ARTICLE OPEN

Check for updates

Translational Therapeutics

Machine learning-based prediction of luminal breast cancer subtypes using polarised light microscopy

Kseniia Tumanova[1]✉, Mohammadali Khorasani[2], Sharon Nofech-Mozes[3] and Alex Vitkin[1,4]

BACKGROUND: Routine histopathology cannot distinguish between clinically diverse luminal A and B breast cancer subtypes (LBCS), often requiring ancillary testing. Mueller matrix polarimetry (MMP) offers a promising approach by analysing polarised light interactions with complex breast tissues. This study explores the efficacy of using MMP for luminal subtype differentiation.
METHODS: We analysed 30 polarimetric and 7 clinical parameters from 116 unstained breast core biopsies, LBCS classified using the BluePrint® molecular assay. These features were used to train various machine learning models: logistic regression, linear discriminant analysis, support vector machine, random forest, and XGBoost to distinguish luminal subtypes. Receiver operating characteristic curve (ROC) analysis was used to each to assess diagnostic performance using area under the curve, accuracy, sensitivity, and specificity.
RESULTS: Using the top six most prognostic polarimetric (three) and clinical (three) biomarkers ranked by feature importance, the best-performing random forest model achieved an accuracy of 81% (area under ROC = 86%), with both sensitivity and specificity at 75% on an unseen test set, indicating moderately promising, clinically informative performance.
CONCLUSIONS: MMP, particularly its selected Mueller matrix elements, combined with clinical biomarkers show promise in distinguishing LBCS as validated against BluePrint®. By detecting subtle differences in tissue morphology, this approach may enhance breast cancer prognosis and help guide treatment decisions.

British Journal of Cancer (2025) 133:1916–1925; https://doi.org/10.1038/s41416-025-03150-x

INTRODUCTION

Breast cancer continues to be a significant cause of cancer-related morbidity and mortality worldwide [1–3]. Breast carcinoma is a biologically diverse disease with variable prognosis and response to systemic therapies, classified into four intrinsic distinct molecular subtypes by gene expression profiling [4, 5]. Among these, luminal breast cancers—characterised by oestrogen receptor (ER) positivity and human epidermal growth factor receptor 2 (HER2) negativity—constitute a major category [4, 6]. These cancers are further divided into luminal A and luminal B subtypes, each with unique prognosis and therapeutic implications [4, 6, 7]. In the last two decades (since the emergence of the molecular classification), the definition of luminal subgroup by surrogate immunohistochemical markers has shifted; currently accepted definitions are that luminal carcinomas characteristically express oestrogen receptor (ER) with variable cell proliferations [4, 8]. HER2 overexpression is the hallmark of ERBB2-overexpressing tumours that also lack ER and progesterone receptor (PR) expression [4, 9]. Basal-like carcinoma fails to express ER, PR, or HER2 [4, 5, 9].

Luminal A tumours are typically ER-positive, PR-positive, low in Ki-67 proliferation index, and have a better prognosis with less aggressive clinical behaviour [5, 6, 9]. In contrast, luminal B tumours, although also ER-positive, tend to have a higher Ki-67 index, lower PR expression, higher histologic grade and are often more aggressive with a poorer prognosis compared to luminal A tumours [5, 6, 10]. The differentiation between luminal A and B subtypes is crucial for tailoring treatment strategies [6, 11]. Luminal A patients often respond well to hormonal therapies, while luminal B patients may require additional cytotoxic chemotherapy due to their more aggressive nature [5, 6, 11]. Despite the importance of this distinction, traditional histopathological methods frequently fall short, as both subtypes can appear remarkably similar on hematoxylin and eosin (H&E) stained sections (Fig. 1) [8]. That said, molecular tools based on commercially available multigene assays have been most quite successful in predicting tumour biologic behaviours and therapeutic responses in early-stage luminal breast carcinoma patients, and their use was adopted by National Comprehensive Cancer Network [6, 12, 13].

One of these assays is the BluePrint® and MammaPrint® tandem based on 80- and 70-gene sets, respectively [12–15]. The former is designed to identify the molecular subtype by analysing gene

[1]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada. [2]Department of Surgery, University of British Columbia, Victoria, BC, Canada. [3]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada. [4]Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada.
✉email: k.tumanova@mail.utoronto.ca

Received: 18 December 2024 Revised: 14 July 2025 Accepted: 4 August 2025
Published online: 3 October 2025

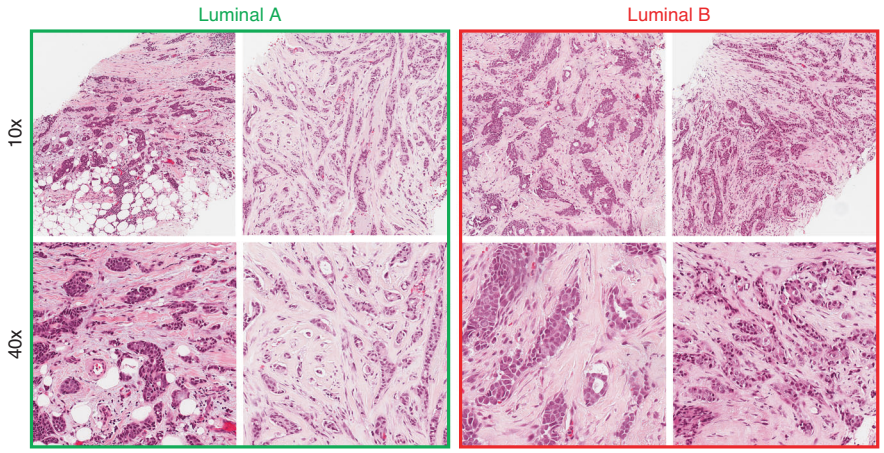

Fig. 1 Representative hematoxylin and eosin (H&E) stained sections of luminal A and luminal B breast cancer biopsies at 10× (top row) and 40× magnification (bottom row). Luminal A sections (first two columns, green border) and luminal B sections (last two columns, red border) display intermediate-grade invasive ductal carcinoma (no special type). Note that tumour architecture and nuclear features are similar across subtypes, underscoring the difficulty of A vs B differentiation. Hormone receptor status: ER+/PR+/HER2−.

expression patterns involved in signalling pathways that promote cancer growth, providing insights into the tumour biology of luminal breast cancers [12, 13, 15]. MammaPrint® evaluates the risk of cancer recurrence, specifically the risk of metastasis by analysing the expression of 70 genes and predicts response to adjuvant and neoadjuvant chemotherapy beyond histopathology and immunohistochemical surrogates [12–15]. Despite their advanced capabilities, these methods come with significant drawbacks. They are expensive, involve time delays, require specialised infrastructure, and may not be accessible in all clinical settings, particularly in resource-limited regions [16, 17]. These challenges highlight the need for alternative affordable diagnostic approaches that are both effective and accessible.

Mueller matrix polarimetry (MMP) presents a promising approach for assessing the structural properties of tissues through polarised light measurement [18–22]. Studies have demonstrated its potential utility in oncology, where it has been used to correlate polarimetric parameters with clinical outcomes such as recurrence [23, 24], survival rates [25], and tumour-stroma interactions, particularly in collagen-rich tumour microenvironments [23, 26]. In breast cancer specifically, MMP has shown potential for distinguishing healthy, benign, and malignant tissues and for providing quantitative insights into tissue microstructures that are key to cancer diagnosis and progression [22, 27]. For instance, Pham et al. applied Mueller matrix imaging to human breast tissue, using MM elements such as M23, M32, and M44 to distinguish tissue types, with M23 showing higher intensity in healthy tissues while M32 and M44 exhibited reduced intensity in malignant samples [27]. Similarly, Xia et al. classified breast cancer cells based on receptor protein expression with 88% accuracy by analysing polarimetric data using convolutional neural networks and demonstrated that MMP can visualise intracellular fibrous actin, a critical factor in tumourigenesis and metastasis [28]. Dong et al. further highlighted the utility of MMP in tracking breast ductal carcinoma progression by analysing polar decomposition parameters and texture features, providing quantitative evidence for monitoring disease stages and treatment outcomes [29]. Together, these findings underscore the versatility of MMP as a promising powerful tool for breast cancer assessment, offering quantitative insights that complement existing diagnostic methods by distinguishing healthy from cancerous tissues and aiding in disease progression analysis.

The present study thus aims to evaluate whether MMP-derived polarimetric parameters alone, or in combination with clinical biomarkers, can aid in distinguishing luminal A from luminal B breast cancer subtypes at the diagnostic stage. By integrating supervised machine learning (ML), we seek to optimise feature selection and improve classification performance. To test this, 116 unstained human pre-operative breast cancer core biopsy samples were analysed using MMP, with findings correlated against genomic classifications provided by BluePrint® and MammaPrint®. To ensure the polarimetric results reflect the full structural complexity of the tissue, *the entire biopsy slide was analysed* without user-dependent input from pathology for region of interest (ROI) selection, allowing for a comprehensive, independent and unbiased objective evaluation of the full core biopsy tissue sample.

METHODS

Human breast cancer samples

Institutional ethics approval was obtained from the participating hospital institutions, University Health Network and Sunnybrook Hospital, both located in Toronto, Canada. The requirement for patients' consent to use the breast cancer biopsy samples was waived by the ethics board due to the retrospective nature of the study and the anonymization of patient information.

Consecutive (sequentially identified without selection bias) ER+/HER2− invasive breast carcinoma diagnosed on core biopsies were identified from the institutional pathology database (N = 116 patients). Each sample contained between 1 to 3 tissue cores, resulting in varying size of imaged tissue areas. The luminal A cohort included 66 patients, while the luminal B cohort comprised 50 patients. The subset of patients who had clinical data useful for ML analysis was smaller ($N_c = 68$ patients), with 42 classified as luminal A and 26 as luminal B. The investigated clinical parameters were: Nottingham grade scores obtained from both the diagnostic biopsy and post-surgical resection specimens, primary tumour size, patient age at diagnosis, histologic subtype, number of lymph nodes with confirmed metastases, and number of sentinel ('hot') lymph nodes identified as suspicious and subsequently examined for metastatic involvement. Luminal A vs B classification of these samples was performed using genomic profiling (BluePrint® and MammaPrint®). Unstained 4.5 µm thick sections were prepared on charged microscope slides from FFPE (formalin-fixed, paraffin-embedded) tissue blocks that were used for the molecular assay. Minimal sample preparation involved chemical dewaxing to avoid potential polarisation imaging artifacts [25, 30–32]. No further processing was required for polarimetric imaging.

Polarimetric image acquisition and feature extraction

Mueller matrix (MM) imaging was utilised to capture the structural characteristics of unstained breast cancer biopsy samples, providing comprehensive polarisation data across the entire tissue section [23, 33].

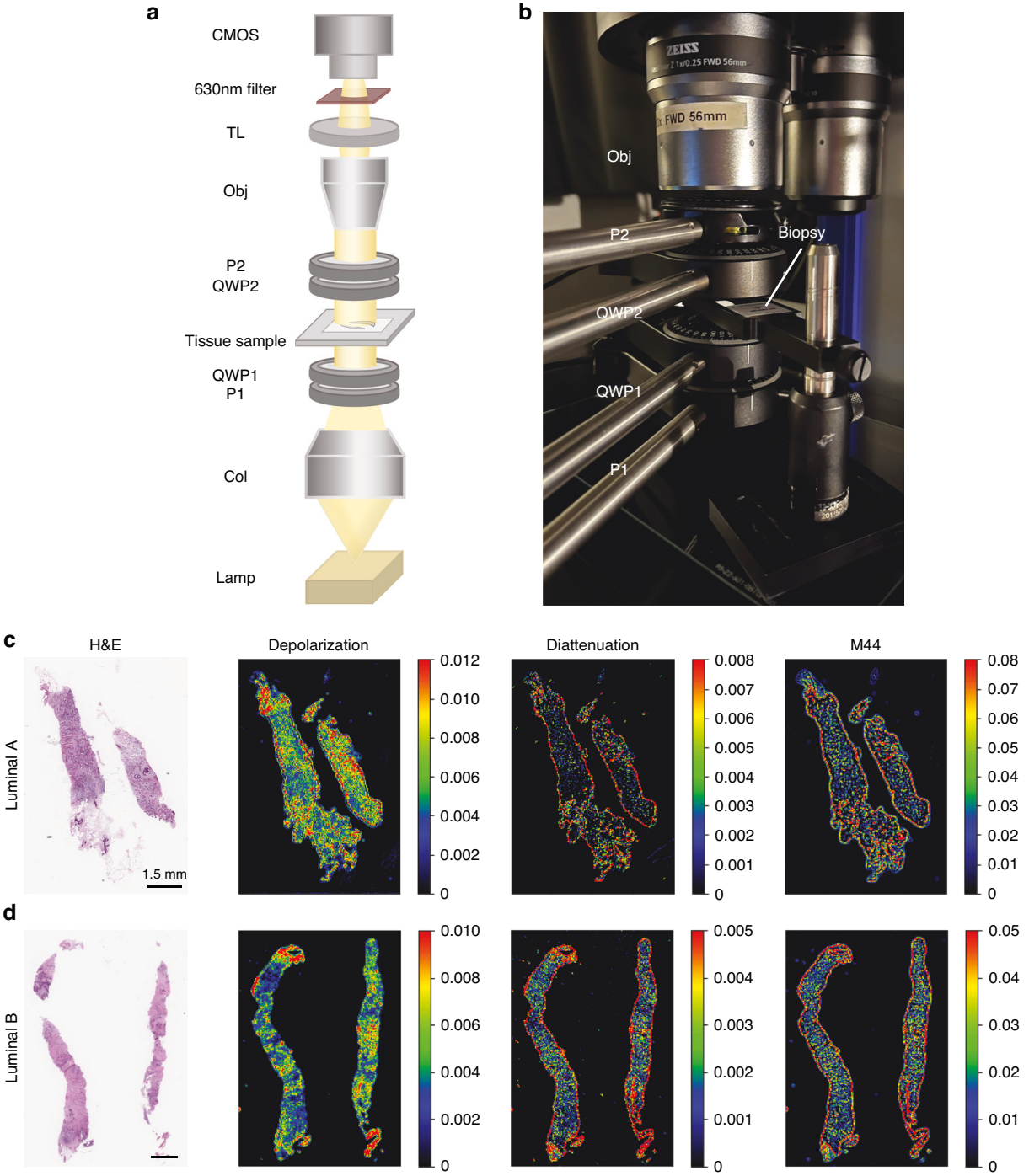

Fig. 2 Mueller matrix microscopy system and representative polarimetric images of breast cancer biopsies. a Schematic of the Mueller matrix microscopy setup. Light from the source (Lamp) passes through a collimator (Col), a linear polarizer (P1), and a quarter-wave plate (QWP1) to form the polarisation state generator (PSG). After interacting with the sample, transmitted light passes through the polarisation state analyser (PSA; QWP2 and P2), a microscope objective (Obj), a tube lens (TL), a 630 nm filter, and is captured by a CMOS camera. **b** Photograph of the Mueller matrix microscopy system. **c**, **d** Representative polarimetric images of human breast core biopsy samples, showing H&E staining, depolarisation, diattenuation, and the M44 Mueller matrix element for luminal A (c) and luminal B (d) cases. Each image includes a 1.5 mm scale bar. Visual differentiation using these representative polarisation images remains challenging due to tissue heterogeneity, motivating the application of supervised machine learning; see text for details.

Importantly, this method ensured full slide coverage without the need for pathologist-assisted (and thus somewhat subjective) ROI selection, allowing for an unbiased assessment of the tissue's polarisation properties. For each biopsy, 24 polarimetric images were acquired using different configurations of the polarisation state generator (PSG) and polarisation state analyser (PSA), following the approach outlined by Tumanova et al. [23]. This image acquisition methodology increased the robustness and signal-to-noise ratio (SNR) of the polarimetric measurements [23, 33]. The experimental setup is illustrated in Fig. 2a, b.

The polarimetric system consisted of a Mueller matrix polarimetry module integrated into a standard stereo zoom microscope (Axio Zoom V16, Zeiss) equipped with a Plan Neofluar Z 1X/0.25 NA objective lens. The

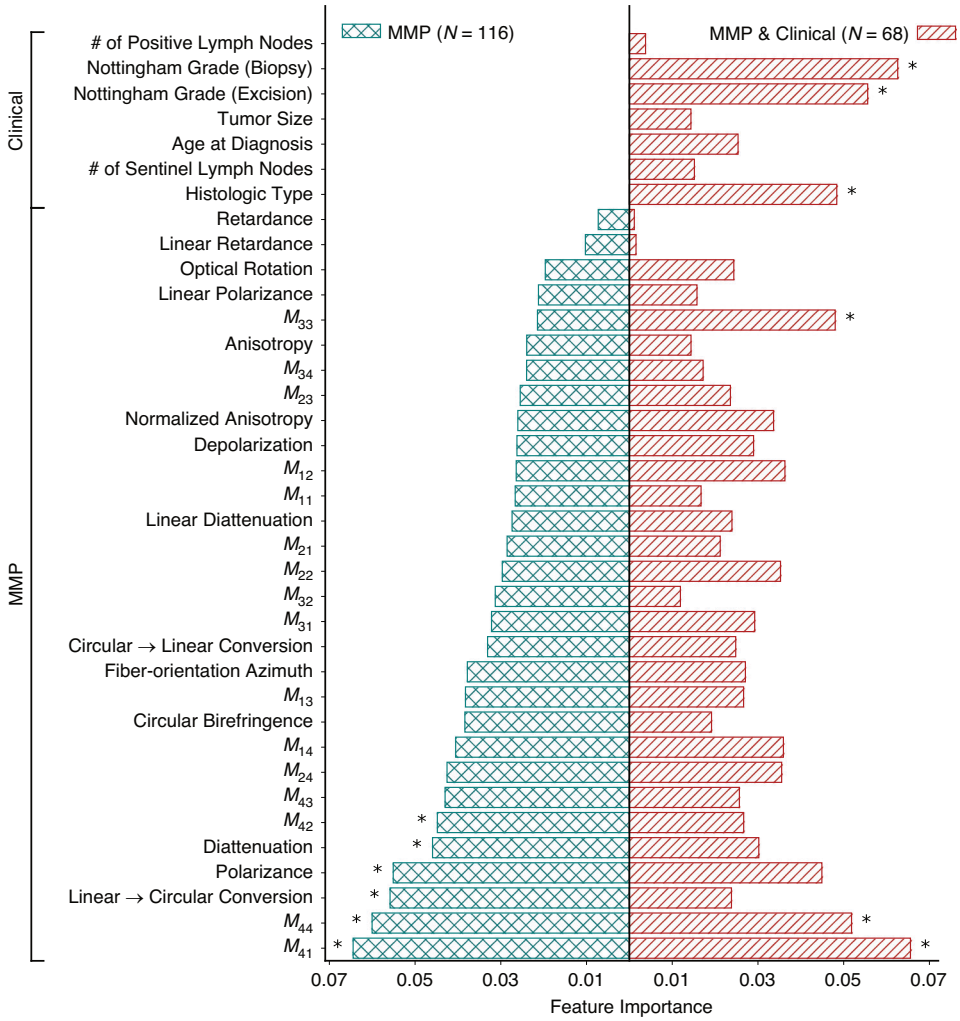

Fig. 3 Relative feature importance plot obtained using the random forest classifier. Features are Mueller Matrix polarisation (MMP) parameters and clinical variables as listed on the y-axis with their respective importance scores on the x-axis. Asterisks indicate the top six features contributing most to the classification models.

PSG, which included a rotatable linear polarizer (LPVISE100-A, Thorlabs) and a rotatable quarter-wave plate (AQWP05M-600, Thorlabs), modulated the incident light polarisation, while the PSA, composed of the same elements but in reverse order, analysed the polarisation state of the transmitted light after interaction with the tissue. The microscope's LED light source [Illuminator HXP 200C (D), Zeiss] emitted 310 W of uncollimated white light, which was passed through a filter centred at 630 nm (ET630/75 or ZET630/10, Chroma). Images were acquired using a Hamamatsu ORCA-Flash4.0 V3 Digital CMOS camera with a 2048×2048 pixel array and a pixel size of $6.5 \times 6.5\ \mu m^2$. At a magnification of 7×, this setup provided a field of view of $18.9 \times 18.9\ mm^2$ and an optical lateral resolution of 13.9 µm. The resolution was validated using a 1951 USAF resolution target (R3L3S1N, Thorlabs) [33]. To calibrate the system, we measured the input Stokes vectors at each pixel using a blank (air) sample, and subsequently calculated the output Stokes vectors with the reference samples (air and a retarder with known optical properties) in place [33]. This pixel-by-pixel calibration approach corrects for spatially dependent distortions arising from off-axis rays and ensures high spatial fidelity [33–36]. The system acquires 24 intensity images for different PSG/PSA configurations, enabling direct calculation of the Mueller matrix via Stokes vector formalism. While this exceeds the minimum 16-image requirement, the 24-image acquisition improves signal-to-noise ratio, enhancing overall data quality [31, 33–36]. In this study, calibration accuracy was considered acceptable when the difference between measured and theoretical Mueller matrices fell within a ~1–5% range.

Once the polarimetric images were acquired for the breast biopsy samples, pixel-by-pixel calculations of the Mueller matrix were performed using the Stokes formalism, as previously detailed [23]. From these matrices, key polarimetric parameters—depolarisation, retardance, and diattenuation—were extracted using the Lu–Chipman MM polar decomposition (MMPD), providing insights into tissue anisotropy and depolarisation [37]. In parallel, additional polarimetric parameters were derived using the Mueller matrix transformation (MMT), which complements the MMPD by characterising scattering and other structural tissue properties [38]. Both MMPD and MMT approaches are widely used in the bio-polarimetry research for detailed morphological analysis of biological tissues [18–21, 23]. The median value of each parameter was used as the representative metric for each slide, ensuring standardisation across samples regardless of the number of tissue cores present and the size of the surface area analysed from each patient [39, 40]. For consistency, all polarimetric parameters listed in the following sections thus refer to their median values across all pixels within each slide, unless otherwise specified.

The various derived polarimetric parameters offer biophysical insights into tissue microstructure, with a particular emphasis on the organisation and characteristics of birefringent structures, such as collagen [18–21]. For example, using the Lu–Chipman decomposition, parameters such as depolarisation (Δ), retardance (R), and diattenuation (D) were calculated, providing information about tissue scattering, fibre-related birefringence, and absorption anisotropy, respectively [18–21, 23, 37]. Additionally, the MMT parameters were employed to quantify the anisotropic properties of the tissue, including fibre alignment and birefringence [19, 20, 23, 38]. The 30 resultant polarimetric parameters were derived following the methodology outlined in previous studies [18–24], comprising 16 Mueller matrix elements, 8 MMPD, and 6 MMT metrics. A comprehensive list of these parameters is provided in Fig. 3.

Machine learning methods

The classification task at hand necessitates a supervised ML approach, given the categorical nature of the problem (luminal A vs B). Feature selection through relative feature importance was undertaken as a preprocessing step to identify and select the most relevant features for improving model performance and interpretability. Then several established ML algorithms were considered, each offering distinct methodological advantages and trade-offs [41]. In this study, we systematically evaluated logistic regression (LR), linear discriminant analysis (LDA), support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost) to determine the optimal classifier for distinguishing between luminal breast cancer subtypes.

LR is a widely used method for binary classification, modelling the probability of an outcome using the logit function, which relates the log-odds of the outcome to a linear combination of the input features [41]. LDA, an extension of this approach, seeks to maximise class separability by identifying an optimal linear combination of features while assuming normally distributed data with equal class covariances [41]. SVM, a powerful kernel-based method, constructs a hyperplane that optimally partitions classes in a high-dimensional feature space, effectively handling complex decision boundaries [41]. RF, an ensemble-based approach, aggregates multiple decision trees to enhance predictive stability and mitigate overfitting, leveraging feature randomness to improve generalisability [41]. XGBoost, an advanced gradient boosting framework, iteratively refines decision trees by minimising classification errors through gradient descent, often achieving superior performance in structured biomedical datasets [42, 43].

Each method contains specific assumptions and design characteristics with varying applicability to the structure of the dataset used in this study. LR and LDA are computationally efficient and interpretable, though their assumptions may be limiting in datasets with non-linear boundaries or mixed feature distributions [41]. In particular, the assumption of equal covariances in LDA may not hold in the presence of heterogeneous clinical and polarimetric variables [41]. SVM does not assume any specific feature distribution and is flexible in modelling non-linear patterns, although its performance can depend on kernel selection and hyperparameter tuning [41]. Tree-based models such as RF and XGBoost are non-parametric, do not require feature normalisation, and are robust to multicollinearity and outliers. These properties make them applicable in settings with modest sample sizes, categorical outcomes, and potentially correlated features. However, these models are more complex and can be less transparent in their decision-making processes [41–43].

Handling outliers

Outliers were retained in the analysis, as no discernible clinical or biological rationale (e.g. age, ER/PR expression, lymph node status) justified their exclusion. While outlier removal is often employed to optimise machine learning model performance by refining decision boundaries and mitigating noise, its indiscriminate application risks eliminating biologically meaningful variability inherent in real-world clinical datasets [44]. Retaining these data ensures that the model is trained on a representative distribution, capturing the full spectrum of patient heterogeneity rather than overfitting to an artificially constrained dataset. Furthermore, excluding such cases may inadvertently bias the model by omitting rare but clinically significant phenotypes, ultimately compromising its generalisability and translational utility [44, 45].

Dataset partitioning and model training

Given the relatively small dataset, a 70% (train+validation)/30% (test) split ratio was implemented to optimise model learning capacity and generalisation. This partitioning strategy aims to train the model on a sufficiently large dataset while preserving an independent (not seen) test set for robust performance evaluation. To prevent class imbalance from skewing the model, stratified sampling was applied during data splitting, ensuring that the proportion of luminal A and luminal B cases remained consistent across all subsets [46].

To enhance model robustness and reduce the risk of overfitting, feature selection was performed as a preprocessing step using relative feature importance scores obtained from the random forest model [47]. RF was chosen for this task due to its ability to handle non-linear relationships, accommodate mixed feature types, and estimate variable importance directly during model training [41, 47]. This method ranks features based on how much they contribute to improving classification performance across the ensemble of decision trees. While alternative approaches to feature selection may yield somewhat different rankings, RF was selected to reflect the structure of our dataset, which includes potentially interacting polarimetric and clinical features [41, 47]. Subsequently, a five-fold cross-validation strategy (five different training-validation set combinations) was employed, whereby the 70% of the dataset was repeatedly partitioned into training and validation folds. Stratified sampling ensured class balance within each fold, and each sample contributed to both training and validation phases. This approach helps mitigate selection bias and provides a more reliable estimate of model performance under limited sample size conditions [48, 49].

Before final model testing, a retraining step was also conducted, where the best-performing model—selected based on training and validation results—was retrained on the combined training and validation set before being evaluated on the test (unseen data) set. This approach is widely used in machine learning, particularly in clinical datasets with limited sample sizes [50]. By incorporating additional (validation) data into the training phase, this method enhances statistical power and improves model generalisability without introducing data leakage, towards a more reliable estimate of real-world performance [50, 51].

Performance evaluation

To rigorously quantify classification performance on unseen data, a 2×2 confusion matrix (true subtype vs predicted subtype) was constructed. Standard classification metrics, including sensitivity, specificity, accuracy, and the area under the receiver operating curve (AUROC), were computed to provide a comprehensive evaluation of model effectiveness.

RESULTS

In this study, polarimetric parameters were analysed to characterise the interaction of polarised light with breast cancer biopsy tissues, aiming to differentiate between luminal A and B subtypes. A total of 30 parameters were examined, including the 16 Mueller matrix elements and additional parameters derived from MMPD and MMT. For demonstration, Fig. 2c, d presents parametric images based on potentially important metrics of depolarisation, diattenuation, and the M44 Mueller matrix element. The depolarisation metric reflects cellularity and microstructural heterogeneity, often associated with complex tissue architectures [20, 21, 23]. Diattenuation provides insights into tissue anisotropy, indicating the directional dependence of optical properties [19, 20, 23]. The biophysical meaning of the M44 element is more difficult to define explicitly, as the direct physical interpretation of any individual Mueller matrix element (except M11) is inherently ambiguous [18, 19, 52–54]. But overall, as seen in Fig. 2c, d, while polarimetric imaging of unstained breast biopsy slides provides some interesting contrast, visual differentiation of luminal subtypes remains challenging, necessitating quantitative analysis to reveal subtle differences essential for subtype classification.

To assess whether polarimetric features alone can differentiate luminal A and B subtypes, or whether incorporating clinical variables offers additional value, all extracted features were ranked by importance using an RF classifier. As shown in Fig. 3, models trained solely on 30 MMP parameters (green bars, left) were compared to those trained on an extended set of 37 features that included both polarimetric *and* clinical variables (red bars, right). For subsequent model development, the top six features (marked by asterisks in Fig. 3) were selected as the feature selection preprocessing step. Interestingly, when examining the combined feature set, the six highest-ranking features were evenly split between MMP-derived parameters (median values of M41, M44, and M33) and clinical variables (Nottingham grade from biopsy, Nottingham grade from excision, and histologic type), suggesting that both feature categories contribute relevant and non-redundant information to the A vs B classification task.

To evaluate classification performance, five supervised learning algorithms—LR, LDA, SVM, RF, and XGBoost—were trained and validated using five-fold cross-validation. Model performance was assessed via its accuracy, sensitivity, specificity, and AUROC. The

Table 1. Classification performance of five supervised learning algorithms for distinguishing luminal A and luminal B breast cancer subtypes.

Training + Validation						Test			
Model	**LR**	**LDA**	**SVM**	**RF**	**XGBoost**	**RF**	**$RF_{retrained}$**	**SVM**	**$SVM_{retrained}$**
Polarimetric features									
Accuracy	62 ± 10	62 ± 11	67 ± 10	68 ± 12	61 ± 13	60	60	51	51
Sensitivity	54 ± 26	40 ± 26	57 ± 13	57 ± 18	49 ± 19	60	60	33	33
Specificity	67 ± 17	78 ± 12	74 ± 20	76 ± 21	70 ± 13	60	60	65	65
AUROC	60 ± 15	62 ± 14	65 ± 11	69 ± 20	65 ± 13	72	72	62	62
Polarimetric and clinical features									
Accuracy	66 ± 14	68 ± 7	70 ± 9	79 ± 13	71 ± 16	62	81	–	–
Sensitivity	48 ± 27	53 ± 24	33 ± 9	67 ± 9	57 ± 8	38	75	–	–
Specificity	75 ± 15	76 ± 8	93 ± 13	86 ± 19	79 ± 21	50	75	–	–
AUROC	75 ± 10	70 ± 15	78 ± 12	77 ± 18	68 ± 18	74	86	–	–

Cross-validation results (left block) are reported for two input feature sets: polarimetric features alone (N = 116, top rows; see Fig. 4a) and a combined set of polarimetric + clinical features (N_c = 68, bottom rows; see Fig. 5a). Each entry is the mean ± standard deviation across folds for accuracy, sensitivity, specificity, and AUROC. The right block lists the corresponding metrics on the independent test sets: models trained on polarimetric features were evaluated on n = 35 cases, whereas the RF model with polarimetric + clinical features was evaluated on n = 21 cases. '$RF_{retrained}$' and '$SVM_{retrained}$' denote the RF and SVM refit on the full training + validation data before testing.

results are summarised in Table 1 and visualised in Figs. 4 and 5, with error bars representing performance variability across validation folds.

When trained using only polarimetric features, the five supervised learning algorithms exhibited broadly similar overall performance, though they differed in some key classification metrics (Fig. 4a; Table 1, top left section). LR and LDA both achieved a mean accuracy of 62% but diverged in class-specific performance. LDA yielded the highest specificity (78%) while showing limited sensitivity (40%), suggesting a strong bias toward correctly identifying luminal A cases but poor detection of luminal B. LR achieved slightly more balanced performance (sensitivity 54%, specificity 67%), though with reduced overall discriminative power.

SVM produced a higher mean accuracy (67%) and the highest sensitivity (57%) among all models, indicating improved detection of luminal B cases. However, this came with somewhat reduced specificity (74%) and greater variability across cross-validation folds, highlighting a trade-off between recall of the minority class and model stability. XGBoost performed comparably in terms of accuracy (61%), with balanced but moderate sensitivity (49%) and specificity (70%), though its higher variance across folds indicated reduced reliability.

RF demonstrated the most balanced and consistent performance across all metrics, achieving a mean accuracy of 68%, sensitivity of 57%, specificity of 76%, and AUROC of 69%, with reasonably low inter-fold variability. While no single model outperformed across all evaluation criteria, RF and SVM were selected for subsequent evaluation on the independent test set: RF for its balanced metric profile, and SVM because its accuracy, sensitivity, specificity, and AUROC were comparable to those of RF.

Then the generalisability of the two top-performing models—RF and SVM—was evaluated on an independent test set of unseen patient data (n = 35). As shown in Fig. 4b, RF achieved an AUROC of 0.72, outperforming SVM, which reached an AUROC of 0.62. RF also demonstrated more balanced classification performance, with 60% accuracy, 60% sensitivity, and 60% specificity (Fig. 4c, Table 1). In contrast, SVM showed reduced generalisation, particularly in sensitivity (33%), despite slightly higher specificity (65%), resulting in an overall lower accuracy of 51% (Fig. 4d, Table 1). Retraining the RF and SVM models on the combined training + validation set prior to testing left all test metrics unchanged, indicating that predictive limits likely stem from the polarimetric feature set rather than training sample size. Consistent with cross-validation trends, RF still outperformed SVM on accuracy, sensitivity, specificity, and AUROC. Together, these findings suggest that while polarimetric features capture microstructural information, may be insufficient on their own for reliable subtype classification, suggesting the potential benefit of incorporating clinical data to improve model robustness.

Given the moderate performance of all models using polarimetric features alone, we next evaluated whether incorporating clinical variables could improve classification outcomes. As seen, overall improvement was not evident across the five models (Fig. 5a; Table 1, bottom left). LR and LDA showed moderate classification accuracy (66% and 68%, respectively) but low sensitivity (48% and 53%), reflecting a limited ability to identify luminal B cases. SVM achieved the highest specificity (93%) but exhibited poor sensitivity (33%), suggesting bias toward the majority class. XGBoost achieved an accuracy of 71% and sensitivity of 57%, representing an improvement over its performance with polarimetric features alone, though these gains were accompanied by higher variability across folds. This variability suggests that XGBoost's performance was more sensitive to the composition of the training data, which may limit its reliability in small-sample settings.

In contrast, RF offered the most balanced performance across metrics, with an accuracy of 79%, sensitivity of 67%, specificity of 86%, and AUROC of 77%. While other classifiers excelled in isolated metrics—such as SVM in specificity or XGBoost in certain folds—RF consistently performed well across all evaluation criteria, supporting its selection as the final best-classification model (Fig. 5a and Table 1).

The trained RF model was then applied to an independent test set (n = 21 patients). As anticipated, performance declined somewhat relative to cross-validation, with test accuracy, sensitivity, and AUROC decreasing to 62%, 38%, and 74%, respectively (Fig. 5b). This reduction is consistent with typical patterns observed in small-cohort machine learning settings, where limited training data can lead to overfitting and reduced generalisation.

To address generalisability, the best-performing RF model was retrained on the full training and validation cohort (n = 47) and subsequently again evaluated on the test set (n = 21). Notably, although this retraining step was applied to a smaller subset (Nc = 68) of the original dataset (N = 116) (due to the limited availability of clinical features), classification performance improved, yielding 81% accuracy, 75% sensitivity, 75% specificity, and an AUROC of 86% (Fig. 5b–d, Table 1). These gains likely reflect the incorporation of

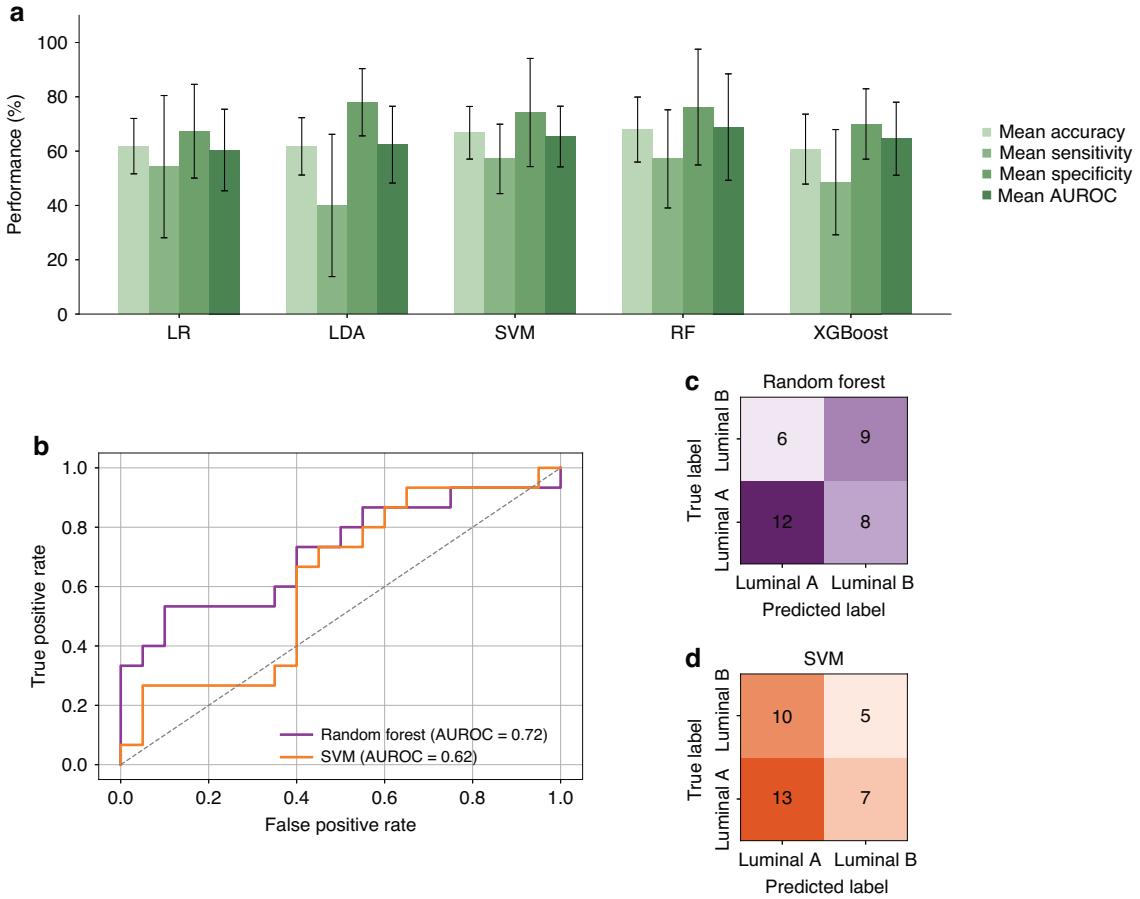

Fig. 4 Classification performance using polarimetric features. a Mean five-fold cross-validation performance on the training + validation set for LR, LDA, SVM, RF and XGBoost. Metrics shown include accuracy, sensitivity, specificity, and AUROC; see Table 1 (top left section) for detailed values. Error bars represent the standard deviation across folds. **b** ROC curves on the independent (previously unseen) test set for the two top-performing cross-validation classifiers, RF and SVM. Corresponding test-set confusion matrices for RF (**c**) and SVM **d**, showing classification performance in terms of true and predicted luminal subtypes.

additional, complementary informative features, specifically clinical biomarkers, into the model. Compared to models trained on the full 116 cases using polarimetric parameters only, the addition of clinical variables in this smaller, clinically enriched cohort enabled better separation of luminal subtypes. This highlights the importance of input feature quality and relevance in driving model performance.

DISCUSSION

While molecular profiling methods are commonly used to distinguish luminal A from luminal B subtypes, their routine use has proven challenging due to cost, complexity and time delays issues. Yet standard histopathological techniques alone are unable to differentiate these subtypes due to the absence of distinct morphological markers (Fig. 1). These challenges underscore the need for quantitative approaches that are objective, rapid, cost-effective, and readily implementable.

We thus evaluated whether label-free Mueller matrix polarimetry performed on unstained slides, either alone or in combination with clinical variables, could support the classification of luminal A and B breast cancer subtypes using core biopsy specimens. Models trained exclusively on polarimetric features achieved modest classification performance. Among the five algorithms tested, RF produced the most balanced cross-validation results and demonstrated greater consistency on the independent test set. However, performance remained modest, particularly in identifying Luminal B tumours—reflected in lower sensitivity scores, increased false negatives, and reduced AUROC. While binary classification often involves trade-offs between class-specific performance, favouring one class over the other can have clinical consequences. Correct identification of luminal B tumours, which are typically more aggressive and have different treatment implications, carries greater diagnostic weight. A model that identifies Luminal A well at the cost of consistently misclassifying Luminal B offers limited practical value. These findings suggest that although polarimetric imaging captures useful structural optical contrast, it may not offer sufficient discriminatory power on its own for reliable subtype classification in breast biopsy samples.

Integration of clinical variables into the models led to improved performance across all metrics, with RF achieving the highest test set accuracy and AUROC, following retraining on the full training and validation cohort. This improvement highlights the complementary nature of polarimetric and clinical data. The top-ranked features included both types of input—specifically, three polarimetric parameters (median values of M41, M44, and M33) and three clinical variables (biopsy- and excision-derived Nottingham grades and histologic type)—indicating that both domains contribute non-redundant and likely independent information.

To reduce dimensionality, we ranked features by random-forest importance using the training data only and retained the six highest-scoring variables for all subsequent classifiers. Because selection was confined to the training set, no information leaked into validation or test evaluations [55]. We acknowledge that this

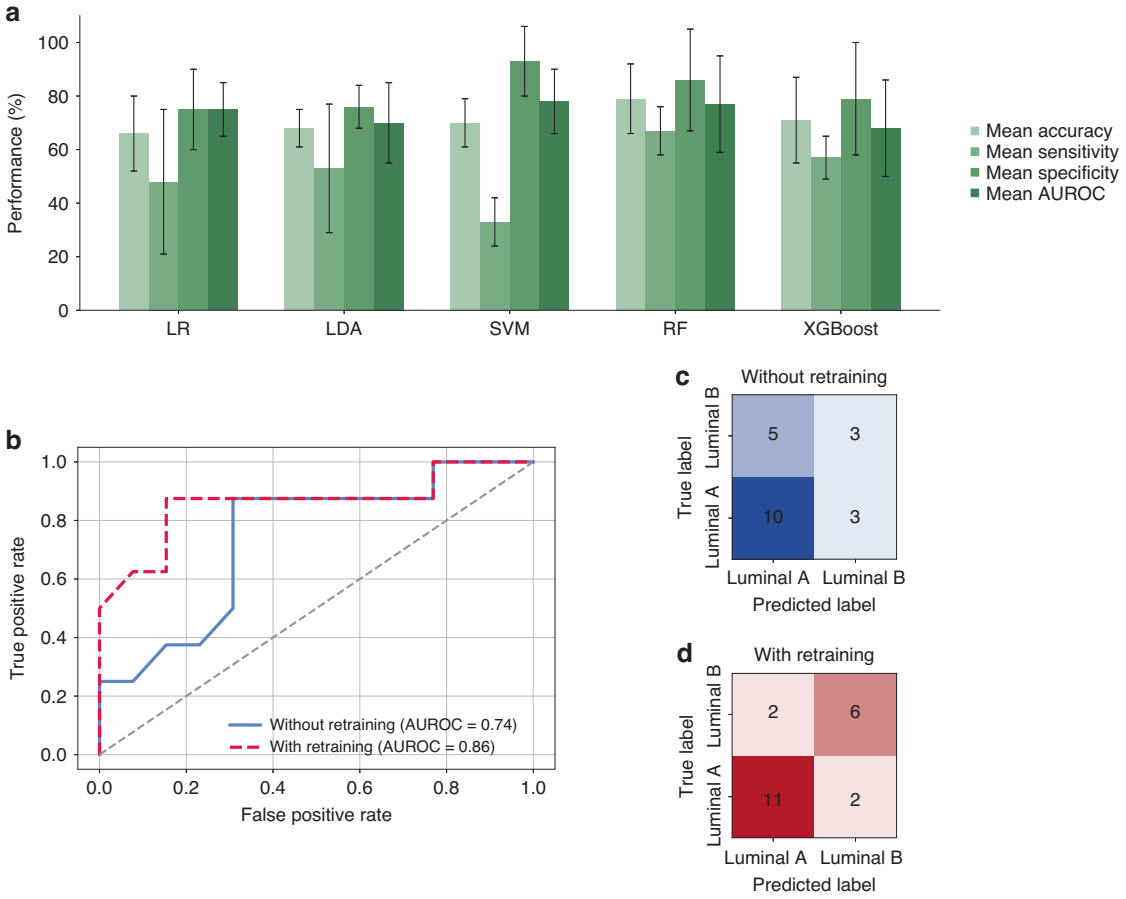

Fig. 5 Classification performance using the combined feature set of polarimetric and clinical variables. a Mean five-fold cross-validation performance on the training + validation set for LR, LDA, SVM, RF, and XGBoost. Metrics include accuracy, sensitivity, specificity, and AUROC; error bars show fold-to-fold variability (see Table 1 for values). **b** ROC curves on the test set for random forest before retraining (solid blue) and after retraining on the full training + validation data (dashed purple). Test-set confusion matrices for random forest without retraining (**c**) and after retraining (**d**), illustrating improved classification of luminal A and B subtypes. Diagonal segments in (**b**) reflect tied predicted probabilities across true positive and false positive cases, a known behaviour in small test sets; no smoothing or interpolation was applied.

filter-based approach can favour tree-based models—such as random forest itself—because the ranking reflects the relationships that those models exploit most effectively [56]. Nonetheless, with a dataset of only 116 cases, it provided a pragmatic, model-independent (of the downstream classifiers) starting point for this proof-of-concept study; more elaborate selection schemes can be unstable when sample sizes are this small [57]. Future work, utilising larger patient cohorts, will explore permutation-based or model-specific feature-selection strategies to determine whether these approaches improve performance [58].

The selected Mueller matrix parameters are associated with different aspects of tissue optical properties. M44 and M33 have been previously linked to depolarisation and tissue heterogeneity, while M41 is related to linear retardance and may reflect structural anisotropy [18, 19]. Although the direct physical interpretation of individual Mueller matrix elements is non-trivial, prior studies suggest that these features can capture microstructural differences such as collagen content, fibre alignment, and extracellular matrix organisation [18, 19, 52–54]. In our dataset, luminal B tumours tended to exhibit lower M44 values than luminal A, possibly reflecting reduced collagen alignment or increased matrix disruption. These findings are consistent with previous reports that associate luminal B tumours with greater architectural disorganisation and higher proliferative activity [5, 6, 10].

Clinical variables also contributed meaningfully to model performance. Both biopsy and excision Nottingham grades ranked among the top features, which may reflect the complementary nature of these histological assessments. Biopsy samples typically capture a limited, focal region of the tumour, potentially underestimating overall grade due to sampling variability. In contrast, excision specimens provide a broader and more representative view of tumour architecture. The presence of both as informative features suggests that they offer non-redundant information, likely capturing different aspects of tumour biology that are relevant for subtype differentiation. Histologic type also emerged as another important feature. Although it does not define molecular subtype, certain histological patterns are more frequently associated with specific molecular profiles. For example, invasive lobular carcinomas, typically oestrogen receptor-positive with lower proliferation indices, are more often luminal A, whereas luminal B tumours are generally higher grade and more proliferative [5, 11, 59]. These associations may explain why histologic type provided additional predictive value in the combined model.

The present analysis combined polarimetric biomarkers with all clinically available data (including variables recorded both at initial diagnosis and after surgical resection), with the intend to gauge maximal predictive potential. In routine practice, however, only clinical information obtainable at the diagnostic stage should be used to discriminate luminal A from luminal B, as this is the timeframe when treatment decisions are made. Accordingly, although our results indicate that polarimetry combined with clinical features can aid subtype classification, they remain

exploratory and require external validation on larger cohorts that rely solely on pre-operative clinical variables.

In parallel, we are also exploring the potential of polarimetric biomarkers to predict sentinel lymph node biopsy (SLNB) positivity or negativity—another essential issue in breast cancer staging and therapeutic decision-making [60]. The ability to assess SLNB status non-invasively through core biopsies could reduce the need for more extensive surgical procedures, thereby streamlining treatment planning in early-stage breast cancer. If polarimetric analysis can reliably indicate lymph node involvement, it could support more intelligent decisions regarding adjuvant therapies while minimising unnecessary interventions.

In conclusion, this study provides proof-of-concept demonstration that core biopsies could be used to obtain quantitative polarimetric biomarkers towards distinguishing between luminal A and B breast cancer subtypes. Importantly, this approach involves no subjective pathology involvement in defining specific ROIs for analysis. Clinically, a non-expensive, relatively simple, rapid, and easily implementable tool such as MMP that derives prognostic information from core biopsy samples would be highly desirable to provide early, minimally invasive insights to help inform robust treatment decisions prior to surgery or neoadjuvant therapy.

DATA AVAILABILITY

Data underlying the results presented in this paper are not publicly available at this time but may be obtained upon reasonable request to KT (k.tumanova@mail.utoronto.ca).

CODE AVAILABILITY

Code underlying the results presented in this paper is not publicly available at this time but may be obtained upon reasonable request to KT (k.tumanova@mail.utoronto.ca).

REFERENCES

1. Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, et al. Current and future burden of breast cancer: Global Statistics for 2020 and 2040. Breast. 2022;66:15–23.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–49.
3. Wilkinson L, Gathani T. Understanding breast cancer as a global health concern. Br J Radiol 2022;95:20211033.
4. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001;98:10869–74.
5. Yersal O, Barutca S. Biological subtypes of breast cancer: prognostic and therapeutic implications. World J Clin Oncol 2014;5:412–24.
6. Ignatiadis M, Sotiriou C. Luminal breast cancer: from biology to treatment. Nat Rev Clin Oncol 2013;10:494–506.
7. Colleoni M, Rotmensz N, Maisonneuve P, Viale G, Mastropasqua MG, Luini A, et al. Outcome of special types of luminal breast cancer. Ann Oncol 2012;23:1428–36.
8. Erber R, Hartmann A. Histology of luminal breast cancer. Breast Care. 2020;15:327–36.
9. Zhang X. Molecular classification of breast cancer: relevance and challenges. Arch Pathol Lab Med 2023;147:46–51.
10. Cheang MC, Chia SK, Voduc D, Gao D, Leung S, Snider J, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J Natl Cancer Inst 2009;101:736–50.
11. Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, et al. Clinical implications of the intrinsic molecular subtypes of breast cancer. Breast. 2015;24(Suppl 2):S26–35.
12. Toss A, Cristofanilli M. Molecular characterization and targeted therapeutic approaches in breast cancer. Breast Cancer Res. 2015;17:60.
13. Prat A, Ellis MJ, Perou CM. Practical implications of gene-expression-based assays for breast oncologists. Nat Rev Clin Oncol 2011;9:48–57.
14. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. N Engl J Med 2016;375:717–29.
15. Viale G, de Snoo FA, Slaets L, Bogaerts J, Rutgers E, van't Veer LJ, et al. Immunohistochemical versus molecular (BluePrint and MammaPrint) subtyping of breast carcinoma. Breast Cancer Res Treat 2018;167:123–31.
16. Myburgh EJ, de Jager JJ, Murray E, Grant KA, Kotze MJ, de Klerk H. The cost impact of unselective vs selective MammaPrint testing in early-stage breast cancer in Southern Africa. Breast. 2021;59:87–93.
17. Kromberg JGR, Sizer EB, Christianson AL. Genetic services and testing in South Africa. J Community Genet 2013;4:413–23.
18. Ghosh N, Vitkin IA. Tissue polarimetry: concepts, challenges, applications, and outlook. J Biomed Opt 2011;16:110801.
19. He H, Liao R, Zeng N, Liu S, Sheng W, Liu S, et al. Mueller matrix polarimetry—an emerging new tool for characterizing the microstructural feature of complex biological specimens. J Light Technol 2019;37:2534–48.
20. Sheng W, Li W, Qi J, Liu T, He H, Dong Y, et al. Quantitative analysis of 4 × 4 Mueller matrix transformation parameters for biomedical imaging. Photonics. 2019;6:34.
21. Sun M, He H, Zeng N, Du E, Guo Y, Liu S, et al. Characterizing the microstructures of biological tissues using Mueller matrix and transformed polarization parameters. Biomed Opt Express. 2014;5:4223–34.
22. Liu B, Yao Y, Liu R, Ma H, Ma L. Mueller polarimetric imaging for characterizing the collagen microstructures of breast cancer tissues in different genotypes. Opt Commun 2019;433:60–67.
23. Tumanova K, Serra S, Majumdar A, Lad J, Quereshy F, Khorasani M, et al. Mueller matrix polarization parameters correlate with local recurrence in patients with stage III colorectal cancer. Sci Rep. 2023;13:13424.
24. Majumdar A, Lad J, Tumanova K, Serra S, Quereshy F, Khorasani M, et al. Machine learning based local recurrence prediction in colorectal cancer using polarized light imaging. J Biomed Opt 2024;29:052915.
25. Lad J, Serra S, Quereshy F, Khorasani M, Vitkin A. Polarimetric biomarkers of peritumoral stroma can correlate with 5-year survival in patients with left-sided colorectal cancer. Sci Rep. 2022;12:12652.
26. Novikova T, Pierangelo A, De Martino A, Benali A, Validire P. Polarimetric imaging for cancer diagnosis and staging. Opt. Photonics News. 2012;23:26–33.
27. Pham TT, Ngoc Quach TN, Vo QH. Analysis of polarization features of human breast cancer tissue by Mueller matrix visualization. J Biomed Opt 2024;29:052917.
28. Xia L, Yao Y, Dong Y, Wang M, Ma H, Ma L. Mueller polarimetric microscopic images analysis based classification of breast cancer cells. Opt Commun 2020;475:126194.
29. Dong Y, Wan J, Si L, Meng Y, Dong Y, Liu S, et al. Deriving polarimetry feature parameters to characterize microstructural features in histological sections of breast tissues. IEEE Trans Biomed Eng 2021;68:881–92.
30. Westreich J, Khorasani M, Jones B, Demidov V, Nofech-Mozes S, Vitkin A. Novel methodology to image stromal tissue and assess its morphological features with polarized light: towards a tumour microenvironment prognostic signature. Biomed Opt Express. 2019;10:3963–73.
31. Jones B, Thomas G, Westreich J, Nofech-Mozes S, Vitkin A, Khorasani M. Novel quantitative signature of tumor stromal architecture: polarized light imaging differentiates between myxoid and sclerotic human breast cancer stroma. Biomed Opt Express. 2020;11:3246–62.
32. Jones B, Thomas G, Sprenger J, Nofech-Mozes S, Khorasani M, Vitkin A. Peritumoural stroma collagen organization of invasive ductal carcinoma assessed by polarized light microscopy differs between OncotypeDX risk group. J Biophotonics. 2020;13:e202000188.
33. Gribble A, Pinkert MA, Westreich J, Liu Y, Keikhosravi A, Khorasani M, et al. A multiscale Mueller polarimetry module for a stereo zoom microscope. Biomed Eng Lett 2019;9:339–49.
34. Smith MH, Woodruff JB, Howe JD. Beam wander considerations in imaging polarimetry. SPIE Proc. 1999;3754:50–4.
35. Thévenaz P, Ruttimann UE, Unser M. A pyramid approach to subpixel registration based on intensity. IEEE Trans Image Process. 1998;7:27–41.
36. Compain E, Poirier S, Drevillon B. General and self-consistent method for the calibration of polarization modulators, polarimeters, and Mueller-matrix ellipsometers. Appl Opt. 1999;38:3490–502.
37. Lu SY, Chipman RA. Interpretation of Mueller matrices based on polar decomposition. J Opt Soc Am A. 1996;13:1106–13.
38. He H, Zeng N, Du E, Guo Y, Li D, Liao R. et al. A possible quantitative Mueller matrix transformation technique for anisotropic scattering media. Photonics Lasers Med. 2013;2:129–37.
39. Khorana A, Pareek A, Ollivier M, Madjarova SJ, Kunze KN, Nwachukwu BU, et al. Choosing the appropriate measure of central tendency: mean, median, or mode? Knee Surg Sports Traumatol Arthrosc. 2023;31:12–15.
40. Whitley E, Ball J. Statistics review 1: presenting and summarising data. Crit Care. 2001;6:66.
41. Alpaydin E. Introduction to machine learning. 4th ed. Cambridge, MA, USA: The MIT Press; 2020.
42. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16 (New York, NY, USA: Association for Computing Machinery); 2016. Vol. 22, pp. 785–794.

43. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 2021;54:1937–67.
44. Gress TW, Denvir J, Shapiro JI. Effect of removing outliers on statistical inference: implications to interpretation of experimental data in medical research. Marshal J Med 2018;4:9.
45. Ben-Gal I. Outlier detection. In: Maimon O, Rokach L, editors. Data mining and knowledge discovery handbook: a complete guide for practitioners and researchers. Boston, MA, USA: Springer; 2005. pp. 131–46.
46. Hatoum FT, Tomek RM, Whitney HM, Giger ML. Importance of stratified sampling for use in the development of training and test sets: medical imaging AI applications. Proc SPIE Med Imaging. 2025;13407:134070Z.
47. Rogers J, Gunn S. Identifying feature relevance using a random forest. In: Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J, editors. Subspace, latent structure and feature selection. SLSFS 2005. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2006. Vol. 3940.
48. Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognit. 2015;48:2839–46.
49. Sivakumar M, Parthasarathy S, Padmapriya T. Trade-off between training and testing ratio in machine learning for medical image processing. PeerJ Comput Sci. 2024;10:e2245.
50. Tsamardinos I. Don't lose samples to estimation. Patterns. 2022;3:100612.
51. Qian J, Tanigawa Y, Du W, Aguirre M, Chang C, Tibshirani R, et al. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. PLoS Genet. 2020;16:e1009141.
52. Novikova T, Meglinski I, Ramella-Roman JC, Tuchin VV. Special section guest editorial: polarized light for biomedical applications. J Biomed Opt 2016;21:071001.
53. Tuchin VV. Polarized light interaction with tissues. J Biomed Opt 2016;21:71114.
54. Singh MD, Ghosh N, Vitkin IA. Mueller matrix polarimetry in biomedicine: enabling technology, biomedical applications, and future prospects. In: Ramella-Roman JC, Novikova T, editors. Polarized light in biomedical imaging and sensing: clinical and preclinical applications. Cham, Switzerland: Springer International Publishing; 2022. pp. 61–103.
55. Smialowski P, Frishman D, Kramer S. Pitfalls of supervised feature selection. Bioinformatics. 2010;26:440–3.
56. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinform. 2007;8:25.
57. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics. 2004;20:374–80.
58. Dernoncourt D, Hanczar B, Zucker J-D. Analysis of feature selection stability on high dimension and small sample data. Comput Stat Data Anal 2014;71:681–93.
59. Orrantia-Borunda E, Anchondo-Nuñez P, Acuña-Aguilar LE, Gómez-Valles FO, Ramírez-Valdespino CA. Subtypes of breast cancer. In: Mayrovitz HN, editor. Breast cancer. Brisbane, Australia: Exon Publications; 2022.
60. Gentilini OD, Botteri E, Sangalli C, Galimberti V, Porpiglia M, Agresti R, et al. Sentinel lymph node biopsy versus no axillary surgery in patients with small breast cancer and ultrasonographically node-negative axilla: the SOUND randomised clinical trial. JAMA Oncol. 2023;9:1557–64.

ACKNOWLEDGEMENTS

The authors thank James Jonkman and the staff at the Advanced Optical Microscopy Facility (University Health Network, Toronto, Canada) for their technical support and discussions. We also thank Marija Pizurica (Ghent University, Ghent, Belgium) for helpful discussions and insights related to machine learning. KT acknowledges support from the Ontario Graduate Scholarship and the Princess Margaret Hospital Foundation Graduate Fellowship in Cancer Research.

AUTHOR CONTRIBUTIONS

KT, MK, SN-M and AV designed the research study. SN-M and KT acquired data. KT analysed the data. KT wrote the manuscript. All authors revised the manuscript and approved the final version.

FUNDING

New Frontiers in Research Fund (NFRFE-2019-01049); Canadian Institutes of Health Research (PJT-156110); Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-04930).

COMPETING INTERESTS

The authors declare no competing interests.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Institutional ethics approval was obtained from University Health Network and Sunnybrook Hospital in Toronto, Canada, with a waiver of patient consent due to the study's retrospective design and anonymized data. All methods were performed in accordance with the relevant guidelines and regulations.

CONSENT FOR PUBLICATION

Not applicable. This study uses de-identified retrospective data, and the requirement for patient consent for publication was waived by the ethics boards.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Kseniia Tumanova.