Trends in AI-powered classification of thyroid neoplasms based on histopathology images, a systematic review.

Haitham KUSSAIBI M.D.[1*], Noor Alsafwani M.D.[2]

[1, 2] Department of Pathology, college of medicine, Imam Abdulrahman bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia.

*Corresponding author: hkussaibi@iau.edu.sa

Abstract:

Background: Assessment of thyroid nodules histopathology using AI is crucial for an accurate diagnosis. This systematic review aims to analyze recent works employing deep learning approaches for classifying thyroid nodules based on histopathology images, evaluate their performance, and identify limitations.

Methods: Eligibility criteria focused on peer-reviewed English papers published in the last 5 years, applying deep learning to categorize thyroid histopathology images. The PubMed database was searched using relevant keywords combination.

Results: Out of 103 articles, 11 studies met inclusion criteria. They used convolutional neural networks to classify thyroid neoplasm. Most studies aimed for basic tumor subtyping; however, 3 studies targeted the prediction of tumor-associated mutation. Accuracy ranged from 77% to 100%, with most over 90%.

Discussion: The findings from our analysis reveal the effectiveness of deep learning to identify discriminative morphological patterns from histopathology images thus leading to enhancing the accuracy of thyroid nodule histopathological classification. Key limitations were small sample sizes, subjective annotation, and limited dataset diversity. Further research with larger diverse datasets, model optimization, and real-world validation is essential to translate these tools into clinical practice.

Other
Funding: This study did not receive any funding.
Registration: The procedural instructions for this systematic review were officially recorded within the PROSPERO database under registration number **RD42023457854** https://www.crd.york.ac.uk/Prospero/

1. Introduction:

Thyroid nodules are frequently observed in clinical practice, and accurate diagnosis is crucial for proper patient management. The accurate characterization and classification of thyroid tumors, including distinguishing malignant from benign lesions, remains a significant challenge in pathology. Several diagnostic modalities, including radiology US (ultrasound), CT (computerized tomography) scan, MRI (magnetic resonance imaging), cytopathology, and histopathology, are in place to achieve this goal. While

histopathology examination is considered the gold standard for diagnosis, it can be sometimes challenging due to the subjective nature of interpretation by pathologists. Manual evaluation of thyroid histopathology slides is prone to interobserver variability especially in the borderline lesions such as NIFTP (Noninvasive Follicular Thyroid Neoplasm with Papillary-like Nuclear Features) (1). A common theme is developing computational approaches to complement pathologists' expertise and reduce the subjectivity and interobserver variability inherent to manual slide review and conventional microscopy. The utilization of AI (artificial intelligence) methods has demonstrated potential in enhancing the precision and efficiency of thyroid nodule histopathology diagnosis. Emerging DL (deep learning) techniques offer the potential to improve the efficiency and consistency of analyzing complex morphological patterns in thyroid histologic images. Recent studies have investigated the utility of deep CNNs (convolutional neural networks) and related DL approaches for the computational pathology of thyroid lesions (2, 3). They have focused specifically on developing DL systems to classify thyroid tumors/nodules into different subtypes including malignant vs. benign categories to guide surgical decision-making. The application of digital pathology and DL for analysis of thyroid histopathology images is an emerging area of research. While the approaches and aims vary, the collective work highlights the significant potential of these techniques to uncover novel insights from high-resolution thyroid histology WSIs (whole slide images).

Other than DL, some studies took a traditional image analysis approach for thyroid nodules classification based on nuclear morphology characterization (4, 5). **Valentim** et al. used computerized nuclear morphometry and texture analysis of histology images coupled with CRT (Classification and Regression Trees) to distinguish follicular pattern thyroid lesions. They demonstrated computerized quantitative image analysis can also provide discriminating nuclear features to classify thyroid tumors (6). While accurate for the study dataset, extensive feature engineering and limited diversity may affect scalability.

Although, DL-assisted classification of thyroid nodules based on radiology (7) or cytopathology (8-10) images is extensively explored in the literature, however, no recent review evaluated the performance of such techniques in classifying histopathology images. Herein, we aim, in this systematic review, to identify and summarize recent studies (published within the past 5 years) employing DL approaches in classifying thyroid neoplasms based on histopathology images. The primary objectives of the research are to evaluate the effectiveness of these methods, identify prevailing trends and patterns, and pinpoint areas that require further investigation.

2. Method:

2.1 Search Strategy:

The current systematic review is in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) statement (11). The procedural instructions for this systematic review were officially recorded within the PROSPERO database under registration number **CRD42023457854**. We conducted an extensive search in the Medline PubMed database, focusing on the last 5 years starting from January 2019 till September 2023. We used specific combinations of keywords such as 'thyroid', 'histopathology', 'images', 'classification', 'diagnosis', 'prediction', 'artificial intelligence', 'machine learning', 'deep learning', 'computer-aided/assisted', etc... Our search was limited to English original papers published in high-quality, peer-reviewed journals indexed in PubMed. To ensure the inclusion of relevant research, the PubMed query was finetuned several times using different combinations of keywords. Here is the final optimized fine-tuned query utilized:

(Thyroid[Title] AND (Histopatholog*[Tiab[1]] OR Histolog*[Tiab] OR Patholog*[Tiab]) AND (image*[Tiab] OR slide*[Tiab]) AND (classif*[Tiab] OR diagnos*[Tiab] OR predict*[Tiab]) AND (artificial intelligence[Tiab] OR machine learning[Tiab] OR deep learning[Tiab] OR convolutional[Tiab] OR neural network*[Tiab] OR automated[Tiab] OR computer-assisted[Tiab] OR computer-aided[Tiab]))

[1](Tiab stands for title/abstract)

2.2 Study Selection:

The first step of refining the search results included the screening of titles and abstracts to select the relevant works based on predetermined inclusion and exclusion criteria.

Inclusion criteria were as follows: (1) Original English-written studies published in the last 5 years indexed in PubMed that: (2) employed AI/ML/DL techniques, for classification of thyroid neoplasms based on histopathology images, (3) reported the performance metrics of the classification methods and (4) have free full text available.

Exclusion criteria included studies based on: (1) cytology images (2) radiology images (US, CT scan, MRI …) (3) pure morphometry and/or feature engineering.

2.3 Data Extraction:

We utilized a standardized form to extract pertinent information out of the selected studies from the screening process. The extracted data encompassed details such as author, title, objectives, data used (source, type, number), study design (approach), ML/DL model architecture with input images size and output classes, evaluation metrics, and limitations as reported by the authors. We conducted a descriptive analysis of the extracted data to uncover prevalent patterns and trends.

3. Results:

3.1 Overview of the included studies (Fig. 1)

Based on the above-mentioned methodology, first, we obtained a total 103 articles, refined to 95 articles after exclusion of none-English and none-original articles. By screening the titles and abstract according to the pre-defined inclusion and exclusion criteria, we excluded papers utilized images other than histopathology (e.g. cytopathology, US, CT, and/or MRI images …) to left out with 15 papers, 11 of them have free full text.

3.2 Study Characteristics:

A total of 11 studies meeting the inclusion criteria were identified. These studies were published between January 2019 and August 2023. Across the studies, the key trends included: 1) leveraging WSIs from histopathology glass slides. 2) extracting image patches/tiles to generate large, labeled datasets amenable for DL. 3) employing state-of-the-art CNNs pre-trained on natural images (ImageNet) and fine-tuned on histology data. 4) incorporating image preprocessing and augmentation strategies to boost model performance. 5) complementing DL with traditional image analysis and ML where limited labeled data is available. 6) most studies aimed for basic classification of thyroid neoplasms. 7) three studies showed promise in predicting specific genetic mutations linked to malignant behavior and prognosis from histomorphology patterns alone. 7) some authors experimented with training CNN models with hyperspectral images rather than traditional RGB images. 8) Frozen rather than FFPE (formalin-fixed paraffin-embedded) sections were utilized by 3 of the studies.

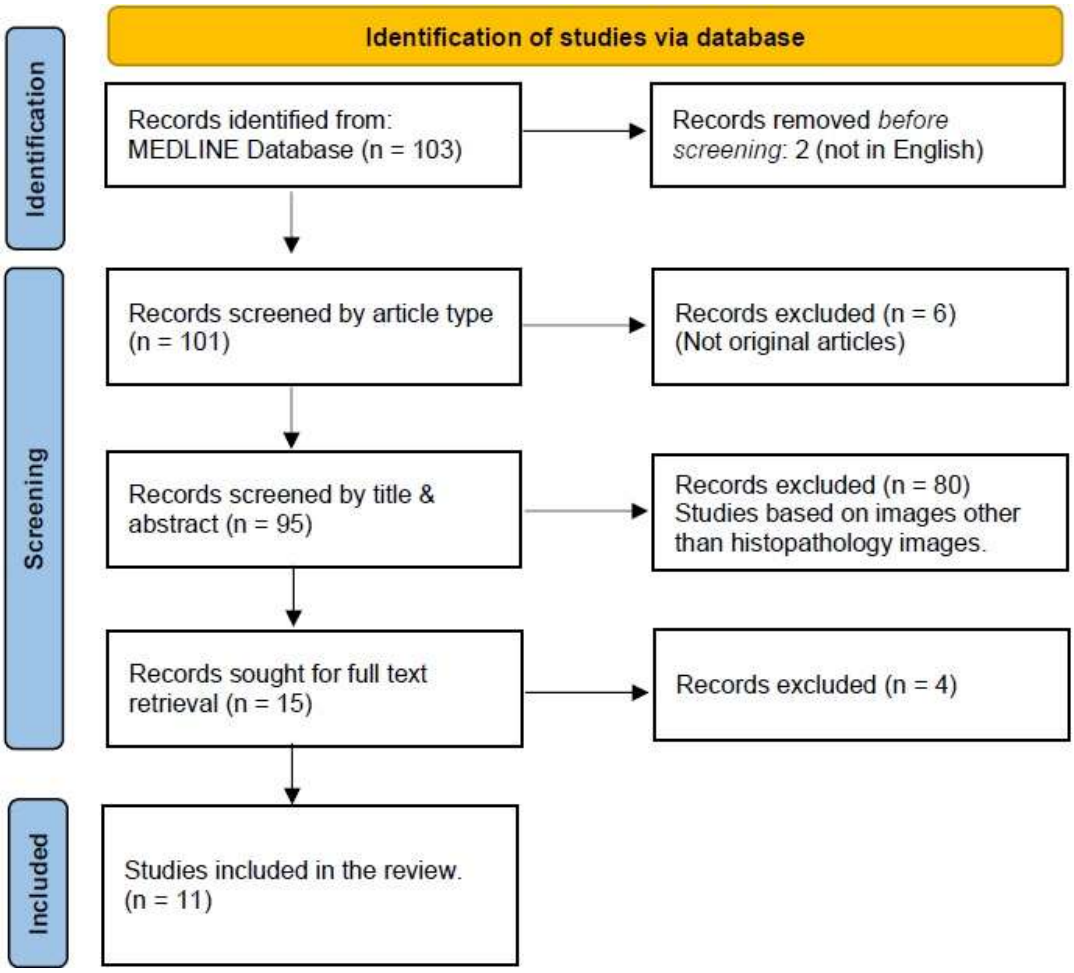

Figure 1 Prisma Flow Diagram

In addition to the selected 11 studies, there were other valuable studies sharing the same objectives, however, lacking a free full text to analyze. A study by **Anand** et al. 2021 described a weakly supervised technique to train a DNN (deep neural network) to predict BRAF V600E mutational status with high accuracy, in unannotated H&E-stained images of thyroid cancer tissue (12). Two studies by **Esce** et al. 2021 & 2023 demonstrated the efficacy of CNN in predicting the probability of nodal metastases in PTC (papillary thyroid carcinoma) based on the histopathology features of the primary tumor (13, 14). Finally, a study by **Nojima** et al. 2023 compared the prediction accuracy of 3 CNN models in distinguishing FA (follicular adenoma) from FTC (follicular thyroid carcinoma) based on histopathology images (15).

Table 1: overview of the selected studies including their objectives, methodologies and data used.

Paper	Title	Objectives	Data	Approach
Deng et al 2023	The value of deep neural networks in the pathological classification of thyroid tumors	Comparing different DNN in Classifying thyroid nodules into one of six categories.	799 thyroid images from 559 patients	-Manual tumor annotation -Manual patches extraction -Four DNN trained and compared in diagnostic efficiency
Bohland et al. 2021	Machine learning methods for automated classification of tumors with papillary thyroid carcinoma-like nuclei: A quantitative analysis	Classify thyroid tumors as PTC-like vs non-PTC-like	- Tharun & Thompson dataset (40x, 156 WSIs) - Nikiforov dataset (20x, 133 WSIs)	-Manual extraction of 10 neoplastic patches from each WSI. -Feature-based classification: nuclei segmentation, extraction of nuclei features (color & shape), preprocessing & classification) -Deep learning-based classification (training & classification)
Wang et al. 2019	Using deep convolutional neural networks for multi-classification of thyroid tumor by histopathology: a large-scale pilot study	Classifying thyroid nodules into (normal, goiter, FA, PTC, FTC, MTC, or ATC) subtypes	-100x ROIs from 806 patients. -11,715 patches	-Capturing ROI and labeling by pathologist -Automated Patches extraction -Nuclei detection by LOG[1] -Supervised Deep learning CNN
Zhu et al 2022	Deep Learning-Based Recognition of Different Thyroid Cancer Categories Using Whole Frozen-Slide Images	Recognition of Different Thyroid Cancer Categories Using Frozen sections WSIs by deep learning	1,374 WSIs of frozen sections from thyroid lesions	-Manual tumor annotation -Automated patch extraction -Patch-based UNet model was trained to classify between malignant and benign for segmentation -Decision tree on probability, size, diameter, and perimeter of lesions to identify a "rare" category
Li et al. 2020	Rule-based automatic diagnosis of thyroid nodules from intraoperative frozen sections using deep learning	Classifying thyroid frozen sections into benign, uncertain, or malignant	Frozen sections 40x WSIs (349 for training, 259 for evaluation) Number of patches (125000 for training, 25000 for evaluation)	-Training phase (manual region annotation, automated patch cropping, model training). -Prediction phase (tissue localization, patch splitting and prediction, rule-based fusion of patch prediction).
Chen et al 2020	Interactive Thyroid Whole Slide Image Diagnostic System using Deep Representation	CNN Classification of frozen sections histopathology images	345 thyroid frozen section slides scanned into 40x WSIs	-Manual lesion annotation -Automated patch extraction -CNN classification of patches -Fused patch prediction passed to traditional classifier (DT, SVM…) for ROI diagnosis.
Kim et al. 2023	Deep Learning Prediction of TERT Promoter Mutation Status in Thyroid Cancer Using Histologic Images	Predicting TERT promoter mutation status in thyroid cancer	25 institutional 40x WSIs (13 TERT mutant, 12 TERT wild-type)	Step1: Tumor detection by CNN (3 different models), automated image tiling, color transformation (HSV, HED). Step2: extract features from tumor tiles by RestNet152, then integrate features with RNN, finally classify TERT status with MLP.
Dolezal et al. 2021	Deep learning prediction of BRAF-RAS gene expression signature identifies noninvasive follicular thyroid neoplasms with papillary-like nuclear features	Classify thyroid tumors into BRAF-like or RAS-like mutation status	- 115 institutional 40x WSIs - 497 TCGA slides	-Tumor annotation by pathologist -Automated tiles extraction -Training a CNN model on institutional slides to predict tumor subtype -Training a CNN model on TCGA slides to predict BRS (BRAF-RAS score) -Evaluation the BRS model on the institutional slides
Tsou & Wu 2019	Mapping Driver Mutations to Histopathological Subtypes in Papillary Thyroid Carcinoma: Applying a Deep Convolutional Neural Network	Classifying PTC into BRAF vs RAS mutated subtypes	-TCGA 40x WSIs (51 RAS-mutant & 52 BRAF-mutant) -2575 patches extracted	-Tumor annotation by pathologist -Automated extraction of patches. -Training Inception V3 for prediction
Tran et al 2022	Thyroid Carcinoma Detection on Whole Histologic Slides Using Hyperspectral Imaging and Deep Learning	Thyroid Carcinoma Detection on WSIs Using Hyperspectral Imaging and Deep Learning	-33 institutional samples to produced 3 datasets: -RGB 40x WSIs -Hyperspectral images -HS-synthesized RGB images	-Manual tumor annotation -Automated patch extraction -Three separate CNNs (VGG19) were trained using the three datasets to predict tumor vs normal.
Halicek et al 2020	Tumor detection of the thyroid and salivary glands using hyperspectral imaging and deep learning	Evaluates hyperspectral imaging (HSI) for salivary and thyroid tumor detection	-200 thyroid and 16 salivary 40x WSIs	-Manual tumor annotation -HSI-synthesized RGB -Automated Patch extraction -Patch-based training of CNN for tumor detection

[1] Laplacian of the Gaussian function,

4. Discussion:

The current systematic review demonstrates the applications of DL and digital pathology techniques in automated analysis of thyroid histopathology images. The collection of works highlights a diversity of approaches while sharing the common aim of leveraging digital pathology imaging and artificial intelligence for enhanced analysis of thyroid histopathology. While all the studies utilized

histopathological tissue images as datasets, there were some key differences in their approaches which differed in terms of the objectives, type of samples, images' format, and analytical methodology, nevertheless, thyroid histopathology images proved to be a valuable dataset for demonstrating the utility of ML/DL-based techniques for enhanced classification and diagnosis.

4.1 Comparing Objectives: Exploring Simple Subtyping vs. Predicting Associated Mutations:

The specific classification tasks of the chosen studies differ. The objectives were not identical. Although the principal aim of most studies was basic diagnostic classification of thyroid neoplasms with generic tumor subtyping which has direct diagnostic relevance. However, three of the studies focused on predicting specific mutations associated with thyroid tumors, addressing the novel goal of predicting genetic mutations like TERT, BRAF or RAS directly from histomorphology features (16-18), demonstrating the potential to infer molecular characteristics from histology-imaging features.

4.2 Tissue sample preparation, processing, and staining:

The samples source and preparation methods varied among the studies. While most of them utilized FFPE thyroid tissue sections, three studies (19-21) used frozen sections. All stained with the conventional H&E protocol. Although H&E staining is the standard in pathology, Ozolek et al. 2014 (5) explored special stains such as Feulgen stain which preferentially binds DNA. This stain highlights nuclear chromatin patterns better than H&E stain. This is particularly relevant in thyroid tumors, where nuclear features are key for precise classification. Although H&E remains the typical staining technique, evaluations of complementary staining methods like Feulgen may further enhance the nuclear details essential for computational pathology applications in this setting (5).

4.3 Data sources: Institutional versus public cases:

While most of the studies retrieved institutional samples to build their datasets, two of them, also added public dataset (2, 16). One of the studied (18) used **public** dataset only. Tsou and Wu leveraged the large public TCGA thyroid cancer dataset consisting of H&E stained WSIs from PTC patients, they retrieved 51 and 52 samples with RAS and BRAF mutations, respectively. They obtained WSIs along with diagnostic labels, reducing the need for extensive in-house data collection and annotation (18). Dolezal et al. and Bohland et al. used both institutional and public cases. Bohland et al. evaluated two datasets. In the 1st dataset (Tharun dataset), they aggregated cases from two centers to obtain 154 slides spanning 5 thyroid tumor subtypes. The second dataset was the public Nikiforov collection (138 cases) (2). Dolezal et al. used 115 in-house WSIs in addition to 497 TCGA public cases (16).

The other studies needed to compile primary datasets due to the lack of suitable public resources for their particular research questions. Kim et al. used 25 institutional thyroid cancer cases (13 TERT-mutant, 12 TERT-wildtype) (17). Wang et al also used in-house samples from 806 patients (3). Similarly, Tran et al., Deng et al., and Halicek et al. collected institutional cases (33, 559 and 200 thyroid samples, respectively) (22-24). The remaining three studies gathered institutional frozen section slides, Li et al. (349 WSIs), Zhu et al. (1,374 WSIs), and Chen et al. (345 WSIs) (19-21).

All the mentioned studies scanned WSIs from patients' samples at 40x, or 20x magnification. The scale of the datasets varied based on the study needs, from around 500 TCGA, WSIs in Dolezal et al. study to just 25 WSIs analyzed by Kim et al. as an early proof-of-concept study. The creation of sufficiently large and representative datasets is crucial in developing ML systems for pathology applications. Larger sample sizes enabled more robust DL model development but require greater resources for annotation and training.

Both institutional samples and public datasets like TCGA can provide representative histopathology data for developing and evaluating DL solutions. Overall, careful data preparation tailored to the research questions remains critical to progress in this emerging field.

4.4 Image acquisition technique (RGB vs Hyperspectral):

RGB was the standard format for images dataset across most of the studies. To improve model performance, authors experimented with different image acquisition techniques by HS (hyperspectral) cameras. Two of the studies adopted acquisition of HS images (22, 23) which are captured across a wide range of wavelengths, beyond what the human eye can see. Each pixel in the image contains a continuous spectrum of measurements, rather than just RGB values. This allows for very detailed spectral information to be collected from the sample providing a powerful spectroscopic tool for histopathology analysis. Tran et al. (23) designed a CNN model called HyperDeep for pixel-wise prediction of thyroid cancerous regions on HS images. HyperDeep achieved an average accuracy of 95.3% which outperformed benchmark CNN models using just RGB images. Likewise, Halicek et al (22) used HS images to develop a CNN model called the Hyperspectral Tumor Detection Network (HTDN) which achieved a tumor detection accuracy of over 93% outperforming conventional histopathological assessment. Overall, the combination of rich HS data and DL algorithms showed promise for improving cancer diagnosis and evaluation.

4.5 Image preprocessing techniques

Many studies incorporated image preprocessing strategies to boost model performance. Kim et al. (17) attempted with images color transformations HSV (Hue, Saturation, Value) and HED (Hue, Exposure, and Dark) in different combinations to enhance histological images details. The best performance achieved with DenseNet161 (HSV-strong) + CRNN (HSV-strong). Wang et al (3), used LOG (Laplacian of the Gaussian function) to identify nuclei that appear dark under H&E staining.

4.6 Tumor segmentation/annotation (manual vs automated):

Both automated and manual approaches were used for tumor delineation. Automated techniques like CNN-based segmentation expedited these steps, while manual annotation ensured patches contained representative tumor regions. Kim et al. used CNN-based tumor detection and segmentation (17), while the other studies relied on manual tumor annotation by pathologists. In Bohland et al study (2), multiple pathologists were engaged to provide diagnoses and manual annotations of ROIs (region of interest) on the WSIs. Complexities arise due to the high-resolution WSIs used, the need for expert annotations, and variability in imaging conditions and protocols. However, the diagnostic and slide-level labels provided in public datasets (16) allow training of deep neural networks without new annotations.

4.7 Extraction of patches/tiles (manual vs automated):

Cropping multiple patches from ROIs or the entire WSIs, was commonly used across the studies to simplify analysis and expand limited datasets. Patch-based approaches, either manual or automated may suffer from selection bias. Only two studies followed manual patch extraction guided by pathologists (2, 24), while all the other studies adopted automated patch extraction approaches. Due to tumor heterogeneity, automated extraction of patches requires additional steps for tumor localization and annotation. The size of patches ranged from 2392 x 2392 pixels (20) down to 256 x 256 pixels (17) depending on the magnification and neural network input requirements. To match the network's default input size, the large patches need resizing which may lead to loss of fine-grained details or cropping which discard relevant information. Both can impact the model's performance and can affect its ability to

learn and generalize from the data. On the other hand, using very small patches could lead to overfitting, as the model may memorize details specific to those patches rather than learning more general features that are applicable to the entire dataset. Ultimately, the choice of patch size is guided by the trade-offs between capturing relevant information, computational constraints, and the DL models requirements.

4.8 Classification approaches: DL vs traditional image analysis:

Most of the studies relied on CNNs for classification of thyroid histopathology images. State-of-the-art CNN architectures like DenseNet, ResNet, VGG and Inception were commonly employed. Pre-training on natural images (ImageNet) provides a feature learning initialization before fine-tuning on histology data. While convenient, the biological domain shift remains a consideration.

While DL CNN models can learn discriminative features on their own, providing engineered features as additional inputs may improve performance for some histopathology image analysis tasks. Some works opt for classical image analysis pipelines based on hand-crafted features and ML classifiers, suggesting hybrid approaches may be beneficial where labeled data is scarce (2). Bohland et al. compared the performance of traditional feature engineering with that of DL models. They performed nuclear segmentation with U-Net model followed by nuclear features extraction and traditional ML classification. Aiming to simulate manual pathological nuclear scoring, they extracted a comprehensive set of 36 color, shape, and spatial features for each nucleus. With 700,000 segmented nuclei per case, they aggregated features by calculating the median and standard deviation across all nuclei in each patient sample to reduce the feature space while preserving relevant statistics. While achieving a slightly better performance, feature engineering approach is time consuming and necessitates tremendous efforts.

4.9 Base models:

Most of the studies tended to leverage ImageNet-pretrained CNNs solely or combined with other neural network components like LSTMs (Long Short-Term Memory) and MLP (Multi-Layer Perceptron) or with more traditional classifiers like SVM (Support Vector Machine). Kim et al. (17) used DenseNet161, VGG16, and EfficientNet_b4 for tumor detection, followed by a CRNN (Convolutional Recurrent Neural Network): ResNet152 + LSTM + MLP for TERT mutation prediction. MLP was also used by Halicek et al. (22) along with InceptionV4 for tumor detection. Zhu et al. (21) utilized a U-Net for segmentation with DT (decision tree) for classification. Bohland et al. (2), in their feature engineering-based approach, compared 6 different traditional classifiers (SVM, DT, …) for classification of nuclear features extracted from nuclear segmentation by a U-Net model. While, in their DL approach, they compared 3 CNN models ResNet18, ResNet50, and ResNet101. Comparison of multiple CNN models was also adopted by Deng et al (24) who compared the classification performance of 4 CNN models (ResNet50, ResNext50, EfficientNet, and DenseNet121). Likewise, Chen et al. (19) studied 3 CNN models (InceptionV3, VGG16, and ResNet50). While **Wang et al**. (3) evaluated 2 models (VGG19 and Inception-ResNet-v2). The remaining studies followed simple and direct methods by using a single CNN model such as Xception-based in Dolezal et al. study (16), InceptionV3 in **Li et al. and Tsou** studies (18, 20), and VGG19 in Tran et al. study (23).

The choice of model depends on factors like the complexity of the task, available computational resources, and the performance requirements of the analysis. The differences among these CNN models primarily lie in their base architectures, which vary in terms of depth, efficiency, and design philosophy. These base architectures are often pre-trained on large datasets like ImageNet and then fine-tuned or extended with additional layers for specific histopathology image analysis tasks. Recently, a couple of works described a more effective way for training a CNN model entirely from scratch on histopathology images rather than ImageNet (25).

Table 2: Summary of the models' architecture used across the studies with their best performance.

Paper	Patch/tile size	Model Architecture	Output classes	Performance
Deng et al 2023	2500x3200px	ResNet50, ResNext50, EfficientNet, and DenseNet121	PTC, MTC, FTC, Adenomatous goiter, Adenoma, Normal	-AUC (0.822-0.994) -Accuracy (92.18-97.53%) -Specificity (88.52-100%)
Bohland et al. 2021	1916×1053px 2272×2272px	SVC (Feature-based), ResNet101 (Deep learning)	PTC-like, non-PTC-like	-AUC (0.93) on Tharun dataset -Accuracy (89.7% & 89.1%)
Wang et al. 2019	448×448px	VGG-19 CNN, Inception-RestNetv2	Normal, Adenoma, Goiter, PTC, FTC, MTC, ATC	-Accuracy (97.34%) with VGG-19 -(94.42%) with Inception-RestNetv2
Zhu et al 2022	512×512px	U-net, Decision tree	Tumor Segmentation	-AUC (0.98) for common tumors -AUC (0.94) for rare tumors
Li et al. 2020	2392×2392px	ImageNet_InceptionV3	Benign, Uncertain, Malignant	-Accuracy (< 0.8) Patches level -(95.3% - 100 %) Slide level
Chen et al 2020	1024×1024px	InceptionV3, VGG16, ResNet50	Benign, Uncertain, Malignant	-Accuracy (> 0.96) ROI-based
Kim et al. 2023	256×256px	DenseNet161, vgg16, EfficientNet-B4 (Tumor detection), RestNet152 (Feature extraction), RNN (Feature integration), MLP (TERT prediction)	TERT mutant, TERT wild-type	-Accuracy (0.95) -AUC (0.94)
Dolezal et al. 2021	299×299px	Xception-based	BRAF-like, RAS-like	-AUC (0.935)
Tsou & Wu 2019	2048×2048px	Inception-v3	BRAF, RAS	-AUC (0.94)
Tran et al 2022	400×400px	VGG19	Normal, Tumor	-AUC (0.96) HS -AUC (0.93) RGB -AUC (0.94) HS-synthesized RGB
Halicek et al 2020	1040×1392px	Inception-v4, MLP	Tumor detection	-AUC (0.9) for HSI-synthesized human-eye RGB multiplex

4.10 Output classes:

In the 11 reviewed works, it's evident that researchers tailored their classification tasks to match the clinical context and objectives. Both binary classifications focusing on key diagnostic distinctions and multi-class schemes were utilized across studies, according to the specific histopathology image datasets and clinical applications. Most of the studies approached histopathology image classification as a binary task, either for subtyping, mutation prediction, or tumor segmentation. For subtyping, **Bohland** et al. (2) classified thyroid nodules as PTC-like or non-PTC-like. Concerning mutation prediction, **Kim** et al. (17) classified TERT promoter mutation status as positive or negative, while, **Tsou** et al. (18) intended to distinguish between RAS vs. BRAF mutated subtypes. Similarly, **Dolezal** et al. (16) classified thyroid nodules as RAS-like or BRAF-like mutated subtypes. Aiming to tumor segmentation, **Tran** et al., **Halicek** et al. and **Zhu** et al. (21-23), classified patches into tumor vs. normal.

In contrast to binary classes, the remaining studies defined multi-class tasks. **Li** et al. and **Chen** et al. (19, 20) classified images into three categories as benign, malignant, or uncertain. **Wang** et al. (3) defined 7 classes: normal, goiter, FA, PTC, FTC, MTC (medullary thyroid carcinoma), and ATC (anaplastic thyroid carcinoma). Similarly, **Deng** et al. (24) classified thyroid lesions into one of 6 classes (PTC, MTC, FTC, FA, adenomatous goiter, or normal).

Although binary classification is simpler to implement and understand compared to multi-class classification, the choice between binary and multi-class classification is driven by the clinical question and the nature of the data. It's important to consider the clinical relevance and potential consequences of model's predictions. For example, if the diagnostic plan significantly differs between subtypes, a multi-class approach may be more appropriate as multi-class classification allows for finer granularity in predictions. On the other hand, if the clinical decision is ultimately binary, especially in an imbalanced dataset, using a binary-output model can be more straightforward and interpretable.

4.11 Analysis of Results/Evaluation of model performance (Table 2):

The studies consistently demonstrated the potential of DL to extract discriminative features from thyroid histology images, supporting accurate classification and prediction. The studies, in our review, have utilized various evaluation metrics to assess models' performance on image classification tasks. Accuracy and AUC were the most used metrics. Accuracy measures the overall correctness of predictions, while AUC provides a measure of the model's ability to discriminate between classes.

The accuracy across the studies ranged from 77-100%. The performance varied substantially, possibly attributable to differences in study design, and DL methods utilized and due to differences in datasets, sample size, and classification tasks. However, most of the studies reported a high accuracy of 90% or above. The top-performing models consistently achieved high accuracy scores in the range of 80-95%.

Kim et al. (17) evaluated models using precision, recall, F1 score, accuracy, and AUC, and selected the best model based on validation accuracy. They achieved 0.95 accuracy and 0.94 AUC when adopting color transformation technique in images preprocessing stage. **Bohland** et al. (2) using accuracy and AUC metrics, reported on Tharun's dataset (89.7% vs. 89.1% accuracy) and on Nikiforov's dataset (83.5% vs. 77.4% accuracy) with features-based vs. DL methods, respectively. They also compared model performance to expert pathologists. **Dolezal** et al. (16) model achieved 0.935 AUC. **Tsou et Wu** (18) obtained 0.878 AUC on validation and 0.951 AUC on testing data. **Halicek** et al. (22) obtained the best performance (0.9 AUC) for HS-synthesized human-eye RGB multiplex. **Tran** et al. (23) reported AUC for their model (0.96 with HS, 0.93 with RGB, and 0.94 with HS-synthesized RGB). Finally, **Zhu** et al. (21) model achieved 0.98 AUC for common tumors, and 0.94 AUC for rare tumors. **Li et al.** (20) reported two levels of performance assessment, patch-level accuracy (<80% for all models) and slide-level accuracy (95.3% - 100% with rule-based method). **Chen** et al. (19) achieved the best classification accuracy (> 0.96) on ROI-based diagnosis. **Wang** et al. (3) achieved 97.3% and 94.4% accuracy with VGG-19 and Inception-ResNet-v2, respectively. **Deng** et al. (24) achievement comparison among the 4 tested models ranged from 0.822-0.994 AUC, 92.18-97.53% accuracy, and 88.52-100% specificity.

Overall, the analysis of the results of this systematic review offers promising insights into the potential of DL techniques in advancing the diagnosis and management of thyroid nodules. However, further research is needed to address the limitations identified.

5. Limitations with suggested key solutions:

The selected studies demonstrated promising applications of DL for thyroid histopathology image classification. However, some key limitations were observed that restrict the real-world utility and clinical translation of these approaches. Small sample sizes, subjective annotation, and lack of case diversity were common limitations. The works highlight the need for larger standardized datasets, expanded diversity in subclasses, comparison across multiple sites and rigorous evaluation on heterogeneous real-world data.

5.1 Small, imbalanced dataset:

One fundamental challenge is the small, imbalanced datasets used in most studies, which limits model performance and generalizability(16). Furthermore, **diversity** is limited given single-center data (17, 23). DL prerequisites of large, labeled datasets and computational resources may hinder its translation to clinical practice. Expanding datasets through extensive labeling efforts with increased examples per class will be critical moving forward.

5.2 No quality control for data preparation

Objective quality control of data extraction and preprocessing programs is required to limit potential biases (21). Direct comparison of pathologists with versus without assistance of models using reader studies would better demonstrate clinical utility and guide development. There is also a lack of head-to-head benchmarking between techniques on shared public datasets (2).

5.3 Bias in ROI annotation and patch extraction:

Additionally, potential subjectivity and selection bias can arise during tumor/region annotation by pathologists, which may not fully capture the diversity of real-world samples. Also, subjective patch sampling may introduce selection bias. Developing standardized annotation protocols could help mitigate this (17). Most studies focused narrowly on classification tasks, while localization, segmentation, and other problem formulations warrant deeper investigation.

5.4 lacking details for model deployments

As an emerging field, translating these approaches into widespread clinical practice remains challenging. Crucially, clinical implementation remains overlooked, with no clear strategies for model deployment and workflow integration. Conducting usability studies is advised before deployment.

5.5 Model optimization & interpretability

Regarding model development, the field is still in early stages, with limited model architectures and hyperparameters evaluated so far. Broad optimization of network architectures and hyperparameters through extensive experiments could boost performance. Another key aspect is model interpretability, which remains underexplored but critical for clinical acceptance. Implementing methods to explain model reasoning and validate learned features is beneficial.

5.6 Lacking clinical and radiological correlation.

Another issue is the sole reliance on histopathology images, lacking complementary clinical, genomic, and radiographic data that pathologists use to reach diagnostic decisions. Integrating these multiple modalities represents an important frontier. Enabling multi-modal integrations represent fertile directions for future research to unlock the full potential of DL in thyroid histopathology.

6. Future Directions:

The findings reveal the potential of DL approaches in enhancing the accuracy of thyroid neoplasms classification. Integrating DL methods into daily practice looks optimistic in enhancing thyroid neoplasms diagnosis. The high accuracy observed in these studies suggests their potential utility in aiding clinical decision-making. Initial results, though limited in scale, strongly demonstrate the capability to encode morphological patterns in an automated high-throughput manner. This could enhance pathologist productivity and improve patient care through quantitative augmented decision support.

Looking ahead, larger multi-center studies, evaluation on heterogeneous real-world data, and pragmatic clinical integration will be crucial to fully translate these emerging technologies into widespread practice. However, the early successes establish the way for larger initiatives to unite pathologists, data scientists and clinicians in realizing the next generation of AI-powered computational thyroid pathology.

7. Conclusion:

To conclude, this systematic review offers comprehensive insights into recent advancements in DL approaches for the classification of thyroid neoplasms based on histopathology images. The findings underscore the potential of these techniques to enhance accuracy and improve clinical decision-making. While further optimization is warranted, integration of these tools to guide pathologists could markedly improve thyroid cancer screening, diagnosis, prognosis, and treatment selection for better patient care.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT and google Bard in order to improve language and readability. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Kholova I, Haaga E, Ludvik J, Kalfert D, Ludvikova M. Noninvasive Follicular Thyroid Neoplasm with Papillary-like Nuclear Features (NIFTP): Tumour Entity with a Short History. A Review on Challenges in Our Microscopes, Molecular and Ultrasonographic Profile. Diagnostics (Basel). 2022;12(2).
2. Bohland M, Tharun L, Scherr T, Mikut R, Hagenmeyer V, Thompson LDR, et al. Machine learning methods for automated classification of tumors with papillary thyroid carcinoma-like nuclei: A quantitative analysis. PLoS One. 2021;16(9):e0257635.
3. Wang Y, Guan Q, Lao I, Wang L, Wu Y, Li D, et al. Using deep convolutional neural networks for multi-classification of thyroid tumor by histopathology: a large-scale pilot study. Ann Transl Med. 2019;7(18):468.
4. Coelho BP, de Oliveira Valentim F, Miot HA, Jaune DTA, Hayashi CY, de Oliveira CC, et al. Follicular Lesions with Papillary Nuclear Characteristics: Differences in Chromatin Detected by Computerized Image Analysis. Arch Endocrinol Metab. 2021;64(5):630-5.
5. Ozolek JA, Tosun AB, Wang W, Chen C, Kolouri S, Basu S, et al. Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning. Medical image analysis. 2014;18(5):772-80.
6. Valentim FO, Coelho BP, Miot HA, Hayashi CY, Jaune DTA, Oliveira CC, et al. Follicular thyroid lesions: is there a discriminatory potential in the computerized nuclear analysis? Endocr Connect. 2018;7(8):907-13.
7. Ludwig M, Ludwig B, Mikuła A, Biernat S, Rudnicki J, Kaliszewski K. The Use of Artificial Intelligence in the Diagnosis and Classification of Thyroid Nodules: An Update. Cancers. 2023;15(3).
8. Pouliakis A, Karakitsou E, Margari N, Bountris P, Haritou M, Panayiotides J, et al. Artificial Neural Networks as Decision Support Tools in Cytopathology: Past, Present, and Future. Biomedical engineering and computational biology. 2016;7:1-18.
9. Thakur N, Alam MR, Abdul-Ghafar J, Chong Y. Recent Application of Artificial Intelligence in Non-Gynecological Cancer Cytopathology: A Systematic Review. Cancers. 2022;14(14).
10. Wong CM, Kezlarian BE, Lin O. Current status of machine learning in thyroid cytopathology. Journal of pathology informatics. 2023;14:100309.
11. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

12. Anand D, Yashashwi K, Kumar N, Rane S, Gann PH, Sethi A. Weakly supervised learning on unannotated H&E-stained slides predicts BRAF mutation in thyroid cancer with high accuracy. The Journal of pathology. 2021;255(3):232-42.
13. Esce AR, Redemann JP, Sanchez AC, Olson GT, Hanson JA, Agarwal S, et al. Predicting nodal metastases in papillary thyroid carcinoma using artificial intelligence. American journal of surgery. 2021;222(5):952-8.
14. Esce A, Redemann JP, Olson GT, Hanson JA, Agarwal S, Yenwongfai L, et al. Lymph Node Metastases in Papillary Thyroid Carcinoma can be Predicted by a Convolutional Neural Network: a Multi-Institution Study. The Annals of otology, rhinology, and laryngology. 2023:34894231158464.
15. Nojima S, Kadoi T, Suzuki A, Kato C, Ishida S, Kido K, et al. Deep Learning-Based Differential Diagnosis of Follicular Thyroid Tumors Using Histopathological Images. Mod Pathol. 2023;36(11):100296.
16. Dolezal JM, Trzcinska A, Liao CY, Kochanny S, Blair E, Agrawal N, et al. Deep learning prediction of BRAF-RAS gene expression signature identifies noninvasive follicular thyroid neoplasms with papillary-like nuclear features. Mod Pathol. 2021;34(5):862-74.
17. Kim J, Ko S, Kim M, Park NJ, Han H, Cho J, et al. Deep Learning Prediction of TERT Promoter Mutation Status in Thyroid Cancer Using Histologic Images. Medicina (Kaunas). 2023;59(3).
18. Tsou P, Wu CJ. Mapping Driver Mutations to Histopathological Subtypes in Papillary Thyroid Carcinoma: Applying a Deep Convolutional Neural Network. Journal of clinical medicine. 2019;8(10).
19. Chen P, Shi X, Liang Y, Li Y, Yang L, Gader PD. Interactive thyroid whole slide image diagnostic system using deep representation. Computer methods and programs in biomedicine. 2020;195:105630.
20. Li Y, Chen P, Li Z, Su H, Yang L, Zhong D. Rule-based automatic diagnosis of thyroid nodules from intraoperative frozen sections using deep learning. Artif Intell Med. 2020;108:101918.
21. Zhu X, Chen C, Guo Q, Ma J, Sun F, Lu H. Deep Learning-Based Recognition of Different Thyroid Cancer Categories Using Whole Frozen-Slide Images. Front Bioeng Biotechnol. 2022;10:857377.
22. Halicek M, Dormer JD, Little JV, Chen AY, Fei B. Tumor detection of the thyroid and salivary glands using hyperspectral imaging and deep learning. Biomed Opt Express. 2020;11(3):1383-400.
23. Tran MH, Ma L, Litter JV, Chen AY, Fei B. Thyroid Carcinoma Detection on Whole Histologic Slides Using Hyperspectral Imaging and Deep Learning. Proc SPIE Int Soc Opt Eng. 2022;12039.
24. Deng C, Li D, Feng M, Han D, Huang Q. The value of deep neural networks in the pathological classification of thyroid tumors. Diagnostic pathology. 2023;18(1):95.
25. Riasatian A, Babaie M, Maleki D, Kalra S, Valipour M, Hemati S, et al. Fine-Tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. Medical image analysis. 2021;70:102032.