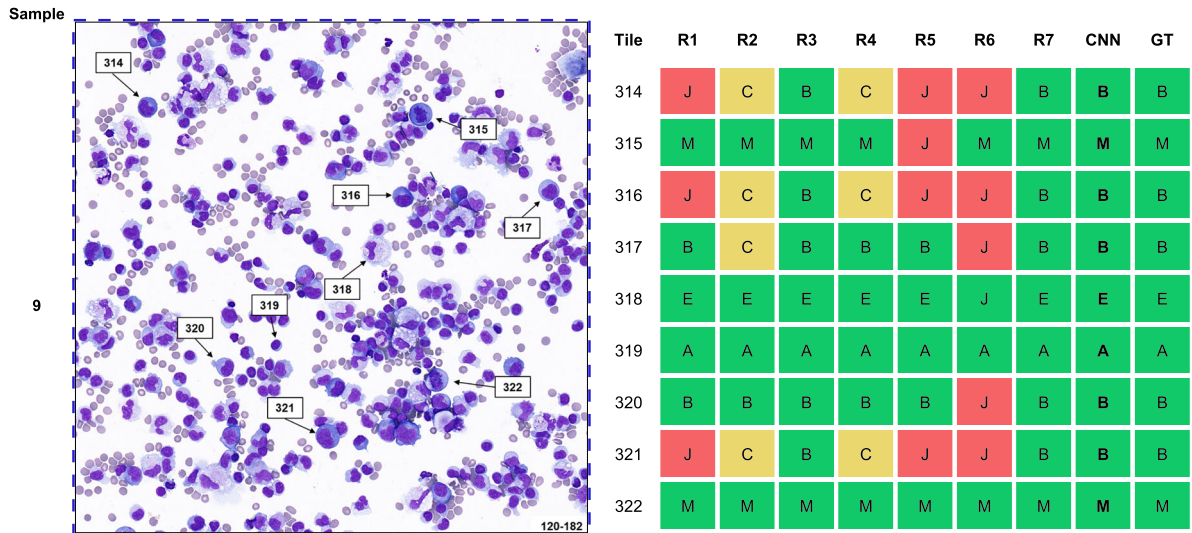| Tile	R1	R2	R3	R4	R5	R6	R7	CNN	GT
314	J	C	B	C	J	J	B	**B**	B
315	M	M	M	M	J	M	M	**M**	M
316	J	C	B	C	J	J	B	**B**	B
317	B	C	B	B	B	J	B	**B**	B
318	E	E	E	E	E	J	E	**E**	E
319	A	A	A	A	A	A	A	**A**	A
320	B	B	B	B	B	J	B	**B**	B
321	J	C	B	C	J	J	B	**B**	B
322	M	M	M	M	M	M	M	**M**	M

FIGURE 7 Legend on next page.

FIGURE 7 Performance of the ensemble CNN on the CSF500 dataset in comparison to neuropathologists. (A) Comparison of diagnostic labels for the 12 CSFs assigned by the raters and the ensemble CNN. Sample 1* was not rated because of the selection of four eosinophilic granulocytes misleading to inflammatory impression in a slightly sanguineous sediment with unspecific changes. For neoplastic CSFs, the number of tumour cells detected compared with the ground truth is provided. Of note, clinical information was not provided to the raters before the evaluation. The ensemble CNN proved capable of correctly classifying diagnostically relevant cell types and assigning the correct diagnostic label in 10/11 cases. Misclassification of two inflammatory samples was based on the identification of one and six tumour cells in cases 3 and 9. (B) The single plasma cell misclassified as a tumour cell by the ensemble CNN is shown as well as the individual prediction and explanation heatmaps of all four CNNs, which highlight a plasma cell with unusual cytoplasmic blebs. (C) An image example of case 9 is depicted with individual votes of human raters, the ensemble CNN and the ground truth (GT). The CNN misclassified fewer cells as tumour cells compared with the human raters. Correctly identified cells are highlighted in green in the table, yellow indicates cells that belong to a morphological continuum and may be considered as correctly labelled, and red corresponds to incorrectly classified cells. Abbreviations: A, lymphocyte; B, activated lymphocyte; C, plasma cell; E, activated monocyte/macrophage; J, tumour cell; M, mitosis

DISCUSSION

Deep learning approaches have shown promising results in the analysis and classification of histopathological images [11, 25]. However, only a few of these have been translated into clinical applications and are currently used in diagnostic settings [11, 26, 27]. In this study, we aimed to develop an AI-based automated cytological/CSF image analysis approach for which we compiled a large dataset with 123,181 annotations of digitised CSF samples from 78 patients and trained a deep neural network capable of multiclass recognition of diagnostically relevant CSF cells and diagnostic labelling of CSF specimens. Our aim was to create a routine diagnostic-grade AI tool and real-world dataset that can serve as a reference to design and improve machine learning solutions for CSF as well as other cytological diagnostics but also to consider and address obstacles in the dataset design and partitioning strategy that may hamper clinical translation.

Several machine learning approaches have been developed for the analysis of peripheral blood smears and bone marrow samples and were capable of differentiating different categories of cells, for example, leukocyte subtypes, as well as normal and neoplastic cells [7, 28]. Furthermore, deep learning models even proved feasible in discriminating between different haematopoietic cancer subtypes [29]. Recently, the proficiency in identifying epithelial cancer cells of a CNN in CSF samples was shown by Yu et al. [9]. The authors demonstrated that the accuracy of the CNN was similar to experts and the CNN was superior in predicting the carcinoma origin compared with humans while reducing the working time by 90%. However, none of these works addressed the full spectrum of diagnostically relevant CSF objects and provided an in-depth explanation of the model behaviour, as presented in this study.

One of the reasons for limited utility in clinical settings is that models are developed on high-quality samples that differ from those encountered in daily diagnostic routines and that do not consider the numerous sources of variance that influence morphological representation in images: laboratory-specific sample handling, staining protocols, scanner setting and especially degree of cell deterioration over time. Instead of compiling a dataset consisting of optimal quality samples which may achieve high-performance metrics, but result in low generalisability of the model, we aimed to design a real-world dataset reflecting the realistic and full spectrum of samples encountered in daily work, including samples of different quality and object density (see Figure 3). We paid special attention to representing the full variability of the data including rare cell types. Hence, we over-represented tumour samples in our training dataset since these feature a high degree of variability, e.g., due to different tumour origins and degrees of differentiation, such that a large number of samples is required to train a deep neural network. For this reason, the prevalence of cell types in the training set does not reflect a real-world situation in daily diagnostics, e.g., tumour cells are usually recognised in only 5–10% of CSF samples [30–32], but were over-represented in our dataset with 56% of CSF samples (see Figure 1). Even though this led to a slight over-sensitivity to tumour cells (see Figure 7), we argue that in a clinical setting, false positive tumour predictions are less severe than false negative ones.

In clinical settings with a wide prevalence distribution of common and very rare conditions (i.e., long-tail distribution) [11], datasets are usually highly imbalanced, which represents a potential further limitation to external validity. In imbalanced datasets, the allocation of a limited number of clinical samples with multiple categories into training, test and validation set is complicated as the assignment of one patient to more than one set may result in overfitting. Compared with previous work [7], we allocated samples of individual patients exclusively to one set (e.g., training, validation or testing set), which is crucial for an accurate performance estimation because samples of one case cannot be considered statistically independent and partitioning per sample rather than per case will optimistically bias performance metrics [33]. However, due to this restriction, it is complicated to divide the dataset in such a way that classes are balanced. We address this issue by proposing a novel data partitioning strategy (see Appendix and Figure S1).

Because of the naturally limited number of rare objects in clinical datasets, there is a need to develop not only suitable partitioning strategies but also intuitive ways to leverage the data-derived representations of learning systems to infer general concepts [11]. Identifying the underlying substructures in a comprehensive image context that resulted in the correct classification of an object by using explanation heatmaps may help to approach new solutions to study algorithmic understanding and transferable conceptualisation. To achieve this, the field of explainable AI has recently advanced with methods such as GradCAM, SmoothGrad and Layer-wise Relevance

Propagation [34, 35]. These methods have already been applied for deep learning in pathology, for instance, tumour classification [36], cytology [7] and morphological biomarker discovery [37].

Using explanation heatmaps allowed us to visualise cellular substructures in correctly and incorrectly classified cells and align them to features relevant to human object recognition in cytological specimens. Furthermore, we identified specific features that provided counter-evidence for a certain prediction (e.g., extracellular deposit reminiscent of hemosiderin in a mitosis example, see Figure 5, 404). The visualisations make the otherwise latent classification rules transparent: They seem to be not only based on the recognition of features in favour of a certain category, but also on information learnt to be highly specific for other categories, resembling in some respects human decision-making of weighing and integrating different sources and layers of information to approximate a conclusion. The identification of comprehensible morphological features as part of the basis of algorithmic predictions rather than elusive criteria of a black box model may further increase the acceptance of such applications in the (neuro)pathological community and overcome resentments due to lack of transparency [11, 38].

To gain further insight into the meaningfulness of the learned features by the CNN and the definition of our predefined annotation categories, we additionally embedded the penultimate layer of the VGG16 containing 4096 features into two dimensions via UMAP. As has been demonstrated by Matek et al [7] before, morphologically defined classes were separated well in the feature space and similar cell types were projected to neighbouring positions. The unsupervised dimensionality reduction and visualisation as image sprite (see Appendix) allowed fast and systematic identification of misannotated images which—in combination with the heatmaps—may serve as an easy screening tool for dataset refinement, with the intention to reduce label noise, which is mainly caused by human errors occurring during highly repetitive and tedious tasks.

We observed a continuous transition in feature space between cells derived from the same developmental lineage (e.g., myeloid cells). On the contrary, normal lymphocytes and activated lymphocytes were represented discretely in the feature map, the latter overlapping with plasma cells and lymphoma cells. Although a biological reason cannot be demonstrated based on our data, it is interesting to note that normal lymphocytes in CSF usually are T-cells and activated lymphocytes forming a morphological continuum with plasma cells belonging to the B-cell lineage. The intrinsic capacity to differentiate distinct lymphoid cell populations may eventually help to further refine the landscape of inflammatory CSF infiltrates in various neurological diseases. The overlap of activated B-cells with lymphoma cells of the B-cell lineage is pointing towards similar or shared features for these cells learnt by the algorithm, which may include enlarged nuclear size and chromatin heterogeneity, skewed nuclear-to-cytoplasmic ratio and increased cytoplasmic basophilia [5, 39].

Among neuropathologists, it is well-known that lymphocytic activation in inflammatory conditions, such as viral meningoencephalitis or neuroborreliosis, may be so pronounced that it mimics malignant lymphoma [1]. Especially in cases where clinical information is not provided, it represents an ongoing dilemma to distinguish between neoplastic and inflammatory lymphocytes [5, 39], which is exemplified in case 9 of the CSF500 set (e.g., viral meningoencephalitis) labelled as neoplastic by three of seven neuropathologists and the CNN. Compared with the CNN, neuropathologists were able to consider context information of the 2000×2000 pixel images containing several other cell types, information on preservation status, overall staining intensity and cellularity of a sample, which is not available to the CNN that classifies images only based on small patches containing a single item. Implementing dynamic patch sizes similar to Hashimoto et al [40], i.e., varying instead of fixed input size and resolution, in the design of the CNN may potentially increase classification accuracy not only for large cell types, which are currently insufficiently enclosed in the fixed patches (see Figure 5, 293) but also by incorporating context information available to human raters.

To obtain differential cell counts, determination of the exact cell type is necessary. However, neuropathological assessment of CSF samples usually does not involve the classification of every single cell in a sample, but relies on summarising the cellular picture in a descriptive diagnostic category often based on the predominance of a cell type, allowing for recognising a mix of cells: some with high and others with low confidence. An important exception is the identification of a single unequivocal tumour cell which results in the diagnosis of neoplastic meningitis (label: neoplastic) [30, 31, 41]. For instance, the misclassification of one plasma cell in sample 3 (Figure 7B) was based on the summary of the classifications of the four individual CNNs. The ensemble predicted "tumour cell" with a reduced probability of $p = 0.41$ followed by the correct category "plasma cell" with a slightly lower probability of $p = 0.38$. However, two of the CNNs provided only intermediate confidence scores ($p = 0.84$ and $p = 0.78$) for the class tumour cell and highlighted cytoplasmic blebs in the explanation heatmap, CNN3 identified an activated lymphocyte with low confidence ($p = 0.57$) and CNN4 the correct label plasma cell with high confidence ($p = 0.99$), highlighting and recognising the decisive perinuclear region in addition to blebs. The result indicates a differing capability of CNNs, trained on a variable selection of the training dataset, to identify specific classes more or less correctly, which suggests that the variance of the data is very high and the model would benefit from even larger and more diverse datasets. Furthermore, providing confidence values together with the class predictions to diagnosticians will help overcome algorithmic aversion and increase the safety of using CNNs as screening tools in CSF diagnostics. A valuable extension of the current model will be the implementation of a calibration model with class-wise thresholds and labelling of unclassifiable objects, which could be directly delegated to human evaluation, similar to what has been introduced with the concept of calibrated classifier scores in the DKFZ brain tumour classifier [26].

In this work, we validated the capability of our model to classify CSF cells based on an independent dataset of 511 cells. Although the performance was already good enough to derive broad diagnostic labels for 11 clinical cases, a thorough and detailed prospective clinical validation period cannot be replaced and needs to be conducted in parallel with routine diagnostics in the future. Our

algorithm provides the basis to develop an automated, transparent and validated diagnostic assist system, which may be available 24/7 also in emergency situations, e.g., at night or the weekend when neuropathological diagnostic service is usually not provided. Furthermore, the estimated time to whole slide image classification in a fully automated workflow is expected to last only a few minutes (e.g., scan time: 1–3 min, whole slide image tiling: 10 seconds, CNN-based cell classification: 2 ms per cell), which will mainly depend on the cellularity of the sample and reduce the turnaround time from lumbar puncture to diagnosis as has been shown by Yu et al. recently [9]. The integration of our system with further improvements in handling images from portable microscopes or even smartphones may translate cytological CSF diagnostics from laboratories to clinical bedside applications.

Besides automation and improvement of diagnostic routine processes, machine learning algorithms have already demonstrated proficiency in refining tumour classification and determining cancer origin based on methylation profiles [26, 27]. Analysing epigenetic patterns in large cohorts of various brain cancers even proved capable of identifying new cancer subtypes and new tumour entities that went unrecognised in small datasets and previous histology-based tumour diagnostics [42, 43]. Similarly, extending the reference dataset of the CSF classifier and applying unsupervised machine learning approaches could result in the identification of new, clinically and prognostically meaningful cell types or delineate novel distinct cell states along a common lineage and differentiation trajectory in neoplastic and inflammatory CSF samples, which could have been missed by human microscopic analysis so far.

ACKNOWLEDGEMENTS

CONFLICT OF INTEREST

PS and MA are currently Aignostics employees. MA, KRM and FK are co-founders of Aignostics. The remaining authors declare no competing interests.

ETHICS STATEMENT

Ethical approval (EA1/060/22) was granted by the Charité Ethics Committee.

AUTHOR CONTRIBUTIONS

LS, PS and FK contributed to the conceptualisation of the study. LS, HK, LB and RJ compiled and annotated the reference dataset. KR contributed to the design of the dataset and the definition of clinically meaningful categories. Acquisition and digitisation were supported by CD. Interpretation of the CSF500 test data was provided by AO, DP, AK, AW, WS, US and PH. LS, PS, HK, RS and AM applied machine learning, statistical analysis and visualisation of the data. LS and PS wrote the initial manuscript. KRM and FK supervised the manuscript writing and provided important intellectual content. All authors were involved in the critical revision and final approval of the manuscript. LS, KRM and FK provided funding.

AFFILIATIONS

[1]Department of Neuropathology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[2]German Cancer Consortium (DKTK), Partner Site Berlin, German Cancer Research Center (DKFZ), Heidelberg, Germany

[3]Machine-Learning Group, Department of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, Berlin, Germany

[4]Aignostics GmbH, Berlin, Germany

[5]Systems Medicine of Infectious Disease, Robert Koch Institute, Berlin, Germany

[6]Department of Pathology, Vivantes Hospitals Berlin, Berlin, Germany

[7]Department of Neuropathology, University Hospital Heidelberg, Heidelberg, Germany

[8]Department of Neurology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[9]Institute of NeuropathologyUniversity Medical Center Hamburg-Eppendorf, Hamburg, Germany

[10]Neurological Institute (Edinger Institute), Goethe University, Frankfurt am Main, Germany

[11]Frankfurt Cancer Institute, Goethe University, Frankfurt am Main, Germany

[12]German Cancer Consortium (DKTK), Partner Site Frankfurt/Mainz, German Cancer Research Center (DKFZ), Heidelberg, Germany

[13]Department of Pediatric Hematology and Oncology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[14]Research Institute Children's Cancer Center Hamburg, Hamburg, Germany

[15]Cluster of Excellence, NeuroCure, Berlin, Germany

[16]German Center for Neurodegenerative Diseases (DZNE) Berlin, Berlin, Germany

[17]Institute of Pathology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[18]Max Planck Institut für Informatik, Saarbrücken, Germany

[19]Berlin Institute for the Foundations of Learning and Data (BIFOLD), Berlin, Germany

[20]Department of Artificial Intelligence, Korea University, Seoul, South Korea

[21]German Cancer Consortium (DKTK), Partner Site Munich, German Cancer Research Center (DKFZ), Heidelberg, Germany

[22]Institute of Pathology, Ludwig-Maximilians-Universität München, Munich, Germany

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/nan.12866.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Zenodo at https://zenodo.org/record/6543147.

ORCID

Leonille Schweizer https://orcid.org/0000-0002-4649-2587
Philipp Seegerer https://orcid.org/0000-0002-4707-7991
Werner Stenzel https://orcid.org/0000-0002-1143-2103
Annika K. Wefers https://orcid.org/0000-0001-9394-8519
Ulrich Schüller https://orcid.org/0000-0002-8731-1121

REFERENCES

1. Rahimi J, Woehrer A. Overview of cerebrospinal fluid cytology. *Handb Clin Neurol*. 2018;145:563-571. doi:10.1016/B978-0-12-802395-2.00035-3
2. Wick M, Gross CC, Tumani H, Wildemann B, Stangel M, on behalf of the German Society of CSF Diagnostics and Clinical Neurochemistry, DGLN e.V. Automated analysis of cerebrospinal fluid cells using commercially available blood cell analysis devices—a critical appraisal. *Cell*. 2021;10(5):1232. doi:10.3390/cells10051232
3. Müller-Jensen L, Diamandis E, Osterloh A, Ruprecht K, Leithner C. CSF cytology in subacute subarachnoid haemorrhage. *Neurology*. Published online. 2020.
4. Nagy K, Skagervik I, Tumani H, et al. Cerebrospinal fluid analyses for the diagnosis of subarachnoid haemorrhage and experience from a Swedish study. What method is preferable when diagnosing a subarachnoid haemorrhage? *Clin Chem Lab Med (CCLM)*. 2013;51(11):2073-2086. doi:10.1515/cclm-2012-0783
5. Perske C, Nagel I, Nagel H, Strik H. CSF cytology—the ongoing dilemma to distinguish neoplastic and inflammatory lymphocytes. *Diagn Cytopathol*. 2011;39(8):621-626. doi:10.1002/dc.21510
6. Prayson RA, Fischler DF. Cerebrospinal fluid cytology: an 11-year experience with 5951 specimens. *Arch Pathol Lab Med*. 1998;122(1):47-51.
7. Matek C, Krappe S, Münzenmayer C, Haferlach T, Marr C. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood, J Am Soc Hematol*. 2021;138(20):1917-1927. doi:10.1182/blood.2020010568
8. Wang Q, Bi S, Sun M, Wang Y, Wang D, Yang S. Deep learning approach to peripheral leukocyte recognition. *PLoS ONE*. 2019;14(6):e0218808. doi:10.1371/journal.pone.0218808
9. Yu W, Liu Y, Zhao Y, et al. Deep learning-based classification of cancer cell in leptomeningeal metastasis on cytomorphologic features of cerebrospinal fluid. *Front Oncol*. 2022;12
10. Matek C, Schwarz S, Spiekermann K, Marr C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nat Mach Intell*. 2019;1(11):538-544. doi:10.1038/s42256-019-0101-9
11. Stenzinger A, Alber M, Allgäuer M, et al. Artificial intelligence and pathology: from principles to practice and future applications in histomorphology and molecular profiling. In: *Seminars in Cancer Biology*. Elsevier; 2021.
12. Hoberger M, von Laffert M, Heim D, Klauschen F. Histomorphological and molecular profiling: friends not foes! Morpho-molecular analysis reveals agreement between histological and molecular profiling. *Histopathology*. 2019;75(5):694-703. doi:10.1111/his.13930
13. Van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med*. 2021;27(5):775-784. doi:10.1038/s41591-021-01343-4
14. Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med Image Anal*. 2019;58:101544. doi:10.1016/j.media.2019.101544
15. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:14091556*. Published online 2014.
16. Bortsova G, Dubost F, Hogeweg L, Katramados I, de Bruijne M. Semi-supervised medical image segmentation via learning consistency under transformations. In: Shen D, Liu T, Peters TM, et al., eds. *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019-22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part VI*. Vol 11769. Lecture notes in computer science. Springer; 2019:810-818. doi:10.1007/978-3-030-32226-7_90
17. Lannelongue L, Grealey J, Inouye M. Green algorithms: quantifying the carbon footprint of computation. *Adv Sci*. 2021;8(12):2100707. doi:10.1002/advs.202100707
18. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv: 180203426*. Published online 2018.
19. Anders CJ, Neumann D, Samek W, Müller KR, Lapuschkin S. Software for dataset-wide XAI: from local explanations to global insights with Zennit, CoRelAy, and ViRelAy. *arXiv preprint arXiv:210613200*. Published online 2021.
20. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*. 2015;10(7):e0130140. doi:10.1371/journal.pone.0130140
21. Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. *Dig Sig Process*. 2018;73:1-15. doi:10.1016/j.dsp.2017.10.011
22. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? *BMC Med Res Methodol*. 2016;16(1):1-10.
23. Wei T, Simko V. R package 'Corrplot': visualization of a correlation matrix; 2021. https://github.com/taiyun/corrplot
24. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach Learn*. 2001;45(2):171-186. doi:10.1023/A:1010920819831
25. Binder A, Bockmayr M, Hägele M, et al. Morphological and molecular breast cancer profiling through explainable machine learning. *Nat Mach Intell*. 2021;3(4):355-366. doi:10.1038/s42256-021-00303-4
26. Capper D, Jones DT, Sill M, et al. DNA methylation-based classification of central nervous system tumours. *Nature*. 2018;555(7697):469-474. doi:10.1038/nature26000
27. Jurmeister P, Bockmayr M, Seegerer P, et al. Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci Transl Med*. 2019;11(509). doi:10.1126/scitranslmed.aaw8513
28. Kimura K, Tabe Y, Ai T, et al. A novel automated image analysis system using deep convolutional neural networks can assist to differentiate MDS and AA. *Sci Rep*. 2019;9(1):1-9. doi:10.1038/s41598-019-49942-z
29. Shafique S, Tehsin S. Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional

neural networks. *Technol Cancer Res Treat*. 2018;17. doi:10.1177/1533033818802789

30. Chamberlain MC. Neoplastic meningitis. *Curr Neurol Neurosci Rep*. 2008;8(3):249-258. doi:10.1007/s11910-008-0038-6

31. Gleissner B, Chamberlain MC. Neoplastic meningitis. *Lancet Neurol*. 2006;5(5):443-452. doi:10.1016/S1474-4422(06)70443-4

32. Le Rhun E, Devos P, Weller J, et al. Prognostic validation and clinical implications of the EANO ESMO classification of leptomeningeal metastasis from solid tumors. *Neuro Oncol*. 2021;23(7):1100-1112. doi:10.1093/neuonc/noaa298

33. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010;11:2079-2107.

34. Alber M, Lapuschkin S, Seegerer P, et al. iNNvestigate neural networks! *J Mach Learn Res*. 2019;20(93):1-8.

35. Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. Explaining deep neural networks and beyond: a review of methods and applications. *Proc IEEE*. 2021;109(3):247-278. doi:10.1109/JPROC.2021.3060483

36. Hägele M, Seegerer P, Lapuschkin S, et al. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Sci Rep*. 2020;10(1):1-12. doi:10.1038/s41598-020-62724-2

37. Seegerer P, Binder A, Saitenmacher R, et al. Interpretable deep neural network to predict estrogen receptor status from haematoxylin-eosin images. In: *Artificial Intelligence and Machine Learning for Digital Pathology*. Springer; 2020:16-37. doi:10.1007/978-3-030-50402-1_2

38. Border SP, Sarder P. From what to why, the growing need for a focus shift toward explainability of AI in digital pathology. *Front Physiol*. 2022;12:821217. doi:10.3389/fphys.2021.821217

39. Xing J, Radkay L, Monaco SE, Roth CG, Pantanowitz L. Cerebrospinal fluid cytology of Lyme neuroborreliosis: a report of 3 cases with literature review. *Acta Cytol*. 2015;59(4):339-344. doi:10.1159/000439160

40. Hashimoto N, Fukushima D, Koga R, et al. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2020:3852-3861.

41. Djukic M, Trimmel R, Nagel I, et al. Cerebrospinal fluid abnormalities in meningeosis neoplastica: a retrospective 12-year analysis. *Fluids Barriers CNS*. 2017;14(1):1-7. doi:10.1186/s12987-017-0057-2

42. Pajtler KW, Witt H, Sill M, et al. Molecular classification of ependymal tumors across all CNS compartments, histopathological grades, and age groups. *Cancer Cell*. 2015;27(5):728-743. doi:10.1016/j.ccell.2015.04.002

43. Sturm D, Orr BA, Toprak UH, et al. New brain tumor entities emerge from molecular classification of CNS-PNETs. *Cell*. 2016;164(5):1060-1072. doi:10.1016/j.cell.2016.01.015

44. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2017:618-626.

45. Hui LY, Binder A. Batchnorm decomposition for deep neural network interpretation. In: *International Work-Conference on Artificial Neural Networks*. Springer; 2019:280-291.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

APPENDIX

Data partitioning for model selection

We aimed at distributing classes evenly across the partitions. This requirement is complicated to fulfil by the fact that the dataset is partitioned on case level in a multi-class setting but the number of cases is relatively low and the classes are not distributed evenly among the cases. Therefore, we used the following partitioning scheme as an approximation: We quantified the amount of imbalance by the mean difference of the Jensen-Shannon divergence (JSD) of the class distribution of a partition to the mean class distribution over all folds. Furthermore, a partitioning was considered invalid if not all partitions contained samples of all classes or if the samples of at least one class came from only one case. The four partitionings with the lowest JSD after a search of over 5000 random partitionings (discarding invalid ones) were selected (Figure S1). Note that this partitioning strategy does not necessarily yield disjoint train, validation and test sets across partitionings (e.g., a case could be used in the train set of both partitions 1 and 2), in contrast to standard stratified cross-validation without grouping (e.g., used in [7]).

Explanation heatmaps

For the computation of LRP heatmaps, we followed the recommendations by Montavon et al. [21]: The ϵ-rule is used for dense layers and the ($\alpha = 2$, $\beta = -1$)-rule for convolutional layers. The input to the LRP-backpropagation is a vector that is 1 for the predicted class and 0 elsewhere, thereby focusing the heatmap on the predicted class. For visualisation, the resulting heatmaps are clipped at the 95th percentile of positive relevance values, such that the colour mapping is robust against outlier values.

Similar to Matek et al. [7], we experimented with GradCAM heatmaps but found that the resolution of these is too coarse to capture the fine details that are important to understand the models [44]. For instance, for the 112×112 input images used in this work, the GradCAM heatmap would be only 7×7 pixels wide ($16 \times$ coarser resolution than the input). In comparison, our LRP heatmaps have the same resolution as the input image and allow us to examine the predictions at the pixel level. Moreover, pixel-level heatmaps of the common ResNet architecture—as the SmoothGrad heatmaps used by Matek et al. [7]—have a peculiar grid pattern; that is, the heat is concentrated on pixels that lie on an equally spaced grid. This issue has been discussed previously in Hui et al. [45] and is not an artefact of heat mapping, but can be directly linked to the architecture. Thus, the grid artefact is visible both in LRP and SmoothGrad heatmaps [7]. Since this artefact impedes the interpretation of the heatmaps, we opted to use the more conservative VGG16 architecture that does not show this artefact.