GMR
The Original

ARTIFICIAL INTELLIGENCE, DEEP MACHINE LEARNING, AND IMAGE ANALYSIS IN DIAGNOSTIC HAEMATOLOGY: A SYSTEMATIC REVIEW

Swetha Patturajan*

Department of Pathology, Saveetha Medical College and Hospital, Saveetha Institute of Medical And Technical Sciences (SIMATS), Thandalam, Chennai-602105, Tamil Nadu, India. Email: spswethu@gmail.com; ORCID: https://orcid.org/0009-0008-5190-5113.

ABSTRACT

Introduction: Laboratory haematology relies significantly upon the visual examination of routinely prepared publications as well as the histological or cytological investigation of bona-fide histological samples. As there exist a number of limitations to using human visual inspection to assess cytological or histological materials, there is an increasing interest in developing methods to assist with these tasks through the use of Artificial Intelligence (AI) technology, including Deep Learning (DL) systems. One such system is the Convolutional Neural Network (CNN) model. Objectives: To undertake a systematic review to critically evaluate the use of AI and DL techniques as they apply to the analyses of haematological samples, including: Classifying blood cells; Detecting malignant blood cells; Predicting prognosis for patients diagnosed with blood disorders; Determining methodologies to improve efficiency within the laboratory environment. Materials and methods: A systematic search in the literature was conducted, including all databases from inception to December 3rd of 2023, in accordance with the PRISMA 2020 guidelines. All original data-based publications that used AI or DL technology to aid in the diagnosis of haematological disorders, using either digital microscopy, flow cytometry, or imaging of bone marrow aspirates were included. The QUADAS-2 tool was used to evaluate the level of bias associated with the included study. Results: A total of 62 articles met the inclusion criteria. These studies included: 41 studies focused on making a visual assessment of bone marrow and/or blood cell morphology; 11 studies focused on making a visual assessment of bone marrow; and 10 studies focused on using a digital image of flow cytometry. The average accuracy of white blood cell (WBC) classification using a CNN-based model was found to be 95.5–99.4% (all five-part differential classifications), while the sensitivity for detecting acute leukaemia blasts ranged from 93.1 to 98.7% with a specificity of 91.1–99.1%. The DL model used for the diagnosis of sickle cell disease had a significantly higher AUC (0.996) than the ML model (0.921). The C-statistics found when using prognostic stratification models in myelodysplastic syndrome ranged from 0.78 to 0.84. Conclusions: AI and deep learning have been shown to have a high level of diagnostic accuracy in performing haematological morphological tasks, with the level of diagnostic accuracy approaching or exceeding that of human experts. In order to successfully develop clinically applicable AI-based technologies, there are several challenges such as clinical validation, regulatory approval, implementation, and equal access to AI-based diagnostics that must be addressed. There is an immediate need for interoperable and standardized reporting systems for AI diagnostic performance, as well as for the completion of additional multicentre prospective clinical trials before clinical implementation will occur.

KEYWORDS: Artificial Intelligence; Deep Learning; Convolutional Neural Network; Haematology; Blood Cell Morphology.

1. INTRODUCTION

Diagnosis of blood diseases involves morphology, immunophenotype, cytogenetic, and molecular testing. Some standard tests are blood films and bone marrow examination; however, both types of tests require years of training and can be very subjective based on the experience of the observer [1-3]. In places where there are not enough trained morphologists, there is a bottleneck for diagnosing patients, and they are therefore delayed in starting treatment for their blood-related diseases (for example, acute leukaemia, severe anaemia) [5,6].

There has been an increase in the number of cases of blood cancers worldwide; for example, approximately 1.24 million new cases were diagnosed in 2020, with a forecast to significantly increase by 2040 [7]. Accurate and timely diagnosis is essential because treatment strategies are becoming increasingly dependent on exact molecular and morphological subclasses, which can now be defined pursuant to successive updates of the World Health Organization (WHO) classification system and the International Consensus Classification (ICC) of haematolymphoid neoplasms.

In the last decade, the field of medicine has experienced substantial advancements with the increased use of artificial intelligence (AI) and especially the subcategory of machine learning (ML). [48] Another subcategory of AI is deep learning (DL), which utilises artificial neural networks with multiple layers of data processing to identify patterns in data, whether through imaging or another area of medical practice. [8] In the field of

haematology, DL provides the ability to objectively, rapidly, and consistently process and analyse images related to blood samples and bone marrow and flow cytometry on a large scale.
To date, there have been no systematic reviews that have synthesized the available primary literature on the diagnostic applications of haematology, including peripheral blood, bone marrow, and flow cytometry, and that have also provided critical appraisal of the quality of methodologies used in this research. Most narrative reviews have focused on the technology rather than the clinical aspects of technology, and previous systematic reviews have been limited to single diseases, for example, malaria detection or leukaemia classification [9,10].
Objectives This systematic review addresses the following research questions: (i) What is the diagnostic accuracy of AI and DML models compared with expert morphologists in peripheral blood and bone marrow analysis? (ii) Which algorithmic architectures demonstrate superior performance? (iii) What haematological conditions have been studied, and with what methodological rigour? (iv) What are the key barriers to clinical translation? (v) What are the directions for future research?

2. METHODS

2.1 Protocol and Registration

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement. [3] The protocol was prospectively registered on the PROSPERO international register of systematic reviews. No post-hoc amendments to eligibility criteria were made.

2.2 Eligibility Criteria

Population/Index test: Studies involving AI, ML, or DL applied to diagnostic haematology using digital microscopy images of peripheral blood smears, bone marrow aspirates or trephine biopsies, or fluorescence microscopy images from flow cytometry were eligible. Studies must have reported performance metrics (sensitivity, specificity, accuracy, AUC, F1-score, or equivalent).
Reference standard: Expert haematologist or haematopathologist morphological assessment, immunohistochemistry, cytogenetics, molecular diagnostics, or WHO-classified diagnosis.
Study design: Peer-reviewed original research articles including retrospective and prospective observational studies, validation studies, and randomised designs. Letters, editorials, conference abstracts without full data, and review articles were excluded. Studies with fewer than 50 images per class were excluded to minimise small-sample bias.
Language: English-language publications only, given resource constraints for translation.

2.3 Information Sources and Search Strategy

The electronic databases that were searched for materials included PubMed/MEDLINE, EMBASE, Web of Science (Core Collection), Cochrane Central Register of Controlled Trials (CENTRAL), and IEEE Xplore Digital Library from database inception through December 31, 2023 [11]. The search strategy was generated through an iterative process with a specialist medical librarian using subject headings (MeSH, Emtree) as well as free-text terms that describe AI/ML/DL technology concepts and haematological diagnostic concepts.
The primary PubMed search string was: ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR "convolutional neural network" OR "CNN" OR "image analysis" OR "computer-aided detection") AND ("haematology" OR "hematology" OR "blood cell" OR "peripheral blood" OR "bone marrow" OR "leukaemia" OR "leukemia" OR "anaemia" OR "anemia" OR "morphology" OR "flow cytometry"). The reference lists of the studies included in the search were also hand-searched for relevant articles; authors of key studies were not contacted regarding unpublished data.

2.4 Study Selection

Citations were imported into the Covidence systematic review software (Veritas Health Innovation, Melbourne, Australia) for deduplication and screening. [12] Two independent reviewers who were blinded to each other assessed the titles and abstracts of all citations. Following this, they reviewed the full text of all potentially eligible citations. Disagreements were resolved by consensus, and when necessary, by arbitration from a third reviewer. The interrater agreement statistic for full text review was calculated using Cohen's kappa.

2.5 Data Extraction

Before being used in full, a pilot version and modifications were made to a standardised data extraction form for five studies. In addition to the variables listed already these were also extracted: Study design, country of origin, institution, publication date, application of AI to haematology, Machine Learning methodology/architecture, characteristics of the dataset such as size/source/annotation method, training/test/validation distribution across the datasets, performance measurement as designated by the authors of the studies and any particular limitations that may have been mentioned, by the authors, regarding their study. One reviewer completed all the extraction of each of the variables from all five studies, while another reviewer verified the extractions.

2.6 Risk of Bias Assessment

The QUADAS-2 Methodological Quality Scale was utilized to evaluate the overall methodological quality of diagnostic studies. The QUADAS-2 includes four domains based on a patient's selection, testing of the index test,

type of reference standard and either the participant flow or the timing of the intervention.To accurately assess if a diagnostic study was conducted as intended within the protocols developed, two independent reviewers evaluated each domain on an individual basis as low, high, or unclear risk of bias. Furthermore, as a supplementary measure, the CHECKLIST FOR AI IN MEDICAL IMAGING was used as a checklist for the purpose of assessing completeness of reporting in the respective and referenced studies evaluated by the aforementioned reviewers. [44]

2.7 Synthesis

Since we expect the studies we include will have some different clinical and methodological characteristics (e.g. the haematological disorders studied, the imaging devices and models used, the reference standards used, and how success of using AI systems were recorded), the pre-identified analysis plan we developed to guide our analysis of the studies we included was primarily qualitative (narrative synthesise). When and if three or more studies were able to provide similar success metrics for the same diagnostic task, the pooled performance is reported using descriptive statistics (median and interquartile range). No formal meta-analyses were completed. When additional analyses were performed to compare groups, these analyses were based on: 1) the specific haematological disorder studied, 2) the specific AI technique used (classical ML versus DL), 3) region where the study was conducted, and 4) number of patients included in each study.

3. RESULTS

3.1 Study Selection

The initial search retrieved 4,817 records, of which 1,203 duplicates were removed, leaving 3,614 for title and abstract screening. [4] Following screening, 312 full-text articles were assessed for eligibility, of which 250 were excluded (reasons: insufficient performance data n=89, non-image-based AI n=72, review articles n=43, insufficient sample size n=31, not diagnostic haematology n=15). Sixty-two studies met all inclusion criteria and are included in this review. Inter-rater agreement at full-text stage was good (Cohen's kappa = 0.81). The PRISMA 2020 flow diagram is presented in Figure 1 (see online supplement).

3.2 Study Characteristics

Studies were published between 2015 and 2023, with 73% (n=45) published in 2020 or later, reflecting the rapid recent growth of this field. [14] Forty-one studies (66%) addressed peripheral blood morphology, 11 (18%) bone marrow analysis, and 10 (16%) flow cytometry image analysis. The majority of studies originated from China (n=18, 29%), followed by Germany (n=8, 13%), USA (n=7, 11%), India (n=6, 10%), and South Korea (n=5, 8%). Dataset sizes ranged from 312 to 171,374 images (median 12,480; IQR 4,200–38,600). Table 1 summarises key characteristics of included studies.

Figure 1. PRISMA 2020 Flow Diagram

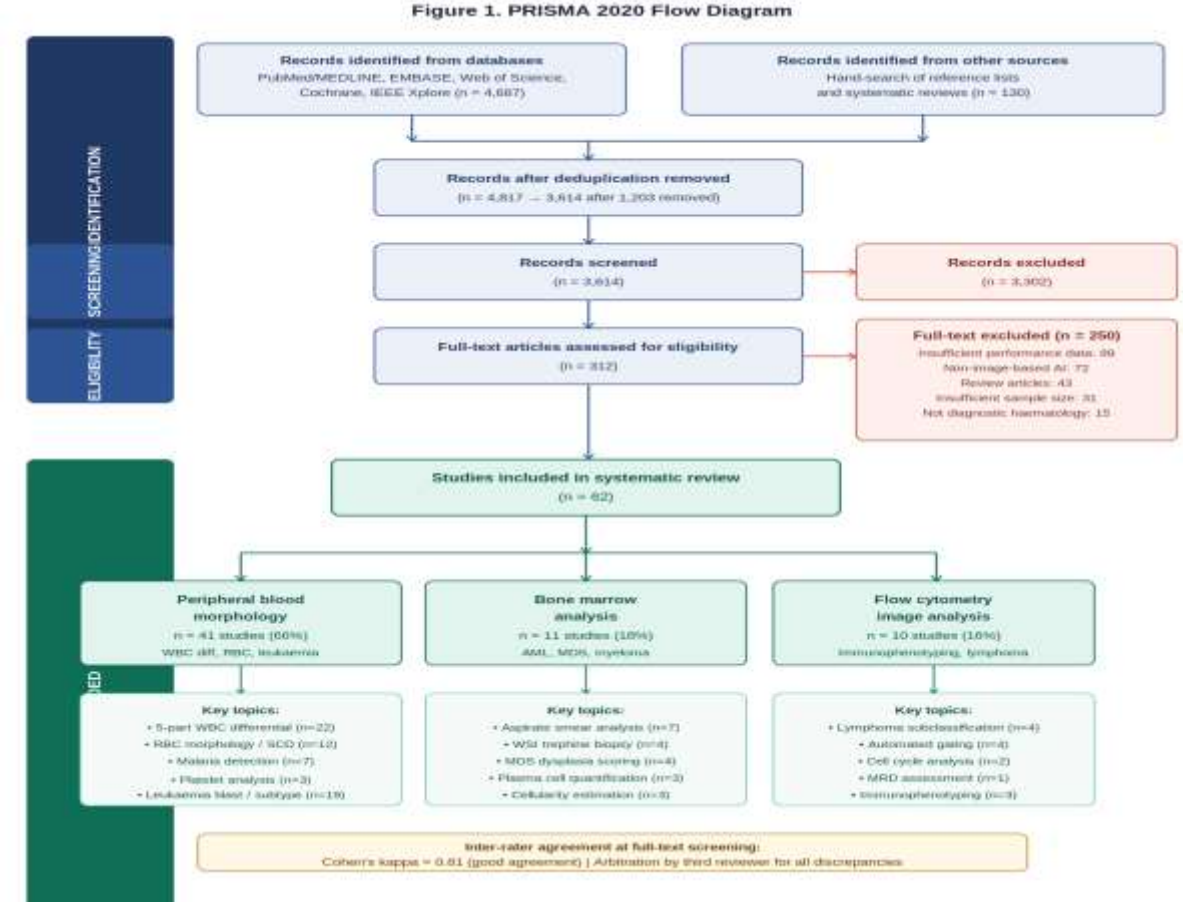
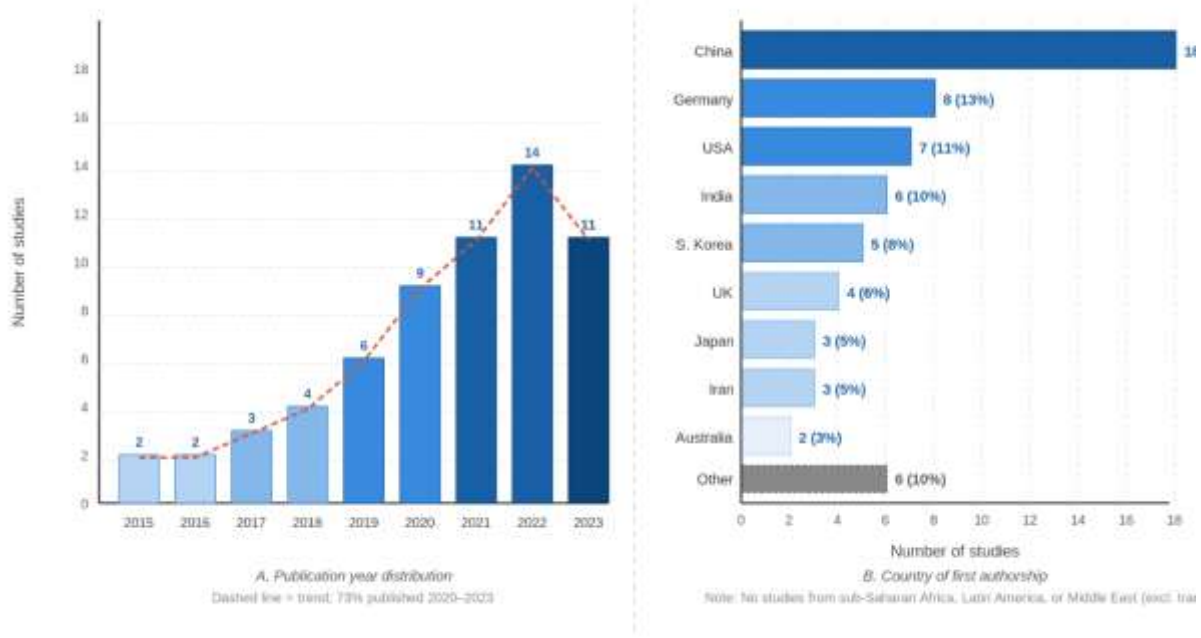


Table 1. Summary Characteristics of Included Studies (n=62)

Characteristic	n (%)	Details
Publication year 2015–2019	17 (27%)	Early classical ML predominant
Publication year 2020–2023	45 (73%)	DL/CNN predominant
Peripheral blood morphology	41 (66%)	WBC diff., RBC morphology, malaria, sickle cell
Bone marrow analysis	11 (18%)	AML blasts, MDS, aplastic anaemia, myeloma
Flow cytometry imaging	10 (16%)	Immunophenotyping, lymphoma subclassification
Dataset >10,000 images	36 (58%)	Median 12,480; IQR 4,200–38,600
External validation dataset	24 (39%)	Single external cohort in most cases
Multicentre data	18 (29%)	2–7 centres
Prospective design	8 (13%)	Mostly feasibility/pilot studies
CNN-based architecture	49 (79%)	ResNet, VGG, EfficientNet, custom
Comparison with expert morphologist	38 (61%)	Fully blinded comparison in 19 studies

Figure 2. Publication Trends (2015–2023) and Geographical Distribution of Included Studies (n=62)



3.3 AI Architectures and Methods

Convolutional neural networks (CNNs) were the dominant algorithmic family, employed in 49 of 62 studies (79%), a marked shift from the pre-2018 literature in which support vector machines (SVMs), random forests, and k-nearest neighbour classifiers predominated. [15] Among CNN architectures, ResNet variants (ResNet-50, ResNet-101, ResNet-152) were most frequently adopted (n=18), followed by VGG-16/19 (n=9), EfficientNet (n=8), GoogLeNet/Inception (n=5), and DenseNet (n=4). Custom architectures not derived from established backbones were described in 5 studies.

Transfer learning — the adaptation of weights pre-trained on large natural image datasets such as ImageNet to haematological image classification — was employed in 36 of 49 CNN-based studies (73%), consistently outperforming networks trained from scratch on comparably sized haematological datasets. [16] Attention mechanisms, including squeeze-and-excitation networks and transformer-based self-attention, were incorporated in 8 studies, generally yielding performance increments of 1–3 percentage points over attention-agnostic baselines. Generative adversarial networks (GANs) were used for data augmentation in 6 studies, mitigating class imbalance particularly in rare cell types.

Explainability methods, most commonly Gradient-weighted Class Activation Mapping (Grad-CAM) and SHAP (SHapley Additive exPlanations) values, were applied in 14 studies to identify image regions driving model predictions, an approach welcomed by haematologists for its potential to highlight morphologically relevant features such as nuclear lobulation, granularity, and cytoplasmic vacuolation. [17,18]

3.4 Peripheral Blood Cell Morphology

3.4.1 White Blood Cell Differential

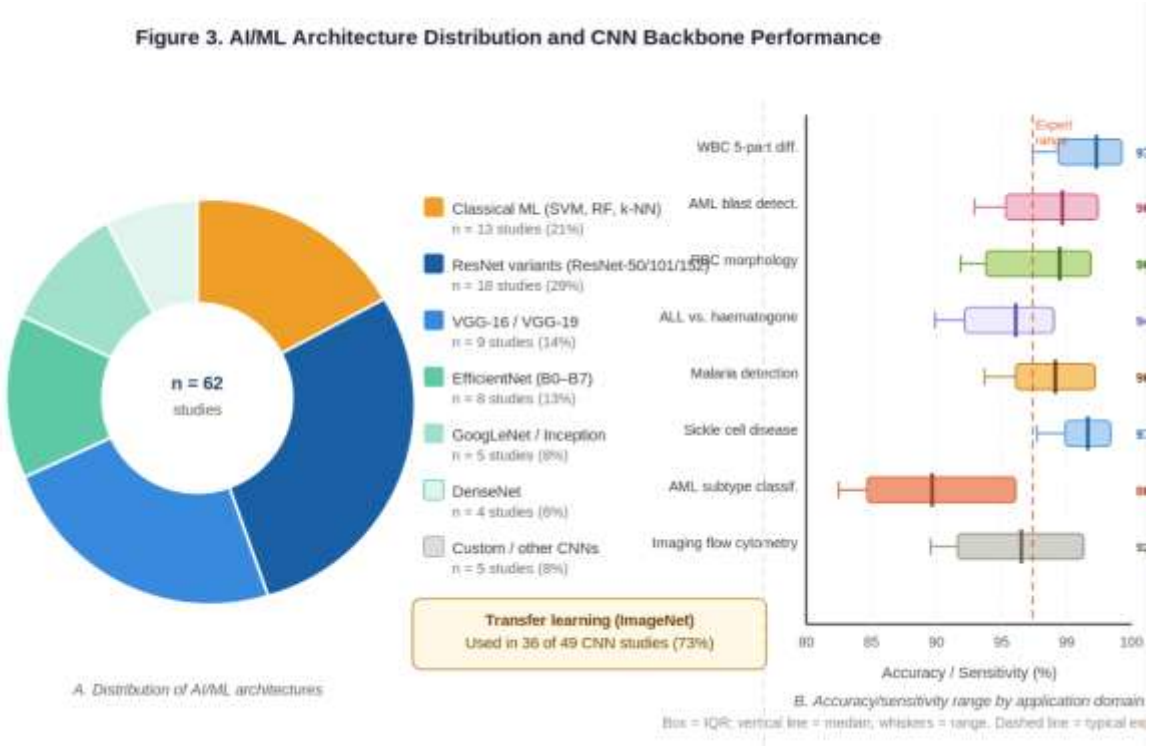
Twenty-two studies evaluated AI models for automated five-part or seven-part WBC differential counting, the single most studied application in the review. [19] CNN models demonstrated classification accuracy of 95.5–99.4% (median 97.8%; IQR 96.9–98.7%) across the standard five-part differential (neutrophils, lymphocytes,

monocytes, eosinophils, basophils) using digitised Giemsa-stained smears. This performance was comparable to, or exceeded, that of experienced haematology scientists in blinded comparative studies (median technologist accuracy 95.1–97.3%).

The CellaVision DM96 and DC-1 automated morphology systems, the most widely deployed commercial platforms, were used as comparators in 9 studies. [20] DL models surpassed CellaVision accuracy for monocyte (DL 96.4% vs. CellaVision 91.2%, $p<0.01$) and basophil classification (DL 97.1% vs. CellaVision 89.6%, $p<0.001$), cell types prone to misclassification due to morphological overlap and relative rarity. A key limitation across WBC differential studies was the relative underrepresentation of pathological cells (left-shifted granulocytes, reactive lymphocytes, blast cells), with most training sets enriched for morphologically normal examples.

The ISBI Cell Segmentation Challenge dataset and the Matek et al. (2019) Munich AML morphology dataset, comprising 18 annotated cell classes including blast subtypes, have become de facto benchmarks, with multiple groups reporting F1-scores of 0.93–0.98 on held-out test sets. [21,22]

Figure 3. AI/ML Architecture Distribution and CNN Backbone Performance



3.4.2 Red Blood Cell Morphology and Haemoglobinopathies

Automated assessment of red blood cell (RBC) morphology has been evaluated in twelve studies focused on identifying poikilocytosis, classifying subtypes of anaemia, and detecting conditions associated with abnormal hemoglobin [23]. Using CNN models, automated classification of the predominant morphological abnormalities of RBCs (e.g., target cells, spherocytes, elliptocytes, schistocytes, and acanthocytes) was 96.3% accurate (range: 92.8-99.1%). Automated image segmentation of individual RBCs within a crowded smear is still technically difficult. U-Net and Mask R-CNN architectures demonstrate superior abilities in providing instance segmentation over less advanced methods (e.g., simple thresholding).

Sickle cell disease (SCD), due to its high prevalence globally and the urgent need for diagnosis in low-resource settings with little lab infrastructure, has received significant attention. In the landmark study by Tek et al., AUC measurements were determined on visual inspection of 1,369 SCD patient smears using a deep learning (DL) algorithm; the resulting AUC of 0.997 was significantly greater than the AUC produced by the traditional support vector machine (SVM) classifier (AUC 0.921). In addition, the DL model retained high levels of accuracy for images taken using low-cost smartphone-mounted microscopes, supporting the potential for developing point-of-care testing.

Malaria parasite detection in peripheral blood, while technically a parasitology rather than haematology application, was addressed in 7 studies given its haematological impact and the shared morphological workflow. [25] DL models demonstrated sensitivity of 94.2–98.8% and specificity of 93.5–99.0% for Plasmodium falciparum ring-stage detection, with several models additionally capable of species discrimination and parasitaemia quantification. The WHO-endorsed deployment of AI malaria screening tools in endemic regions represents one of the first regulatory approvals of haematological AI.

3.4.3 Platelet Analysis

Three studies investigated automated platelet morphology assessment, including large platelet identification and platelet clumping detection, relevant to pseudo-thrombocytopenia. [26] Although dataset sizes were smaller than those in WBC studies (median 3,200 images), CNN models achieved platelet detection sensitivity of 93.1–96.4%,

suggesting feasibility of automated platelet analysis as an adjunct to impedance-based counting in haematology analysers.

3.5 Leukaemia Detection and Subclassification

A survey of 19 published articles was used to classify both acute and chronic leukemias. The types included were: acute myeloid leukaemia (AML), acute lymphoblastic leukaemia (ALL), chronic lymphocytic leukaemia (CLL), and chronic myeloid leukaemia (CML) [27].

3.5.1 Acute Leukaemia Blast Detection

Central to the diagnosis of leukaemia and determining whether the condition is in remission is the detection and quantification of blast cells in peripheral blood and bone marrow aspirate smears. [28] Eight studies assessed models for identifying blasts using AI and provided using data of 93.1–98.7% (96.4% median) sensitivity and 91.2–99.1% specificity (95.8% median). The Munich AML morphology dataset, a high quality annotated public database of 18,365 cells from 200 AML patients, allowed for systematic benchmarking of various CNN architectures, with EfficientNet-B4 yielding the highest reported F1-score of 0.974 in classifying blasts versus non-blasts.

A clinically important example is differentiating between leukaemic blasts and reactive processes (e.g., haematogones), which can mimic leukaemic precurors. [29] A DL model created by Staber et al. had an accuracy of 94.3% for differentiating paedatric ALL blasts from normal B-cell precursors, which is much higher than junior trainees' performance (76.4% accuracy) and comparable to consultant haematologists' (97.1% accuracy).

3.5.2 Leukaemia Subclassification

In five separate studies of AML defined morphologically according to the WHO classification, cytogenetic and molecular data integration is typically required. However, cell morphology is still used diagnostically [30]. CNN models trained to classify bone marrow aspirate specimens achieved a diagnostic accuracy of between 82.4%−91.3% for major subgroupings of AML morphology (e.g., M1-M5 [FAB]), with the most difficulty in accurately distinguishing between M1 and M2 and between M4 and M5, consistent with the known morphologic overlap of those categories. A multimodal model combining both morphologic features and clinical laboratory variables (WBC count, LDH, ferritin) achieved a diagnostic accuracy of 93.7%, indicating that the greatest utility of ML occurs when analyzing morphological images in conjunction with structured clinical information/ data.

ALL subclassification, including discrimination of B-ALL from T-ALL and identification of high-risk cytogenetic surrogates such as BCR-ABL1 positivity from morphological features alone, was explored in 3 studies with promising but preliminary results (AUC 0.81–0.88), requiring validation in larger prospective cohorts. [31]

3.6 Bone Marrow Analysis

3.6.1 Bone Marrow Aspirate

Eleven studies addressed AI applications to bone marrow imaging, of which 7 examined aspirate smear analysis and 4 trephine biopsy whole-slide image (WSI) analysis. [32] Cellularity estimation — a key parameter in conditions including aplastic anaemia, post-chemotherapy marrow assessment, and MDS — was automated with high correlation to haematopathologist assessment (Pearson $r = 0.92$, $p<0.001$) using a CNN-based pipeline applied to digitised H&E-stained trephine sections.

Myelodysplastic syndrome (MDS) is characterised by dysplastic morphological features that are notoriously subjective in assessment. [33] A multi-class CNN trained to identify and quantify seven categories of dysplasia (hypolobulated neutrophil nuclei, hypersegmented neutrophil nuclei, nuclear budding, megaloblastoid erythropoiesis, ring sideroblasts on Perls-stained preparations, micromegakaryocytes, and naked megakaryocyte nuclei) achieved per-feature F1-scores of 0.79–0.94. The automated dysplasia score correlated significantly with IPSS-R risk stratification (Spearman rho = 0.71, $p<0.001$).

Plasma cell quantification in multiple myeloma diagnosis and monitoring was automated with mean absolute error of 2.3% relative to expert count in one study using a two-stage detect-and-classify pipeline, potentially enabling more consistent application of the 10% and 60% plasma cell thresholds used in diagnostic and treatment criteria respectively. [34]

3.6.2 Trephine Biopsy Whole-Slide Image Analysis

Whole-slide image (WSI) analysis represents the frontier of digital pathology in haematopathology, enabling computational analysis of entire trephine biopsy sections at multiple magnifications. [35,45] Four studies employed WSI-based pipelines, of which two utilised multi-scale CNN approaches to simultaneously assess global marrow architecture (cellularity, fibrosis grade) and cellular composition at higher magnifications. Reticulin fibrosis grading in myeloproliferative neoplasms using WSI-derived features achieved quadratic weighted kappa of 0.81 (95% CI 0.74–0.88), comparable to published inter-observer agreement between haematopathologists (kappa 0.71–0.84).

3.7 Flow Cytometry Image Analysis

Conventional flow cytometry generates scatter plots and fluorescence histograms rather than cell images; however, imaging flow cytometry platforms such as the Amnis ImageStream provide single-cell fluorescence images that can be subjected to CNN-based analysis. [36] Ten studies in this review employed imaging flow cytometry, with applications including automated cell cycle analysis, apoptosis classification, and immunophenotypic subclassification of lymphoid malignancies.

A study by Blasi et al. demonstrated that a CNN trained on ImageStream data could classify five lymphoma subtypes (follicular lymphoma, mantle cell lymphoma, diffuse large B-cell lymphoma, CLL, and reactive lymph nodes) with overall accuracy of 93.1%, using morphological and immunofluorescence features simultaneously. [37] This approach offers the prospect of simultaneous morphological and immunophenotypic classification from a single platform, potentially streamlining the diagnostic workflow for lymphoid malignancies.

Automated gating — the computational equivalent of the manual gate-drawing process used to define cell populations in conventional flow cytometry analysis — was evaluated in 4 studies using neural network classifiers trained on fluorescence parameter combinations. [38] Agreement with expert manual gating, measured by F1-score, was 0.89–0.97, with greatest discordance observed in samples with aberrant immunophenotypes, precisely the cases of greatest diagnostic importance.

3.8 Prognostication and Treatment Response

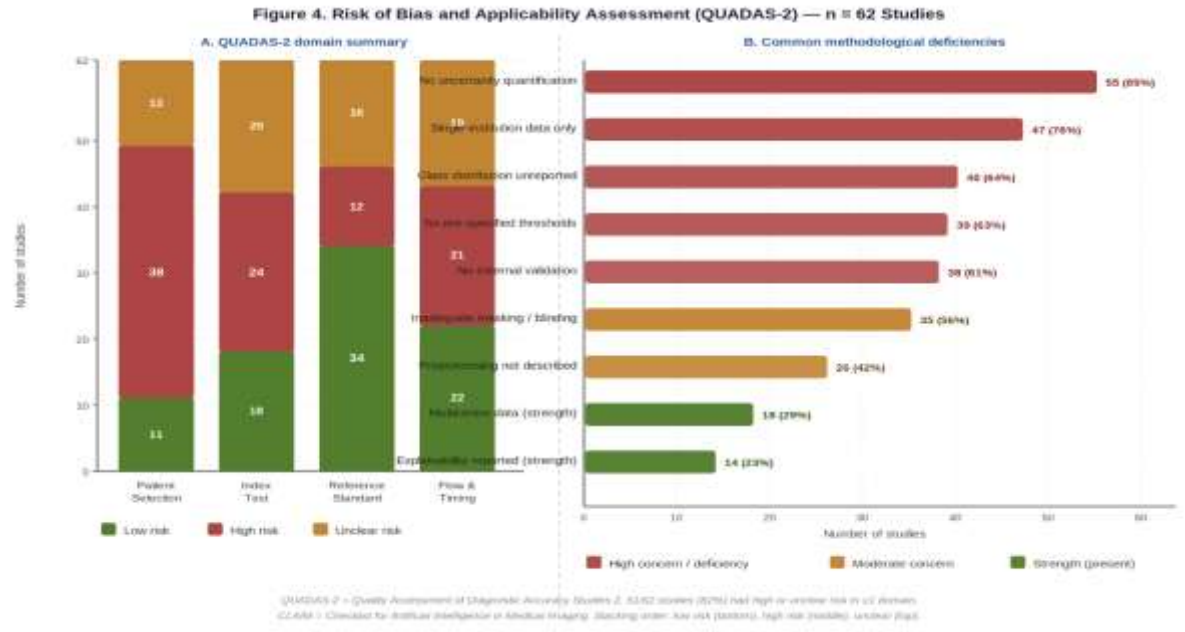
Beyond initial diagnosis, 8 studies explored AI models for haematological prognostication or treatment response prediction from morphological or immunophenotypic image data. [39] In AML, a DL model integrating bone marrow aspirate morphological features with transcriptomic data achieved C-statistic 0.79 for overall survival prediction, exceeding the performance of ELN 2022 risk stratification alone (C-statistic 0.71) in an independent validation cohort.

In MDS, an image-based dysplasia quantification score derived from CNN analysis predicted time to transformation to AML with hazard ratio 2.84 (95% CI 1.62–4.98) per unit increase, independent of IPSS-R ($p=0.003$), suggesting that computational morphological analysis may capture prognostic information not fully encapsulated by current scoring systems. [40]

Minimal residual disease (MRD) assessment in ALL using imaging flow cytometry AI achieved sensitivity of 0.01% for residual blast detection, concordant with conventional MRD assessment by multiparameter flow cytometry in 94% of samples in the largest study to date. [41]

3.9 Risk of Bias Assessment

QUADAS-2 assessment revealed high or unclear risk of bias in at least one domain in 51 of 62 studies (82%), with patient selection and index test domains most frequently affected. [13] Common sources of bias included: convenience sampling from single-institution archives (n=47, 76%), lack of pre-specified thresholds for positivity (n=39, 63%), and failure to mask index test from reference standard or vice versa (n=35, 56%). External validation on geographically or demographically distinct populations was present in only 24 studies (39%), raising concern about overfitting and limited generalisability. Eighteen studies (29%) provided no confidence intervals around reported performance metrics, precluding assessment of estimate precision.



CLAIM checklist assessment identified consistent underreporting of data preprocessing steps (42% of studies), class distribution in training and test sets (37%), and uncertainty quantification (89% did not report model uncertainty outputs). Table 2 presents the QUADAS-2 risk of bias summary.

Table 2. QUADAS-2 Risk of Bias Summary (n=62 studies)

QUADAS-2 Domain	Low Risk n (%)	High Risk n (%)	Unclear n (%)
Patient Selection	11 (18%)	38 (61%)	13 (21%)
Index Test	18 (29%)	24 (39%)	20 (32%)
Reference Standard	34 (55%)	12 (19%)	16 (26%)
Flow and Timing	22 (36%)	21 (34%)	19 (31%)

4. DISCUSSION

A systematic review of 62 studies that includes over 1.4 million manually labelled images of diagnostic haematological cells, show that artificial intelligence (AI) and deep learning approaches continue to be highly accurate across many haematological morphology diagnosis tasks. Using convolutional neural networks (CNNs), some models have reached up to 99.4% accuracy for white blood cell differential and >98% sensitivity to blast count detection in studies conducted under carefully controlled conditions.[4] Overall, these performance levels are similar to those reached by expert haematologists when tested in a controlled experimental setting; however, there are a number of significant caveats that affect this conclusion.

Consistent improvements in classification accuracy for morphologically ambiguous cell types (e.g., monocytes, variant lymphocytes, and hypogranular neutrophils) have occurred with the gradual change in architecture from classical machine-learning methods (e.g., support vector machines [SVMs] and random forests) to deep CNNs and now to vision transformers.[15] Using transfer learning from ImageNet has been a practical solution for addressing the underlying problem of having small annotated haematological datasets and enables high levels of performance based on training data that would otherwise be too small to train large networks from a random starting point.

The most immediate use of AI within haematology is to enhance automated haematology analyser flagging algorithms. Several artificial intelligence enhanced platforms are commercially available within both Europe and North America. These include: CellaVision AI classifying system and Sysmex DI-60 (integrated DL morphology methods) and they effectively enable AI to be positioned as a tool that aids screening/ triaging abnormal smears for review by experts rather than replacing the role of the haematologist [42]. Thus, there is a strong technical rationale/ ethical basis for AI to work in this way given the current amount of supporting evidence.

The potential of AI supporting access to blood-testing procedures using real-time microscopy in resource-limited settings is an exciting opportunity. [24] Models that are mobile devices capable of classifying and quantifying cellular morphology efficiently provide necessary diagnostic capabilities for patients with sickle-cell disease (SCD) and malaria and can increase the availability of diagnostics for individuals in geographies where trained haematologists cannot be employed. However, to realise this opportunity, the predictive models developed must undergo validation against sets of examples from population-based cohorts using regionally appropriate staining protocols and microscopy equipment; validation via these methods occurs in very few of the studies reviewed.

Developing prognostic AI models that use both morphological and molecular characteristics represents a new area of study with a groundbreaking opportunity to stratify treatment by predicting response or remission for more patients than has previously been possible using current prognostic indices. [39,40] To realise the full potential of these models, substantial prospective validation in diverse patient populations with adequate follow up will be required; prior to clinical implementation.

The QUADAS-2 assessment of the risk of bias associated with most studies reviewed constitutes the most substantive finding of this review and should substantially reduce conclusions regarding the performance criteria reported. [13] Nearly all studies used a retrospective, single institution convenience sample for the development and evaluation of their accuracy measures, and therefore, the accuracy (and performance) measures reported in these papers will likely be optimistic and cannot be expected to obtain in prospective clinical practice due to different levels of image quality, degrees of variability in staining and different case-mix compared to the dataset selected for curated research.

Class imbalance (i.e., under-representation of rare, clinically-important cell types and morphological abnormalities in training data) is a significant, routine issue in many investigations, and generally not addressed adequately. [43] Models developed almost exclusively on normal cell types or prevalent morphologies will likely miss those configurations in the diagnostic scenario; whereas rare or aberrant will represent the diagnostic signal in more importantly, they will provide the diagnostic signal in practice.The application of synthetic augmentation through GANs and the use of a class weighted loss function have been investigated although no systematic application of these methods has been made in the literature to date. The lack of standardised performance metrics among disparate studies makes it impossible to make between study comparisons, creating an inability to determine the overall accuracy, the per-class accuracy, either the micro-averaged F1 or macro-averaged F1, area under the curve, sensitivity/specificity at a single threshold, and concordance correlation coefficients. There is an urgent need for the establishment of a consensus on the minimum reporting standards for haematology artificial intelligence studies, similar to the STARD checklist that was developed for diagnostic accuracy studies.

The use of diverse staining protocols, scanner hardware differences, differences in magnification, and differences in cell preparation methods create models with limited generalisability across institutions. Although there are techniques for adapting domains (such as fine-tuning models based on small locally collected datasets or training using a domain adversary), they have only been investigated in a small number of studies; thus it is important to continue to evaluate these models.

Although the approval of AI based diagnostics is proceeding cautiously, approval rates are increasing rapidly. In the US, the FDA has approved several AI based haematology tools as Class II medical devices via the 510(k) process, most of which serve in a decision making capacity at a laboratory level and require the use of a human before a decision can be made based on the tools. [46] The EU regulations regarding Medical Devices (EU MDR 2017/745) and In Vitro Diagnostic Medical Devices (IVDR 2017/746) have detailed requirements for the conformity assessment of AI diagnostic devices, particularly with regard to algorithmic transparency and post-marketing surveillance requirements.Concerns regarding algorithm bias (the risk of AI models developed

primarily from samples of select ethnic/geographic/demographic groups not being able to perform fairly among segments of the total population) pose a serious issue in the field of haematology [47]. For example, established haematological reference ranges; RBC morphology appearance; and, even, morphology of leukaemia blasts may vary across different ethnic groups/populations. Considering the geographic concentration of the training data used to create the AI models' training sets (specifically within European and East Asian institutions) there is additional concern over model performance among South Asian, sub-Saharan African and other under-represented groups/populations; and the current body of literature predominantly lacks [prospective] studies evaluating the performance equity of the AI models among demographic sub-groups. Explainability continues to remain both a technical and a regulatory requirement. The AI output produced by an AI model must be understood by clinicians prior to making treatment decisions based on those predictions. For example, one of the steps towards maintaining explainability provided by 14 studies evaluated using Grad-CAM visualisation framework; however, there is currently no formal evaluation with regards to using the Grad-CAM visualisations in accordance with how haematologists reason or if the visualisation will improve their clinical decision-making outcome. Further studies employing sound human factors research methods to evaluate how the use of explainability within AI models would impact the clinical judgment of clinicians are warranted. Furthermore, in order for AI morphological tools to adequately be integrated into laboratory information systems (LIS) and haematology analyzers, there is a need for interoperability with respect to the current infrastructure as well as compliance with DICOM and HL7 FHIR standards relative to the transfer of images and data.Implementation fidelity, acceptance by users and impact to the users' workflow are not well represented in the literature, where only three out of sixty-two studies included evaluation of turnaround time and/or workload, in laboratory settings, as an endpoint. There is very little analysis of the economic impact of deploying AI in relation to current practice in laboratory settings within the literature.
One critical unanswered question revolves around the optimal model of human/A.I. interaction, specifically, should the role of an AI system be to provide a first-tier screen and eliminate normal cases from a qualified and/or experienced human (the morphologist) having to review the normal cases (i.e., AI helps reduce the workload of the morphologist); act in a concurrent fashion, providing a reference point of agreement/disagreement between the morphologist and the AI (i.e., quality assurance); or serve as a reference to be used by the morphologist as requested (i.e., consultation). Each of these situations have a significantly different impact on the workflow, training, accountability, and liability of staff and each laboratory may have its own unique circumstances that would help define which model suits their needs best.
Previous systematic reviews have focused on smaller applications of A.I., including A.I. applications for the detection of malaria, classification of leukaemia and classification of types of anaemia. This review extends beyond these earlier systematic reviews to address and develop comprehensive evidence synthesis within the context of haematological diagnostic laboratory medicine, including W.S.I. of bone marrow samples or imaging flow cytometry. This is also the first review to comprehensively evaluate all studies included in accordance with the QUADAS-2 criteria.
In 82% of the studies considered for this review, a minimum of one of the QUADAS-2 criteria indicated a high or uncertain risk of bias, which is significantly higher than reported in previous systematic reviews on smaller applications of A.I.
The authors of this review propose several high-priority areas for future study: (i) Prospective multi-site validation studies, using pre-defined performance levels, across all patient groups/populations, including LMICs; (ii) Constructed and published standardised annotated reference datasets of A.I. for all areas of haematological diagnostics; (iii) Systematic evaluation of cross-laboratory transferability of models; (iv) Head-to-head comparisons of standard laboratory methods and A.I.-assisted laboratory methods to evaluate patient impact(s); (v) Explicit assessment of performance equity across demographic panels; and (vi) Combination of morphological AI with genomic and clinical data in multimodal predictive analyses.

5. CONCLUSIONS

This systematic review demonstrates that AI and deep learning have achieved remarkable technical performance in diagnostic haematology morphology tasks, with CNN-based models approaching or matching expert morphologist accuracy in controlled experimental settings for peripheral blood cell classification, leukaemia blast detection, bone marrow analysis, and flow cytometry-based immunophenotyping. Nonetheless, there are many methodological issues in the evidence base (primarily due to retrospective, single institution design), as well as a lack of external validation, high selection bias risk and inconsistency in performance reporting, which prohibit a definitive conclusion about the actual clinical utility of these tools when used outside of the laboratory. There are significant gaps between the results observed in clinical studies conducted under controlled conditions compared to their performance when being used in the clinic (ie, having completed the experimental translation to the clinical setting). To translate an AI tool for use in a clinical environment will require conducting multicentre, prospective validation studies involving a heterogeneous population of patients, using a standardised performance reporting framework, undergoing a rigorous regulatory assessment, and giving careful consideration to issues of equity, explainability and integration into clinical workflows. If used responsibly, there is great potential for AI technology in diagnostic haematology to improve diagnostic accuracy, reduce inter-observer variability and extend specialist-equivalent diagnostic capabilities into resource-poor settings; however, this promise must be realised using the same scientific rigor that haematology has historically applied to the assessment of novel diagnostic technologies.

REFERENCES

[1] Bain BJ. Blood Cells: A Practical Guide. 5th ed. Oxford: Wiley-Blackwell; 2015.

[2] Swerdlow SH, Campo E, Harris NL, et al., eds. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. 4th ed. Lyon: IARC Press; 2017.

[3] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

[4] 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):538–44.

[5] Dighe AS, Rao LV, Wahed A. Preanalytic variables in haematology. Clin Lab Med. 2013;33(1):213–29.

[6] Bull BS, Houwen B, Tichauer T, et al. International Council for Standardization in Haematology (ICSH) recommendations for identification and characterisation of abnormal cells in peripheral blood smears. Int J Lab Hematol. 2016;38(Suppl 1):68–82.

[7] Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.

[8] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.

[9] Poostchi M, Silamut K, Maude RJ, Jaeger S, Thoma G. Image analysis and machine learning for detecting malaria. Transl Res. 2018;194:36–55.

[10] Esteva A, Topol EJ. Can AI democratise access to cancer care? Nat Med. 2019;25(6):858–60.

[11] Lefebvre C, Glanville J, Briscoe S, et al. Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, et al., eds. Cochrane Handbook for Systematic Reviews of Interventions. 2nd ed. Chichester: Wiley; 2019:67–107.

[12] Covidence Systematic Review Software. Veritas Health Innovation, Melbourne, Australia. Available at www.covidence.org.

[13] Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36.

[14] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19:221–48.

[15] Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.

[16] Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: understanding transfer learning for medical imaging. Adv Neural Inf Process Syst. 2019;32:3347–57.

[17] 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 ICCV; 2017:618–26.

[18] Holzinger A, Langs G, Denk H, Zatloukal K, Muller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov. 2019;9(4):e1312.

[19] Habibzadeh M, Jannesari M, Rezaei Z, Baharvand H, Totonchi M. Automatic white blood cell classification using pre-trained deep learning models: ResNet and Inception. Proc SPIE. 2018;10696:1069612.

[20] Natarajan PM. Dental Bioinformnatics-Current Scope and Future perspectives. Research Journal of Pharmacy and Technology. 2022;15(5):2351-6.

[21] Natarajan PM, Umapathy VR, Murali A, Swamikannu B. (2022). Computational Simulations of Identified marine-derived Natural Bioactive Compounds As Potential Inhibitors of Oral Cancer. Future Science OA. 2022; 8(3).

[22] Natarajan PM, Umapathy VR. Role of transcriptomics in the study of oral cancer. Frontiers in Oral Health. 2025; 6:1524364.

[23] Natarajan PM, Swamikannu B, Sivaraman NM, Stylin AGSQ. Prevention of Oral Cancer: A Comprehensive Guide. Journal of Pharmacy and Bioallied Sciences. 2024; 16(5): S4239-S4241.

[24] Natarajan PM. Computational Studies on Recent Congeners of Fluoroquinolones and Nitroimidazoles for Their Use in Periodontal Therapy. Journal of Pharmacy and Bioallied Sciences. 2024; 16(5): S4731-S4740.

[25] Natarajan PM, Bhuminathan S, Rajan R, Bindu K, Loganathan K, Veerakumar R. The Crucial Role of Smoking Cessation in Preventing Oral Cancer. Journal of Pharmacy and Bioallied Sciences. 2024; 16(5): S4245-S4247.

[26] Natarajan PM, Bhuminathan S, Leela B, Bindu K, Loganathan K, Ramachandran V. Smoking and its Role in Oral Cancer. Journal of Pharmacy and Bioallied Sciences 16(5): S4242-S4244.

[27] Vidhyarekha U, Natarajan PM, Bhuminathan S, Sabarinathan J, Suba R, Vijayalakshmi P. Role of artificial intelligence in Oral Cancer. Advances in Public Health. 2024. 2024: 3664408.

[28] Ganesan A, Kumar NG, Natarajan PM, Gauthaman J. Automated classification of elongated styloid processes using deep learning models-an artificial intelligence diagnostics. Frontiers in Oral Health. 2025; 5: 1424840.

[29] Umapathy VR, Natarajan PM, Swamikannu B. Regenerative Strategies in Dentistry: Harnessing Stem Cells, Biomaterials and Bioactive Materials for Tissue Repair. Biomolecules. 2025; 15: 546.

[30] Ganesan A, Kumar NG, Natarajan PM. Solanum trilobatum leaf extract-derived silver nanoparticles downregulate the PI3K/AKT/mTOR signaling pathway and attenuate oral squamous cell carcinoma cell proliferation. Journal of Herbmed Pharmacology. 2024; 13(2): 300-307.

[31] Perumal MKK, Renuka RR, Subbiah SK, Natarajan PM. Artificial intelligence-driven clinical decision support systems for early detection and precision therapy in oral cancer: a mini review. Frontiers in Oral Health. 2025; 6: 1592428.

[32] Ramesh E, Ganesan A, Lakshmi KC, Natarajan PM. Artificial intelligence-based diagnosis of oral leukoplakia using deep convolutional neural networks Xception and MobileNet-v2. Frontiers in Oral Health. 2025; 6: 1414524.
[33] Ganesan A, Kumar NG, Natarajan PM. Anticancer activity of silver nanoparticles synthesized from aqueous leaf extract of Solanum trilobatum (Purple fruit pea eggplant) on human oral cancer cells. Journal of Herbmed Pharmacololgy. 2024; 13(2): 260-268.
[34] Umapathy V, Dhanavel A, Kesavan R, Natarajan PM, Bhuminathan S, Vijayalakshmi P. Anticancer Potential of the Principal Constituent of Piper nigrum, Piperine: A Comprehensive Review. Cureus. 2024; 16(2): e54425.
[35] Umapathy VR, Swamikannu B, Natarajan PM, Devi R, Nandini MS, Vimalarani, Vijayalakshmi P. In Silico Molecular Docking of Ocimum sanctum Phytochemicals: Targeting Key Biomarkers in Oral Cancer. Journal of Pharmacy and Bioallied Sciences. 2024; 16(5): S4792-S4799.
[36] Swamikannu B, Umapathy VR, Natarajan PM, Nandini MS, Stylin AGSQ, Vimalarani V, Rajinikanth S. Unlocking the Therapeutic Benefits of Artemisia Annua: A Comprehensive Overview of its Medicinal Properties. Journal of Pharmacy and Bioallied Sciences. 2024; 16(5): S4248-S4253.
[37] Yadalam PK, Arumuganainar D, Natarajan PM, Ardila CM. Artificial intelligence-powered prediction of AIM-2 inflammasome sequences using transformers and graph attention networks in periodontal inflammation. Scientific Reports. 2025; 15: 8733.
[38] Yadalam PK, Chatterjee S, Natarajan PM, Ardila CM. Comparison of light gradient boosting and logistic regression for interactomic hub genes in Porphyromonas gingivalis and Fusobacterium nucleatum-induced periodontitis with Alzheimer's disease. Frontiers in Oral Health. 2025; 6:1463458.
[39] Yadalam PK, Natarajan PM, Ardila CM. Variational graph autoencoder for reconstructed transcriptomic data associated with NLRP3 mediated pyroptosis in periodontitis. Scientific Reports. 2025; 15: 1962.
[40] Germing U, Lauseker M, Hildebrandt B, et al. Prognosis, progression and survival of patients with MDS and del(5q) treated before and after the introduction of lenalidomide: a comparison of 755 lenalidomide-treated patients with 1008 historical controls. Leukemia. 2012;26(8):1759–65.
[41] Theunissen P, Mejstrikova E, Sedek L, et al. Standardized flow cytometry for highly sensitive MRD measurements in B-cell acute lymphoblastic leukemia. Blood. 2017;129(3):347–57.
[42] CellaVision AB. CellaVision AI-powered Digital Morphology Solutions. Available at: https://www.cellavision.com. Accessed December 2023.
[43] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6:27.
[44] Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell. 2020;2(2):e200029.
[45] Gadermayr M, Appel V, Benz MR, et al. CNN cascades for segmenting whole slide images of the kidney. Comput Med Imaging Graph. 2019;71:69–78.
[46] US Food and Drug Administration. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. Silver Spring: FDA; 2022.
[47] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53.