MDPI

Article

hist2RNA: An Efficient Deep Learning Architecture to Predict Gene Expression from Breast Cancer Histopathology Images

Raktim Kumar Mondol [1], Ewan K. A. Millar [2,3,4,5], Peter H. Graham [3,6], Lois Browne [6], Arcot Sowmya [1] and Erik Meijering [1,*]

1   School of Computer Science and Engineering, UNSW Sydney, Kensington, NSW 2052, Australia;
    r.mondol@unsw.edu.au (R.K.M.); a.sowmya@unsw.edu.au (A.S.)
2   Department of Anatomical Pathology, NSW Health Pathology, St. George Hospital,
    Kogarah, NSW 2217, Australia; ewan.millar@health.nsw.gov.au
3   St. George and Sutherland Clinical School, UNSW Sydney, Kensington, NSW 2052, Australia;
    peter.graham@health.nsw.gov.au
4   Faculty of Medicine and Health Sciences, Sydney Western University, Campbelltown, NSW 2560, Australia
5   University of Technology Sydney, Ultimo, NSW 2007, Australia
6   Cancer Care Centre, St George Hospital, Sydney, NSW 2217, Australia; lois.browne@health.nsw.gov.au
*   Correspondence: erik.meijering@unsw.edu.au

Simple Summary: Breast cancer diagnosis and treatment can be improved by understanding the specific genetic makeup of a patient's tumour. Currently, this genetic information is obtained through expensive and time-consuming molecular tests, which are not widely reimbursed by healthcare systems. To address this issue, we propose a new, computationally efficient deep learning based method called hist2RNA to predict the expression of genes using digital images of stained tissue samples. Our approach successfully predicts gene expression and identifies breast cancer subtypes, enabling personalized treatments, thereby improving patient survival and overall breast cancer management.

Abstract: Gene expression can be used to subtype breast cancer with improved prediction of risk of recurrence and treatment responsiveness over that obtained using routine immunohistochemistry (IHC). However, in the clinic, molecular profiling is primarily used for ER+ breast cancer, which is costly, tissue destructive, requires specialised platforms, and takes several weeks to obtain a result. Deep learning algorithms can effectively extract morphological patterns in digital histopathology images to predict molecular phenotypes quickly and cost-effectively. We propose a new, computationally efficient approach called hist2RNA inspired by bulk RNA sequencing techniques to predict the expression of 138 genes (incorporated from 6 commercially available molecular profiling tests), including luminal PAM50 subtype, from hematoxylin and eosin (H&E)-stained whole slide images (WSIs). The training phase involves the aggregation of extracted features for each patient from a pretrained model to predict gene expression at the patient level using annotated H&E images from The Cancer Genome Atlas (TCGA, n = 335). We demonstrate successful gene prediction on a held-out test set (n = 160, corr = 0.82 across patients, corr = 0.29 across genes) and perform exploratory analysis on an external tissue microarray (TMA) dataset (n = 498) with known IHC and survival information. Our model is able to predict gene expression and luminal PAM50 subtype (Luminal A versus Luminal B) on the TMA dataset with prognostic significance for overall survival in univariate analysis (c-index = 0.56, hazard ratio = 2.16 (95% CI 1.12–3.06), $p < 5 \times 10^{-3}$), and independent significance in multivariate analysis incorporating standard clinicopathological variables (c-index = 0.65, hazard ratio = 1.87 (95% CI 1.30–2.68), $p < 5 \times 10^{-3}$). The proposed strategy achieves superior performance while requiring less training time, resulting in less energy consumption and computational cost compared to patch-based models. Additionally, hist2RNA predicts gene expression that has potential to determine luminal molecular subtypes which correlates with overall survival, without the need for expensive molecular testing.

Keywords: breast cancer; whole slide images; deep neural network; gene expression prediction

1. Introduction

Breast cancer is one of the most common cancers worldwide and a leading cause of death in women [1]. In Australia, over 3000 women die each year from it, although the 5 year survival has increased from 76% to 91% , which has been achieved through early diagnosis and improved treatment regimes [2,3]. Histopathology, clinical findings and imaging (CT, MRI, Ultrasound) are used for breast cancer diagnosis and stage classification [4,5]. Hematoxylin and eosin (H&E) histopathology slides contain essential phenotypic information and are used to provide key diagnostic and prognostic features (size, type, grade, lymph nodal status and expression of the biomarkers ER, PR, HER2, Ki67) to guide treatment decisions [6].

A critical clinical question is the selection of the postmenopausal luminal ER+ patients who may benefit from the addition of chemotherapy to endocrine therapy [7–9]. This difficult decision may be further supported by commercially available multigene tests to assess the risk of recurrence (e.g., Oncotype DX, Prosigna, Mammaprint, Endopredict etc.) and molecular subtype (PAM50: Luminal A, Luminal B, HER2-enriched or Basal). However, molecular profiling is expensive, tissue destructive, takes several weeks to obtain a result, and may not be reimbursed by healthcare providers and therefore not available to all patients with breast cancer [10,11]. Providing a cheap and scalable solution to predict molecular subtype using deep learning could make this important clinical information rapidly available to more patients. Whilst simplified IHC-based intrinsic molecular subtyping can be used as a surrogate for gene expression assays, it is less robust [12]. Therefore, the use of a rapid and low-cost method for gene expression prediction from digital images could offer clinically useful information for more patients.

With the advent of phenotype-rich medical images and next-generation sequencing (NGS) technology, artificial intelligence (AI) has led to the development of newer computer-aided diagnosis (CAD) systems capable of recognising complex patterns to aid in clinical decision making [13,14]. Deep learning (DL), a subset of AI, has demonstrated enormous potential for analysing pathology images and gene expression data, driving towards the goal of personalised treatment and identifying the underlying key molecular features of various diseases [15–17]. Deep learning models can also be used for molecular phenotyping (gene mutations, gene transcripts and proteins), thereby enabling translational research that may eventually facilitate detailed characterisation of individual tumours and inform clinically targeted treatment decisions [18–20]. Intervariability in histology slide staining is a challenge; however, various techniques can be applied to reduce bias and undesirable colour variation [21,22]. In addition, data augmentation [23], multiresolution patches [24] and overlapping patches [25] can be used to improve model generalisation and performance.

Several studies have proposed different approaches to predict molecular phenotypes from whole slide images (WSIs). Hong et al. proposed the use of multiple patches of different resolutions to enable a neural network to learn features at different scales, whereas Schmauch et al. used a clustering approach to reduce the dimensions and give a final prediction per WSI. Various pretrained models, such as ResNet18, ResNet50, ShuffleNet and InceptionV3, have been successfully used for molecular phenotype prediction in recent years [24,26–29]. Notably, ShuffleNet is particularly efficient due to its use of depth-wise separable convolution [30,31]. In addition, Tavolara et al. employed a custom convolutional neural network (CNN) that utilises multiple instance learning (MIL)-based attention pooling to predict gene expression, eliminating the need for expert annotation [32]. In our own research, we have experimented with several state-of-the-art pretrained deep learning models such as ResNet50, InceptionV3, RegNetY320, EfficientNetV2S and DenseNet201 for the prediction of molecular phenotypes from WSIs, and found that existing models typically require excessive training time and computational resources.

In this paper, we propose a new deep learning method for gene expression prediction from breast cancer histopathology images that requires substantially less training time, computational resources and energy usage, without impacting prediction performance. Our proposed hist2RNA model is inspired by bulk RNA sequencing that relies on averaged

gene expression from a homogenised population of cells [33]. Bulk RNA sequencing, however, has some drawbacks because it cannot show the specific composition of each cell type or the heterogeneity of the tissue. Our experimental results show that it is nevertheless possible to use histopathology images to efficiently predict gene expression to gain insight into disease biology and to better predict cancer subtypes for targeted treatment.

2. Materials and Methods

2.1. Study Overview

In this study, we aimed to predict mRNA expression levels of 138 genes using a deep CNN model on WSIs. To prepare the WSIs for analysis, we first annotated the cancer region, extracted image patches, and applied colour normalisation. These preprocessed images were subsequently utilised to train and optimise the deep CNN model by tuning hyperparameters and adjusting the model architecture. Finally, the optimised models were evaluated on a held-out test set and an external dataset.

2.2. Data Collection

Our study consisted of H&E-stained formalin-fixed paraffin-embedded (FFPE) digital slides from TCGA-BRCA (n = 495). We selected a sample of 495 cases out of total 1100 cases due to the complex and labour-intensive nature of annotating histopathological images. In order to maintain high data quality, our expert pathologist carefully reviewed the slides, leading to a smaller but more reliable dataset. WSIs from TCGA-BRCA were downloaded from the GDC Portal (Accessed 20 August 2021). We used samples from all 4 molecular subtypes for our experiment: Luminal A (LumA, 174 samples), Luminal B (LumB, 116 samples), Basal (135 samples) and HER2 (70 samples). All patients had corresponding RNA sequencing data available for analysis. For our study, we used 335 samples for training and 160 samples for testing (held-out test set) while maintaining the relative proportions of all subtypes. We also used an external TMA dataset of 498 patients with invasive breast cancer from a randomised radiotherapy clinical trial at St George Hospital, Sydney, Australia (ClinicalTrials.gov NCT00138814) (Accessed 23 January 2022) [34]. Ethics approval was provided by South East Sydney Local Health District Human Research Ethics Committee at the Prince of Wales Hospital, Sydney (HREC 96/16). Samples of the 498 patients were scattered randomly across 18 glass slides. Each slide contained multiple cores from different patients, and each patient's tumour was sampled with a 3 × 1 mm core using the Beecher Manual Arrayer MTA-1 as previously published [34]. This dataset included a range of clinical information for each patient, such as follow-up duration, overall survival status, tumour grade, tumour size, lymph node status, age at diagnosis and intrinsic subtypes determined by IHC. We tested model performance to predict survival on this external TMA dataset with additional exploratory analysis.

2.3. Data Preparation

2.3.1. Slide Annotation

An expert breast pathologist manually annotated the selected slides using QuPath [35]. The annotation was performed for localisation of the tumour outline, excluding any necrosis but including stroma and tumour infiltrating lymphocytes (TILs). The pathologist was blinded to any molecular or clinical features during annotation.

2.3.2. Image Data Preprocessing

In this study, H&E-stained tissues were taken from both the TCGA and TMA datasets. In order to optimise the balance between image resolution and file size, both datasets were downsampled to 0.25 μm/pixel (approximately ×40 magnification). A semi-automated approach using QuPath software was used to generate tissue masks, which excluded areas with artifacts and non-tissue regions. The annotated tumour regions were then tiled into 224 × 224 pixel patches, yielding approximately 1000 non-overlapping patches per sample. To address staining inconsistencies, vector-based colour normalisation was applied

to both datasets, resulting in improved quantitative results [21] (Figure 1A). Overall, the preprocessing steps for both datasets were conducted with the goal of producing high-quality image patches suitable for subsequent analysis.

Preprocessing of Histopathology Images

Preprocessing of Gene Expression
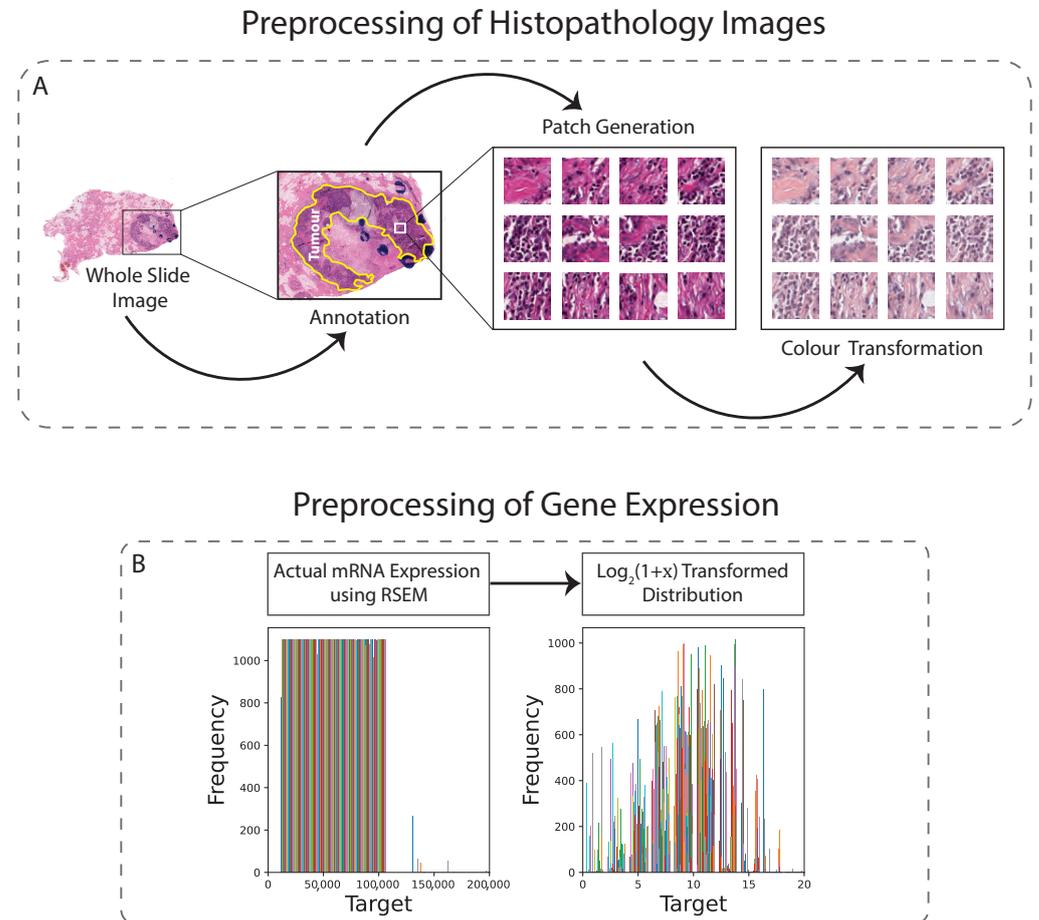
Figure 1. Preprocessing technique for histopathology images and gene expression. (**A**): WSIs are first annotated, then non-overlapping patches are generated from the annotated region, and finally colour transformation is applied. (**B**): Gene values are $\log_2(1 + x)$ normalised.

2.3.3. RNA Sequencing Data Preprocessing

We obtained transcriptome-wide RNA sequencing data representing mRNA expression levels for a total of 20,438 genes in the reference genome from the TCGA dataset, which were processed using RNA sequencing by expectation maximisation (RSEM) [36], and downloaded from the cBioPortal platform. Only patients with both RNA sequencing data and WSIs available were included in the study. We extracted genes used in the following commercially available assays: Oncotype DX, Mammaprint,Prosigna (PAM50), EndoPredict, BCI (Breast Cancer Index) and Mammostrat, which resulted in 138 genes as the final training targets (see Supplementary Materials for gene lists).

In addition to images, the genetic data also required preprocessing because regression analysis directly on raw RNA sequencing data would lead the model to focus only on the most strongly expressed genes, which would lead to poor model performance. As the normalised gene counts (x) can be equal to zero, we shifted them by 1 before log-transformation. Thus, we applied $\log_2(1 + x)$ transformation on the mRNA expression values (Figure 1B). These transformed data are typically less skewed, with fewer extreme values, but may contain unequal variances for the covariates [37]. Despite this, it is unlikely to be a major concern when using deep learning based regression models, as these models

can effectively adapt to unequal variances and other nonlinearities in the data, allowing the capture of complex relationships and patterns.

2.4. Feature Extraction, Aggregation and Model Training

The architecture of the proposed hist2RNA (Figure 2) uses a pretrained neural network to extract features from the image patches. More specifically, we considered five different pretrained models (EfficientNet, RegNet, DenseNet, Inception, ResNet) and evaluated their performance with our model. Feature extraction results in a tensor of dimension $N \times 1 \times F$, where N is the number of patches and F the number of features. Our model aggregates the patch-level features to obtain a slide-level feature representation in the form of a single vector of dimension $1 \times 1 \times F$ for each slide, as follows:

$$z = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

where z is the aggregated feature and each image patch is represented by a feature vector of size $1 \times F$, denoted as x_i, where i is the index of the patch. That is, the resulting slide-level vector has the same number of features as the original patch-level vectors, and represents their average. This aggregated feature is fed to three 1D convolutional layers followed by global average pooling and an output layer with 138 neurons with linear activation functions to build a regression model.
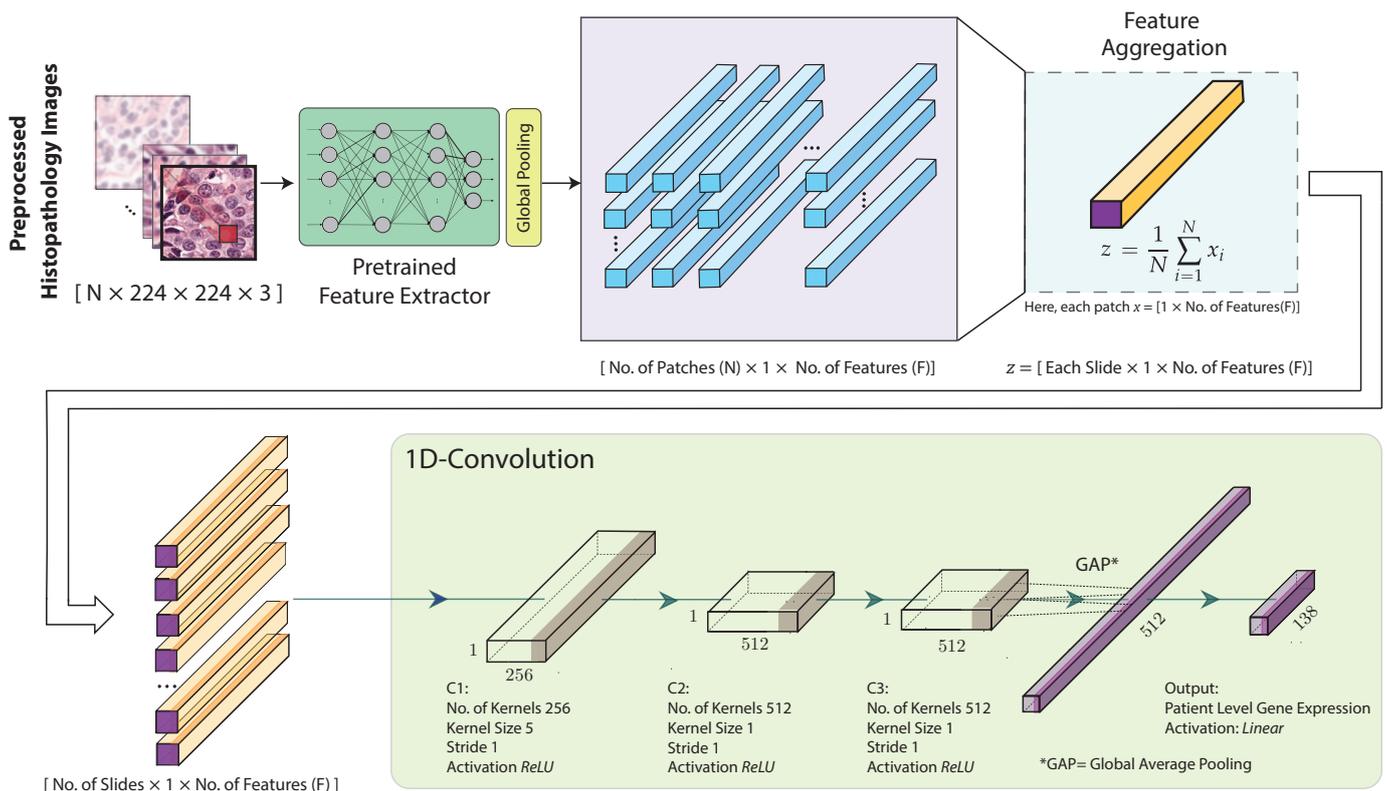
Figure 2. Proposed hist2RNA neural network architecture. After image preprocessing, a pretrained model extracts features from the histopathology image patches. The extracted features are then aggregated for each patient. Next, three convolutional blocks (C1, C2, C3) are used, each consisting of a 1D convolutional layer with a rectified linear unit (ReLU) activation function. C1 uses 256 kernels of size 5×5, while C2 and C3 use 512 kernels of size 1×1. Finally, the output layer consists of 138 neurons with a linear activation function, followed by a global average pooling layer.

We employed the Adam optimiser with the mean-squared error loss function and learning rate of 0.001. We used a minibatch of 12 samples per step and early stopping to monitor change in loss with a patience of 4. We ran continuous optimisation until the early stopping criterion was met or the maximum of 150 epochs were completed. The hyperparameters used in the optimisation process were determined empirically. During optimisation, we stored the best weights with the lowest mean-squared error loss for use in testing. Depending on when the early stopping criterion was met, optimisation runs took approximately 20 to 50 min on a single GPU (NVIDIA Tesla V100 32 GB).

2.5. Evaluation Metrics

Predictions of gene expression and genetic variation (such as mutation status, cancer grade, chromosomal instability, copy number variation and methylation status) are, respectively, regression and classification problems. The performance of the regression models in this project was evaluated using the Spearman rank correlation coefficient and its corresponding FDR (False Discovery Rate) adjusted *p*-value. This *p*-value was calculated using a t-distribution-based approach to determine the statistical significance of the observed correlation between predicted and actual gene expression values. The coefficient of determination (R^2) was also employed as a metric to assess the effectiveness of our regression model. The R^2 score measures the proportion of the variance in the dependent variable that can be explained by the independent variable, with values typically ranging from 0 to 1, with higher scores indicating that the regression model has a better fit to the data. However, in cases where the model performs poorly, R^2 can be negative. These parameters are essential in evaluating the effectiveness of the regression problem. In addition to the analyses, *t*-tests were conducted to evaluate the differences between positive and negative groups for each biomarker (ER, PR and HER2), while ANOVA tests were used to assess the variance in tumour grades (1, 2, 3). These statistical tests are instrumental in determining the presence of statistically significant differences among various groups.

Furthermore, survival analysis was evaluated with the concordance index (c-index), which was computed with the lifelines package [38]. The c-index measures the effectiveness of predicted risk scores on ranking survival times, with 1 indicating perfect concordance and 0.5 representing results from random predictions. Both univariate and multivariate survival models were employed to investigate the relationship between survival information and various predictor variables, such as cancer subtypes, tumour grade, tumour size, age, and lymph node status. The strength of these relationships was assessed using hazard ratios (HRs), which were accompanied by their corresponding confidence intervals (CIs). HR represents the hazard or risk of an event (such as death) occurring in one group relative to another group, while controlling for other variables. In the context of multivariate analysis, the HR for a LumB cancer subtype compared to a LumA subtype would indicate the relative risk of death in LumB patients compared to LumA patients, while controlling for the effects of other factors that may influence survival outcomes such as age, tumour grade and lymph node status. On the other hand, in the univariate analysis, the HR for LumB subtype compared to LumA was calculated without adjusting for other factors, providing a simple assessment of the relationship between cancer subtypes and survival outcomes.

3. Results

#### 3.1. Evaluating hist2RNA Method on Held-Out Test Data

We evaluated the proposed hist2RNA method (Figure 2) with five different pretrained models across patients and genes (Figure 3). Of the five models, EfficientNet appeared favourable, achieving a median Spearman correlation of 0.82 across patients on the held-out test set with few outliers, similar to the other models, while achieving the highest median correlation of 0.29 across genes, with statistically significant results for 105 out of 138 genes ($p < 5 \times 10^{-2}$). Of the top 20 genes (Figure 4) according to their corresponding correlation coefficients, 13 belong to the PAM50 gene set. In addition, 32 genes predicted by our model had a coefficient of determination (R^2) ≥ 0.1 (Figure 5), of which 17 were from PAM50.
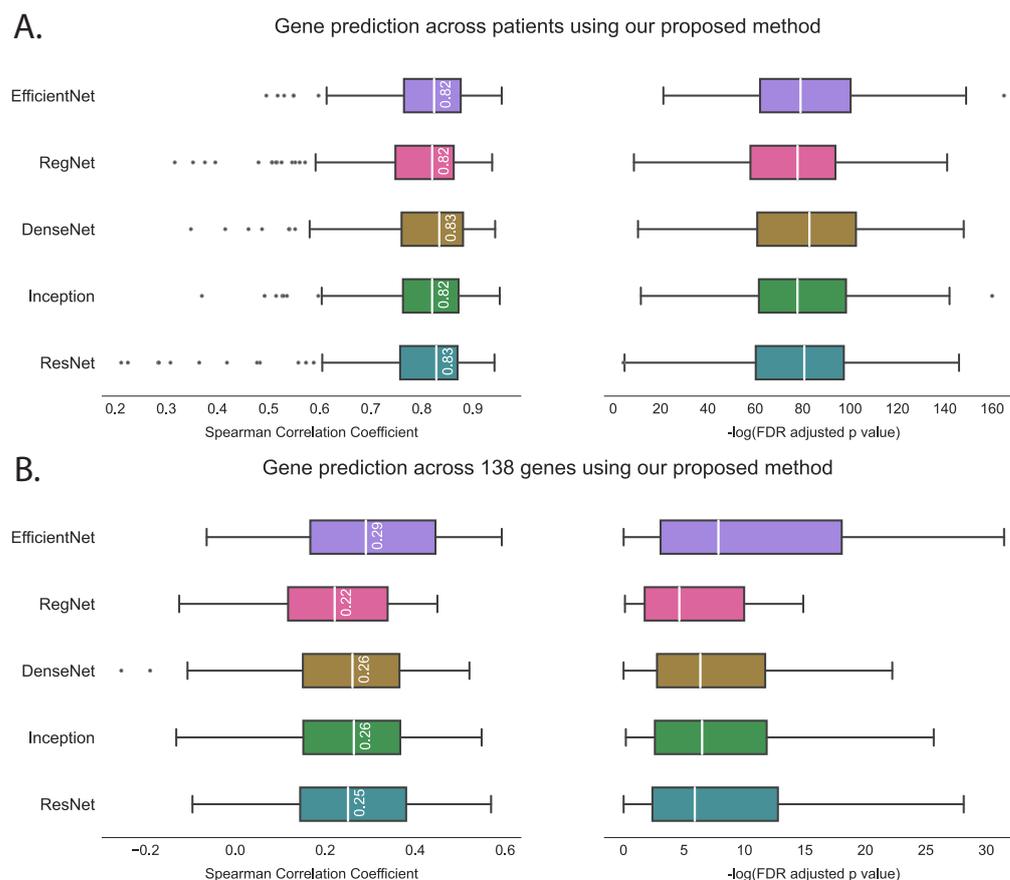
Figure 3. Performance comparison of five pretrained models on the held-out test set using the proposed hist2RNA method. The box plots show the distribution of Spearman correlation coefficients and their corresponding −log FDR-adjusted *p* values (where significance is represented by −log(5% FDR *p* value) > 1.30). The vertical white bar inside each box represents the median value, while the box itself indicates the interquartile range (IQR) from the 25th to the 75th percentile. The whiskers indicate the range of values beyond IQR and any data point that falls outside of the whiskers is considered an outlier and is marked with a gray circle. (**A**): Gene prediction is performed across patients, with the correlation coefficient calculated using predicted and true expression levels of all 138 genes for each patient. (**B**): Gene prediction is performed across 138 genes, with the correlation coefficient calculated for each gene separately, based on the predicted and true expression levels for all patients.

3.2. Classification of Breast Cancer Subtypes and Computational Efficiency of hist2RNA

Next, we aimed to develop a gene-expression-based classification model for breast cancer subtypes using the hist2RNA-predicted PAM50 gene set. We used a voting classifier consisting of four different machine learning algorithms, including Random Forests (RF), MultiLayer Perceptron (MLP), Linear Discriminant Analysis (LDA) and Logistic Regression (LR), with soft voting. In this experiment, the hist2RNA-predicted PAM50 gene set achieved reasonable accuracy and F1 score of 0.55 and 0.54, respectively, for the classification of the four subtypes (LumA, LumB, Basal and HER2). Our model achieved an AUROC (Area Under the Receiver Operating Characteristic curve) score of 0.79, 0.63, 0.89 and 0.78 for the LumA, LumB, Basal and HER2 subtypes. In terms of computational efficiency, the proposed method outperformed the patch-based approach as evidenced by the results (Figure 6), which showed 50× lower energy consumption, which can be attributed to EfficientNet's reduced training time.
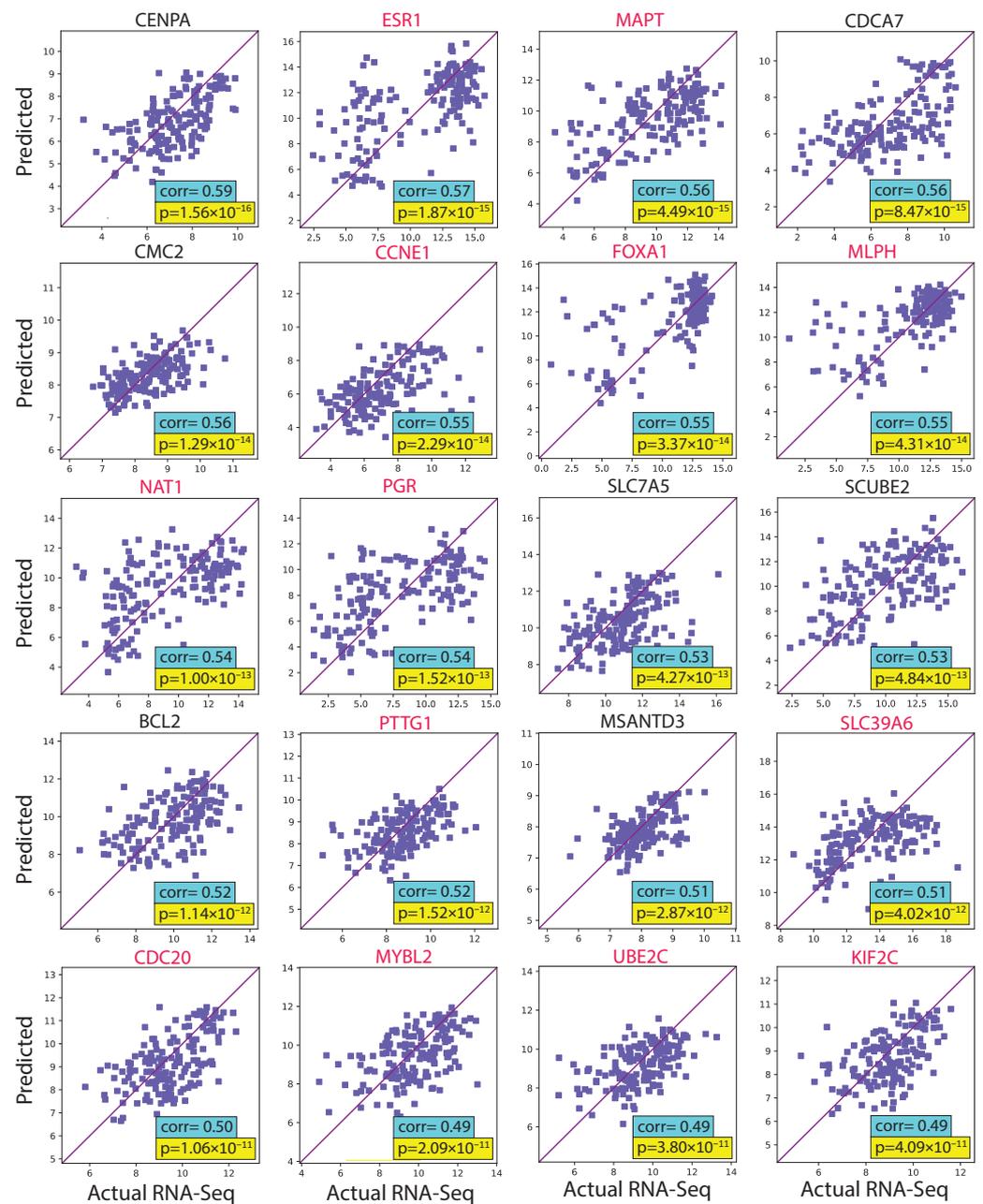
Figure 4. Scatter plots of predicted and actual RNA sequencing gene expression values for the 20 top genes ranked by correlation coefficient in the test set. PAM50 genes are highlighted in red.

### 3.3. Performance of hist2RNA on External TMA Dataset

In this study, we also evaluated the performance of our proposed method on an external TMA dataset. The results presented in the violin plot (Figure 7) indicate that there were statistically significant differences in the predicted median gene expression of ESR1, PGR and ERBB2 for ER, PR and HER2 positive and negative cases, respectively, in both TCGA and TMA datasets. Similarly, median gene expression of MKI67 showed statistically significant differences across the three grades correlated with increasing proliferation and biological aggressiveness. In addition, our analysis revealed a positive correlation ($\rho = 0.31$, $p = 2.81 \times 10^{-12}$) between predicted MKI67 gene and average Ki67% protein expression, as illustrated in Figure 8. It is worth noting that higher cancer grades are often associated with higher Ki67% values, which is confirmed by this study.
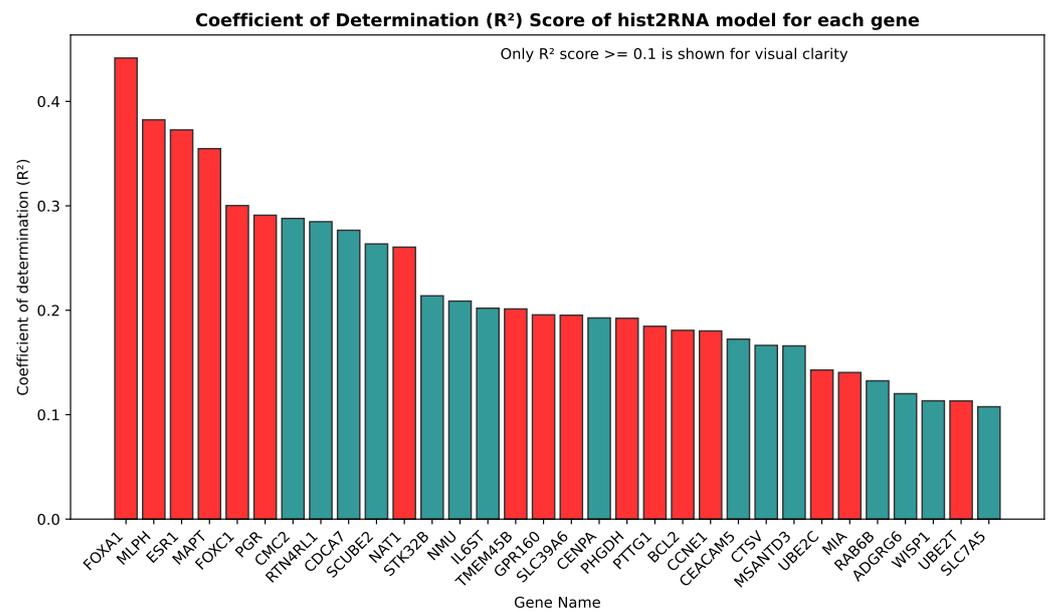
Figure 5. Coefficient of determination (R^2) scores for the top 32 genes predicted by our model. The bar plot includes all genes with an R^2 score greater than or equal to 0.1. PAM50 genes, which are important biomarkers for breast cancer subtyping, are highlighted in red.
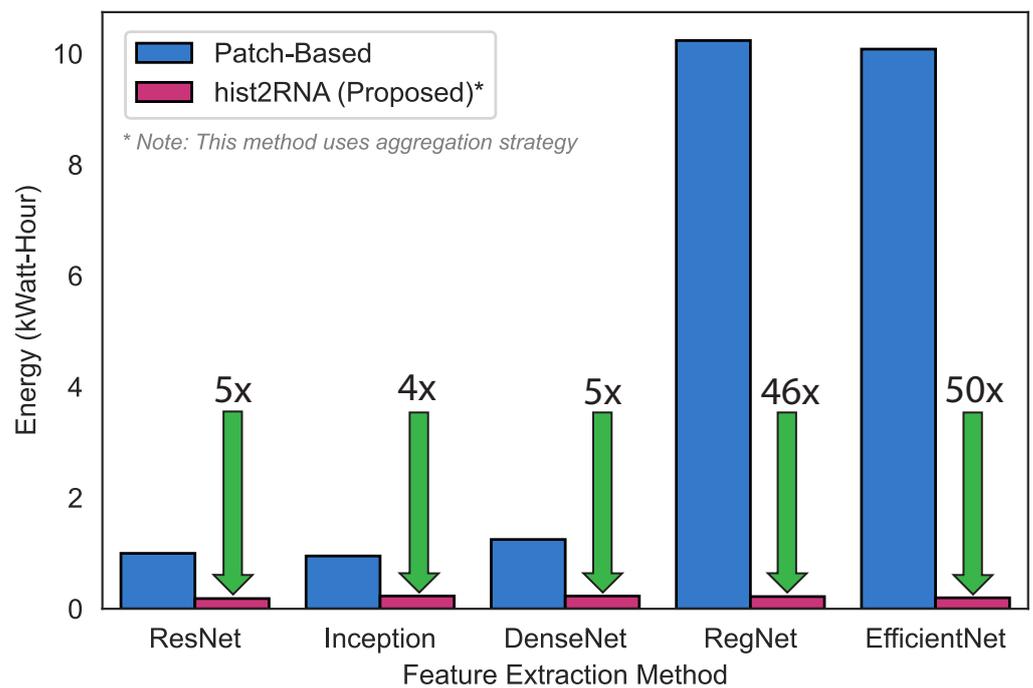
Figure 6. Energy consumption comparison based on the training times of the proposed strategy versus a patch-based strategy. To calculate energy (in kiloWatt-hours, kWh), the power consumption of the used NVIDIA Tesla V100 32 GB GPUs (about 300 W each) was multiplied by the training time (in hours) and divided by 1000. For example, training the EfficientNet patch-based method on 4 GPUs took 8.48 h at maximum power, translating to an energy consumption of $4 \times 8.48 \times 300 = 10{,}176$ kWh. By contrast, training our EfficientNet aggregation-based approach took only 0.66 h on a single GPU, thus consuming only 0.19 kWh, which is $50\times$ lower than the patch-based method.
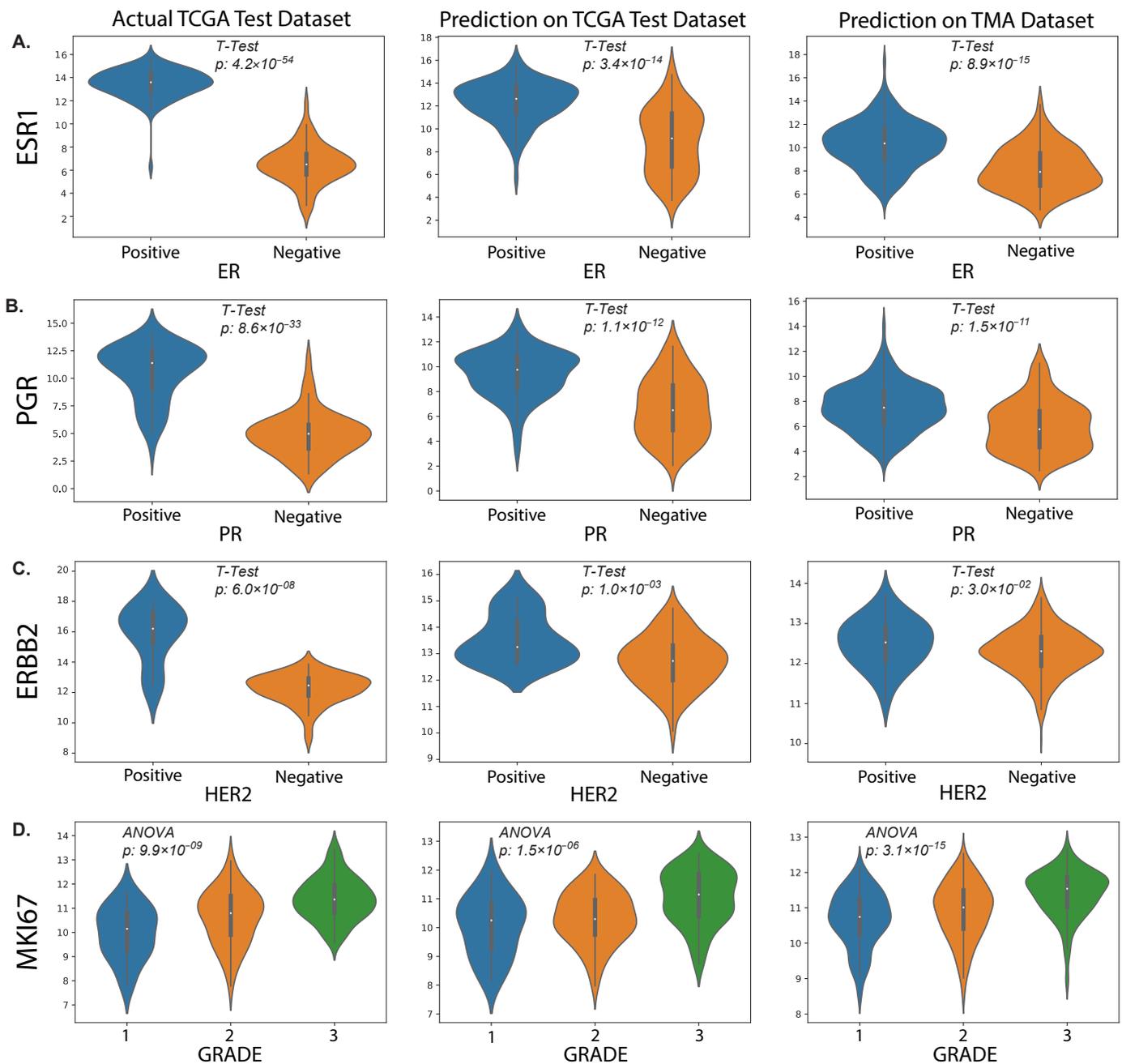
Figure 7. Predicted results from our proposed method. The plots show the distributions of RNA sequencing expression values for routine biomarkers ESR1 (**A**), PGR (**B**), ERBB2 (**C**) and MKI67 (**D**) according to the TCGA dataset (left column) and their predicted values on the held-out test set (middle column) and our external TMA dataset (right column), with respect to clinical status (IHC) of protein expression for the corresponding proteins encoded by each gene.

As mRNA gene expression data were not available for the external TMA dataset, we instead evaluated the surrogate subtype classification based on clinically used IHC-based methods. For this, we reclassified the PAM50 subtype using predicted genes from the TMA dataset, and first conducted patient-level univariate survival analysis solely using the predicted subtypes. The results revealed a c-index of 0.56, with an HR of 2.16 (95% CI 1.12–3.06), and $p < 5 \times 10^{-3}$ (Figure 9 and Table 1). Subsequently, we performed multivariate analysis by incorporating standard clinicopathological variables such as tumour grade, tumour size, age and lymph node status. The subtype prediction demonstrated independent prognostic significance with a c-index of 0.65, HR = 1.87 (95% CI

1.30–2.68) and $p < 5 \times 10^{-3}$ (Table 1). The *p*-values for both methods were statistically significant, indicating that both methods are able to distinguish between LumB and LumA subtypes with a high degree of confidence. IHC-based classification resulted in 96 patients being classified as LumB and 309 patients as LumA. In contrast, the proposed hist2RNA model resulted in 69 patients being classified as LumB and 336 patients being classified as LumA. This represents a shift of 27 patients from LumB to LumA in the proposed method compared to IHC-based classification. With regard to other clinopathological parameters, tumour size and patient age were consistently observed to be significant predictors of overall survival among luminal breast cancer patients, with HRs of around 1.6 and 3, respectively, in both univariate and multivariate analyses (Tables 1 and 2), where $p < 5 \times 10^{-3}$. Despite being a commonly used prognostic factor, tumour grade was not observed as a significant predictor, as indicated by $p > 5 \times 10^{-2}$ in both univariate and multivariate analyses.

Table 1. Univariate and multivariate analysis for overall survival in the TMA dataset of luminal patients using hist2RNA-predicted genes.

Parameter	Risk Group Cut-off Value	No. of Patients in Each Group	Multivariate (n = 406)			Univariate (n = b406)		
			HR	95% CI	*p*	HR	95% CI	*p*
Tumour Grade	1 & 2 vs. 3	324 vs. 81	0.94	0.62–1.42	7.6×10^{-1}	0.92	0.31–1.36	6.7×10^{-1}
Tumour Size	>20 vs. ≤20 (mm)	297 vs. 108	1.57	1.11–2.21	1×10^{-2}	1.56	0.77–2.15	1×10^{-2}
Age	>55 vs. ≤55	139 vs. 266	3.04	2.02–4.58	$<5 \times 10^{-3}$	2.96	1.49–4.43	$<5 \times 10^{-3}$
LN status	pos. vs. neg.	283 vs. 122	1.24	0.89–1.73	2×10^{-1}	1.43	0.67–1.96	3×10^{-2}
hist2RNA Predicted	LumB vs. LumA	69 vs. 336	1.87	1.30–2.68	$<5 \times 10^{-3}$	2.16	1.12–3.06	$<5 \times 10^{-3}$

Table 2. Univariate and multivariate analysis for overall survival in the TMA dataset of luminal patients using IHC-based subtype.

Parameter	Risk Group Cut-off Value	No. of Patients in Each Group	Multivariate (n = 406)			Univariate (n = 406)		
			HR	95% CI	*p*	HR	95% CI	*p*
Tumour Grade	1 & 2 vs. 3	324 vs. 81	0.72	0.45–1.13	1.5×10^{-1}	0.92	0.31–1.36	6.7×10^{-1}
Tumour Size	>20 vs. ≤20 (mm)	297 vs. 108	1.63	1.16–2.30	$<5 \times 10^{-3}$	1.56	0.77–2.15	1×10^{-2}
Age	>55 vs. ≤55	139 vs. 266	3.21	2.13–4.83	$<5 \times 10^{-3}$	2.96	1.49–4.43	$<5 \times 10^{-3}$
LN status	pos. vs. neg.	283 vs. 122	1.27	0.91–1.77	2×10^{-1}	1.43	0.67–1.96	3×10^{-2}
IHC Subtype	LumB vs. LumA	96 vs. 309	2.07	1.42–3.02	$<5 \times 10^{-3}$	1.68	1.20–2.34	$<5 \times 10^{-3}$
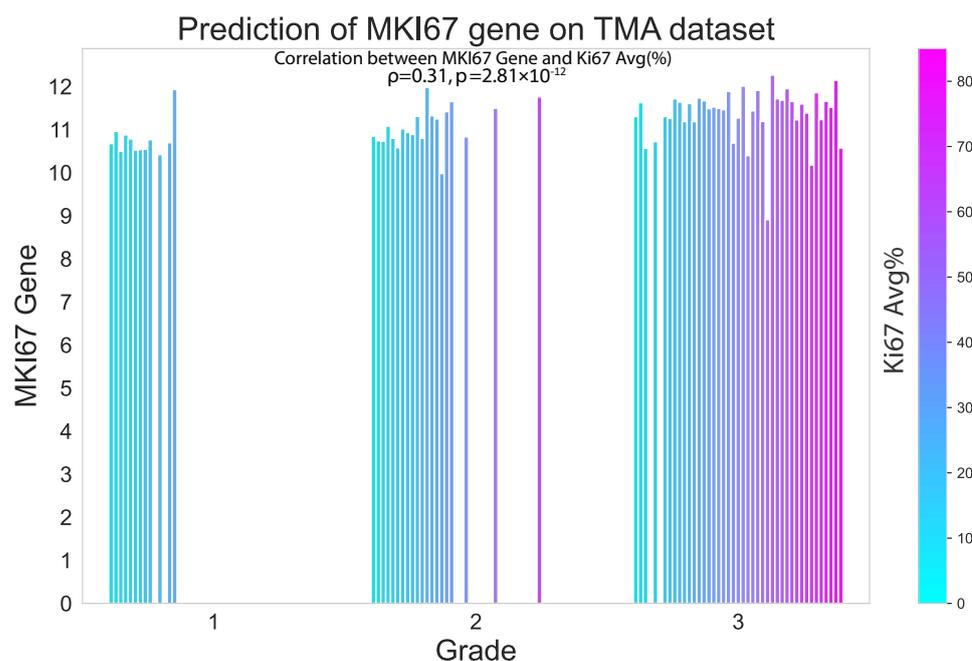
Figure 8. Proposed model performance in predicting MKI67 gene expression on the external TMA dataset. The results are seen to be significantly correlated with the Ki67% value.
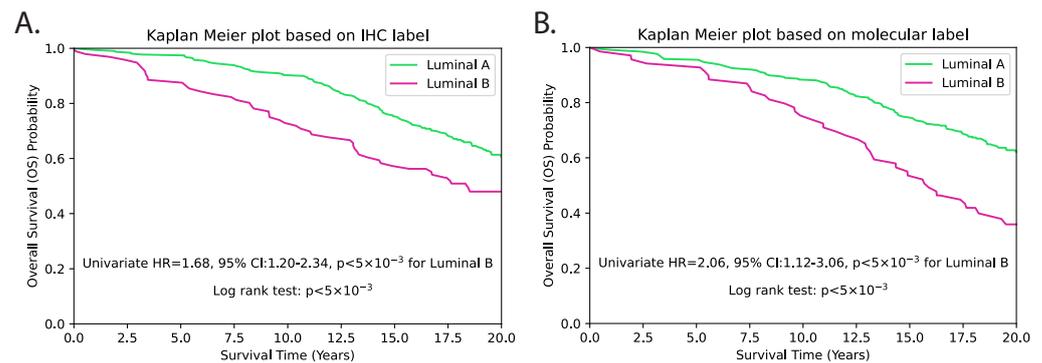
Figure 9. Kaplan–Meier survival estimation of luminal (LumA/LumB) categories on the TMA dataset using the IHC (**A**) and molecular (**B**) subtype label from hist2RNA predicted genes. The log-rank test *p*-value indicates a significant difference in survival between the two subtypes ($p < 5 \times 10^{-3}$).

4. Discussion

4.1. hist2RNA Model Performance in Gene Expression Prediction

We have demonstrated reliable and robust RNA sequencing gene expression prediction from breast cancer WSIs with our proposed hist2RNA model, which was validated on a held-out test set from TCGA. The results suggest that our model using the EfficientNet pretrained model for image feature extraction is well-suited for predicting gene expression across a diverse set of 138 genes related to risk prediction in ER+ breast cancer. The proposed aggregation strategy achieved a median Spearman correlation of 0.29 (FDR adjusted $p = 3.8 \times 10^{-4}$) across genes and 0.82 (FDR-adjusted $p = 4.3 \times 10^{-64}$) across patients, while also being computationally efficient in terms of energy consumption compared to the patch-based approach. The findings of the study reveal that the correlation coefficient for gene prediction across patients was significantly higher (0.82) in comparison to the correlation coefficient for gene prediction across 138 genes (0.29). This outcome can be attributed to the fact that gene prediction across patients takes into account the overall patterns of gene expression for each patient, allowing the model to capture the general trends in gene expression that are shared across patients. Conversely, when evaluating gene prediction across 138 genes, the correlation coefficient is calculated for each gene separately. This approach is more focussed on the individual gene expression patterns, which can exhibit significant variation between patients, thereby resulting in lower correlation coefficients.

4.2. Comparing Breast Cancer Subtype Classification Approaches

In this study, we carried out breast cancer subtype classification by leveraging the PAM50 gene set predicted from hist2RNA-based models. When comparing breast cancer subtype results with other studies, it is important to first acknowledge the differences in approach. Both Kather et al. and Liu et al. utilised direct image-to-subtype classification techniques, whereas our proposed method involves a two-step process: first image-to-gene prediction, followed by gene-to-subtype classification [39,40]. Despite these differences, our study found that the PAM50 gene set predicted by hist2RNA outperformed Kather et al.'s results in terms of AUROC scores across all subtypes: LumA (0.79 vs. 0.78), LumB (0.63 vs. 0.61), Basal (0.89 vs. 0.85) and HER2 (0.78 vs. 0.75). Furthermore, our study achieved an accuracy of 56% and an F1 score of 55% in breast cancer subtype classification, while Liu et al. achieved a higher accuracy of 64.3% and a higher F1 score of 68.5%. Our comparison of results with Liu et al. suggest that direct image-to-subtype approach can yield good performance in breast subtype classification. Therefore, further optimisation of gene expression generation process may be necessary to enhance the classification performance.

4.3. Clinical Implications of hist2RNA Model in Breast Cancer

We applied our findings to an independent external cohort of patients from a randomised clinical trial to demonstrate that our model has direct prognostic significance. The PAM50 gene set was chosen to predict subtypes such as LumA and LumB, and to evaluate its potential impact on improving patient survival. Furthermore, the clinical use of Ki67 as a proliferation marker remains standard practice in many countries due to the high cost of molecular profiling. The positive correlation (0.31) between our predicted MKI67 gene and Ki67% value by IHC staining suggests that our proposed method can serve as an alternative to IHC. We also showed that higher MKI67 expression value is associated with higher grade cancer and found statistically significant differences with lower grade cancer in both test and external data. Our results suggest that predicting ESR1 and PGR gene expression from WSIs may potentially be a good method for differentiating ER and PR protein status in breast cancer patients. The significant differences observed in the predicted gene expression levels between positive and negative cases, along with the strong positive correlations found for ESR1 ($\rho = 0.57$, $p = 1.87 \times 10^{-15}$) and PGR ($\rho = 0.54$, $p = 1.52 \times 10^{-13}$) in the test set, support this conclusion. However, based on our findings, the association between ERBB2 gene and HER2 status was not as strong as that observed for ESR1 and PGR, which may be due to the complex regulation of HER2 expression, the influence of other genetic factors or the relatively small sample size of our dataset.

4.4. Multigene Prognostic Signatures and Risk Stratification

Our multigene prediction model generates crucial prognostic information that is particularly useful for cancer patients where clinical parameters and traditional IHC markers alone lead to equivocal findings. The clinical importance of multigene prognostic signatures in cancer is extensively demonstrated in the current clinical practice guidelines for adjuvant systemic therapy in early-stage breast cancer and the disease staging guidelines [41]. In this study, our model reclassified about 28% of LumB patients as LumA, with corresponding good outcomes. This shift in classification could be due to the inclusion of molecular information that is not captured by IHC-based classification. This additional information may better differentiate between the LumB and LumA subtypes and result in more accurate risk stratification. A closer analysis of the HRs and CIs suggests that the proposed hist2RNA method may be a slightly better predictor of LumB subtype compared to the IHC subtype method. Univariate analysis of the proposed method showed higher risk of death for LumB compared to LumA subtypes, with a hazard ratio (HR) of 2.16 (95% CI 1.12–3.06, $p < 5 \times 10^{-3}$), while the IHC method had a lower HR of 1.68 (95% CI 1.20–2.34, $p < 5 \times 10^{-3}$) for the same subtypes (see Table 2). In contrast, multivariate analysis of LumB versus LumA subtypes indicated that the proposed method yielded a hazard ratio of 1.87 (95% CI 1.30–2.68), whereas the IHC method showed a slightly higher hazard ratio of 2.07 (95% CI 1.42–3.02). However, the CI for the proposed method was narrower, suggesting greater precision in the estimate. This suggests that the proposed model may be more accurate in predicting patient outcomes and risk stratification compared to the IHC-based classification, particularly in identifying patients with the LumB subtype who are at higher risk of recurrence and worse outcomes. The observed shift of 27 patients from LumB to LumA in our study suggests the use of less aggressive treatment recommendations and potential avoidance of chemotherapy toxicities. Overall, the findings of this study are encouraging, and we anticipate that our method may be useful for predicting other cancer types as well as other molecular phenotypes, such as mutation status prediction, copy-number alterations and microbiome prediction. Prediction of molecular phenotypes can enable cost-effective precision medicine that will allow large-scale epidemiological studies that include gene expression phenotypes as exposures for the research domain [26].

4.5. Comparison with Other State-of-the-Art Approaches

In comparing our proposed method with other state-of-the-art approaches, it is important to consider potential challenges arising from variations in datasets, target variables

(number of genes) and evaluation metrics used in previous studies. Therefore, a direct comparison between our method and previous studies may not be feasible. For example, Weitza et al. utilised ResNet18 to extract prostate cancer features and applied hierarchical clustering to reduce the number of CNN features, achieving Spearman correlation of 0.243 with an associated FDR adjusted $p < 1 \times 10^{-4}$ [28]. Schmauch et al. used a pan-cancer dataset from TCGA and employed ResNet50 to extract features, followed by k-means clustering to reduce dimensionality [26]. They achieved tile-level Pearson correlation of 0.19 on T-Cell genes and 0.23 on B-cell genes, with $p < 1 \times 10^{-4}$ in both cases. Tavolara et al. utilised attention-based multiple instance learning on mouse population data to predict the expression of the top five frequently occurring genes, achieving an average Pearson correlation coefficient of 0.59 [32]. Lastly, Wang et al. utilised a patch-based strategy and achieved a predicted median Spearman correlation coefficient of 0.4 with Bonferroni adjusted $p < 5 \times 10^{-2}$. However, their approach took 12–70 h on a single GPU (NVIDIA Tesla V100 32 GB) for each gene [29]. In contrast, our experiments indicated that the patch-based approach inspired by Wang et al.'s method yields a median Spearman correlation coefficient of 0.18 using RegNet on our test set, with 68 genes having $p < 5 \times 10^{-2}$ (Figure 10).
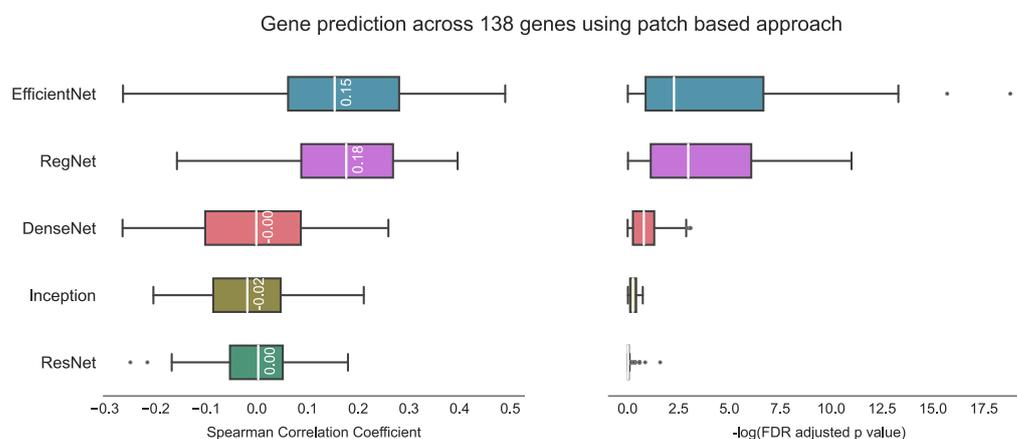
Figure 10. Gene prediction across 138 genes using the patch-based approach [29]. The best result with this approach is a median correlation coefficient of 0.18 using RegNet on our test data, with the corresponding −log FDR adjusted *p*-value indicating statistical significance (where significance is represented by −log(5% FDR *p* value) > 1.30). Notably, higher −log FDR adjusted *p*-values are considered to be indicative of better model performance.

4.6. Limitations, Challenges and Future Directions

Despite promising results, our study has some limitations. Generating gene expression data from images has the potential to introduce extra noise into the subtype classification task due to factors such as tissue heterogeneity, staining variability and image artefacts. Although we used an external dataset and performed exploratory analysis, we did not perform rigorous external validation of our method due to the absence of molecular information in the TMA dataset. Additionally, the external dataset we used had limited representation of the Basal and HER2 subtypes. Therefore, our exploratory analysis was limited to the LumA and LumB subtypes (n = 405). This limitation may affect the generalisability of our findings to other breast cancer subtypes, and further validation on a more diverse dataset is warranted. Six other studies have reported on predicting gene expression from histopathology images, each with its own limitations, including small sample size, significant training time, individual gene-based training, lack of independent external validation cohorts and lack of comparison with other pretrained models suggesting generally that more work is necessary [24,26–29,32].

Currently, it would be premature to propose that image analysis can replace gene expression assays for clinical applications. While there has been promising research on the use of image analysis to predict gene expression, several challenges need to be addressed

before it can be implemented in clinical practice. These challenges include the development of robust and reliable algorithms for image analysis, and the validation of image-based biomarkers in large and diverse patient populations. Finally, regulatory approval and integration of image analysis into clinical practice is needed, which requires collaboration between researchers, clinicians and regulatory agencies.

5. Conclusions

Summarising our findings, we conclude that mRNA expression in solid tumours can be effectively inferred from routine histology alone using deep learning approaches. We employed a public dataset and validated model robustness on a held-out test set, as well as conducted exploratory analysis on an external TMA dataset. Gene prediction from pathology images has substantial impact on patient survival by accurately identifying their molecular subtypes. These findings have the potential to pave the way for more personalised treatment planning for breast cancer patients in a time-efficient and cost-effective manner.

References

1. Wilkinson, L.; Gathani, T. Understanding breast cancer as a global health concern. *Br. J. Radiol.* **2022**, *95*, 1130. [CrossRef] [PubMed]
2. Foundation, N.B.C. Breast Cancer Stats in Australia. 2022. Available online: https://nbcf.org.au/about-breast-cancer/breast-cancer-stats/ (accessed on 20 November 2022).
3. Ji, P.; Gong, Y.; Jiang, C.C.; Hu, X.; Di, G.H.; Shao, Z.M. Association between socioeconomic factors at diagnosis and survival in breast cancer: A population-based study. *Cancer Med.* **2020**, *9*, 1922–1936. [CrossRef]
4. Hirra, I.; Ahmad, M.; Hussain, A.; Ashraf, M.U.; Saeed, I.A.; Qadri, S.F.; Alghamdi, A.M.; Alfakeeh, A.S. Breast cancer classification from histopathological images using patch-based deep learning modeling. *IEEE Access* **2021**, *9*, 24273–24287. [CrossRef]
5. Conti, A.; Duggento, A.; Indovina, I.; Guerrisi, M.; Toschi, N. Radiomics in breast cancer classification and prediction. *Semin. Cancer Biol.* **2021**, *72*, 238–250. [CrossRef]

6. Zhou, X.; Li, C.; Rahaman, M.M.; Yao, Y.; Ai, S.; Sun, C.; Wang, Q.; Zhang, Y.; Li, M.; Li, X.; et al. A comprehensive review for breast histopathology image analysis using classical and deep neural networks. *IEEE Access* **2020**, *8*, 90931–90956. [CrossRef]

7. Paik, S.; Tang, G.; Shak, S.; Kim, C.; Baker, J.; Kim, W.; Cronin, M.; Baehner, F.L.; Watson, D.; Bryant, J.; et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J. Clin. Oncol.* **2006**, *24*, 3726–3734. [CrossRef] [PubMed]

8. Albain, K.S.; Barlow, W.E.; Shak, S.; Hortobagyi, G.N.; Livingston, R.B.; Yeh, I.T.; Ravdin, P.; Bugarini, R.; Baehner, F.L.; Davidson, N.E.; et al. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: A retrospective analysis of a randomised trial. *Lancet Oncol.* **2010**, *11*, 55–65. [CrossRef] [PubMed]

9. Sparano, J.A.; Gray, R.J.; Makower, D.F.; Pritchard, K.I.; Albain, K.S.; Hayes, D.F.; Geyer, C.E.; Dees, E.C.; Goetz, M.P.; Olson, J.A.; et al. Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. *N. Engl. J. Med.* **2018**, *379*, 111–121. [CrossRef] [PubMed]

10. Ki, C.S. Recent advances in the clinical application of next-generation sequencing. *Pediatr. Gastroenterol. Hepatol. Nutr.* **2020**, *24*, 1–6. [CrossRef]

11. Deverka, P.A.; Dreyfus, J.C. Clinical integration of next generation sequencing: Coverage and reimbursement challenges. *J. Law, Med. Ethics* **2014**, *42*, 22–41. [CrossRef]

12. Szymiczek, A.; Lone, A.; Akbari, M.R. Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review. *Clin. Genet.* **2021**, *99*, 613–637. [CrossRef] [PubMed]

13. Lu, Z.; Zhan, X.; Wu, Y.; Cheng, J.; Shao, W.; Ni, D.; Han, Z.; Zhang, J.; Feng, Q.; Huang, K. BrcaSeg: A deep learning approach for tissue quantification and genomic correlations of histopathological images. *Genom. Proteom. Bioinform.* **2021**, *19*, 1032–1042. [CrossRef] [PubMed]

14. Karthiga, R.; Narashimhan, K. Deep Convolutional neural network for computer-aided detection of breast cancer using histopathology images. *J. Phys. Conf. Ser.* **2021**, *1767*, 012042. [CrossRef]

15. Yu, K.; Tan, L.; Lin, L.; Cheng, X.; Yi, Z.; Sato, T. Deep-Learning-Empowered breast cancer auxiliary diagnosis for 5GB remote e-health. *IEEE Wirel. Commun.* **2021**, *28*, 54–61. [CrossRef]

16. Katzman, J.L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; Kluger, Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **2018**, *18*, 24. [CrossRef]

17. Ragab, M.; Albukhari, A.; Alyami, J.; Mansour, R.F. Ensemble deep-learning-enabled clinical decision support system for breast cancer diagnosis and classification on ultrasound images. *Biology* **2022**, *11*, 439. [CrossRef]

18. Qu, H.; Zhou, M.; Yan, Z.; Wang, H.; Rustgi, V.K.; Zhang, S.; Gevaert, O.; Metaxas, D.N. Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. *NPJ Precis. Oncol.* **2021**, *5*, 87. [CrossRef]

19. Sammut, S.J.; Crispin-Ortuzar, M.; Chin, S.F.; Provenzano, E.; Bardwell, H.A.; Ma, W.; Cope, W.; Dariush, A.; Dawson, S.J.; Abraham, J.E.; et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **2022**, *601*, 623–629. [CrossRef]

20. Binder, A.; Bockmayr, M.; Hägele, M.; Wienert, S.; Heim, D.; Hellweg, K.; Ishii, M.; Stenzinger, A.; Hocke, A.; Denkert, C.; et al. Morphological and molecular breast cancer profiling through explainable machine learning. *Nat. Mach. Intell.* **2021**, *3*, 355–366. [CrossRef]

21. Macenko, M.; Niethammer, M.; Marron, J.S.; Borland, D.; Woosley, J.T.; Guan, X.; Schmitt, C.; Thomas, N.E. A method for normalizing histology slides for quantitative analysis. In Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Boston, MA, USA, 28 June–1 July 2009; pp. 1107–1110.

22. Vahadane, A.; Peng, T.; Sethi, A.; Albarqouni, S.; Wang, L.; Baust, M.; Steiger, K.; Schlitter, A.M.; Esposito, I.; Navab, N. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imaging* **2016**, *35*, 1962–1971. [CrossRef]

23. Sandfort, V.; Yan, K.; Pickhardt, P.J.; Summers, R.M. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Sci. Rep.* **2019**, *9*, 16884. [CrossRef]

24. Hong, R.; Liu, W.; DeLair, D.; Razavian, N.; Fenyö, D. Predicting endometrial cancer subtypes and molecular features from histopathology images using multi-resolution deep learning models. *Cell Rep. Med.* **2021**, *2*, 100400. [CrossRef] [PubMed]

25. Bernal, J.; Kushibar, K.; Cabezas, M.; Valverde, S.; Oliver, A.; Llado, X. Quantitative analysis of patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging. *IEEE Access* **2019**, *7*, 89986–90002. [CrossRef]

26. Schmauch, B.; Romagnoni, A.; Pronier, E.; Saillard, C.; Maillé, P.; Calderaro, J.; Kamoun, A.; Sefta, M.; Toldo, S.; Zaslavskiy, M.; et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat. Commun.* **2020**, *11*, 3877. [CrossRef] [PubMed]

27. Schrammen, P.L.; Laleh, N.G.; Echle, A.; Truhn, D.; Schulz, V.; Brinker, T.J.; Brenner, H.; Chang-Claude, J.; Alwers, E.; Brobeil, A.; et al. Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. *J. Pathol.* **2021**, *256*, 50–60. [CrossRef]

28. Weitz, P.; Wang, Y.; Kartasalo, K.; Egevad, L.; Lindberg, J.; Grönberg, H.; Eklund, M.; Rantalainen, M. Transcriptome-wide prediction of prostate cancer gene expression from histopathology images using co-expression-based convolutional neural networks. *Bioinformatics* **2022**, *38*, 3462–3469. [CrossRef]

29. Wang, Y.; Kartasalo, K.; Weitz, P.; Acs, B.; Valkonen, M.; Larsson, C.; Ruusuvuori, P.; Hartman, J.; Rantalainen, M. Predicting molecular phenotypes from histopathology images: A transcriptome-wide expression–morphology analysis in breast cancer. *Cancer Res.* **2021**, *81*, 5115–5126. [CrossRef]

30. Hoang, V.T.; Jo, K.H. PydNet: An efficient CNN architecture with pyramid depthwise convolution kernels. In Proceedings of the 2019 International Conference on System Science and Engineering, Dong Hoi, Vietnam, 20–21 July 2019; pp. 154–158. [CrossRef]

31. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856. [CrossRef]

32. Tavolara, T.E.; Niazi, M.K.; Gower, A.C.; Ginese, M.; Beamer, G.; Gurcan, M.N. Deep learning predicts gene expression as an intermediate data modality to identify susceptibility patterns in mycobacterium tuberculosis infected diversity outbred mice. *eBioMedicine* **2021**, *67*, 103388. [CrossRef]

33. Hegenbarth, J.C.; Lezzoche, G.; Windt, L.J.D.; Stoll, M. Perspectives on bulk-tissue RNA sequencing and single-cell RNA sequencing for cardiac transcriptomics. *Front. Mol. Med.* **2022**, *2*, 839338. [CrossRef]

34. Millar, E.K.; Browne, L.H.; Beretov, J.; Lee, K.; Lynch, J.; Swarbrick, A.; Graham, P.H. Tumour stroma ratio assessment using digital image analysis predicts survival in triple negative and luminal breast cancer. *Cancers* **2020**, *12*, 3749. [CrossRef]

35. Bankhead, P.; Loughrey, M.B.; Fernández, J.A.; Dombrowski, Y.; McArt, D.G.; Dunne, P.D.; McQuaid, S.; Gray, R.T.; Murray, L.J.; Coleman, H.G.; et al. QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **2017**, *7*, 16878. [CrossRef] [PubMed]

36. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef] [PubMed]

37. Zwiener, I.; Frisch, B.; Binder, H. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS ONE* **2014**, *9*, e85150. [CrossRef] [PubMed]

38. Davidson-Pilon, C.; Kalderstam, J.; Jacobson, N.; sean reed.; Kuhn, B.; Zivich, P.; Williamson, M.; Abdeali, J.K.; Datta, D.; Fiore-Gartland, A.; et al. *CamDavidsonPilon/Lifelines: v0.24.15*; Zenodo: Geneva, Switzerland, 2020. [CrossRef]

39. Kather, J.N.; Heij, L.R.; Grabsch, H.I.; Loeffler, C.; Echle, A.; Muti, H.S.; Krause, J.; Niehues, J.M.; Sommer, K.A.; Bankhead, P.; et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **2020**, *1*, 789–799. [CrossRef] [PubMed]

40. Liu, H.; Xu, W.D.; Shang, Z.H.; Wang, X.D.; Zhou, H.Y.; Ma, K.W.; Zhou, H.; Qi, J.L.; Jiang, J.R.; Tan, L.L.; et al. Breast cancer molecular subtype prediction on pathological images with discriminative patch selection and multi-instance learning. *Front. Oncol.* **2022**, *12*, 858453. [CrossRef]

41. Vieira, A.F.; Schmitt, F. An update on breast cancer multigene prognostic tests-emergent clinical biomarkers. *Front. Med.* **2018**, *5*, 248. [CrossRef] [PubMed]