Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1147–1164, DOI 10.12694/scpe.v26i3.4300

DEEP LEARNING BASED STACKED PROBABILISTIC ATTENTION NEURAL NETWORK FOR THE PREDICTION OF BIO MARKERS IN NON-HODGKIN LYMPHOMA

SIVARANJINI NAGARAJAN[*] AND GOMATHI MUTHUSAMY[†]

Abstract. The biomolecular characterization of Non-Hodgkin lymphoma (NHL) impacts the prognosis, therapy planning, and prediction of therapeutic response. The development of cancerous characteristics in lymphoma formation may often be attributed to certain genetic defects and the resulting disruption of oncogenic regulatory processes. The use of advanced technology has made it feasible to identify genetic variations and their corresponding biomarkers. However, the current challenges in histopathology include the identification techniques and the presence of different cell types inside a tumour. Computational techniques are now being used more often to diagnose genetic abnormalities without invasive procedures. This is done by analysing quantitative imaging data. Therefore, we are now deploying a deep learning-based stacking probabilistic attention neural network in this project. In this study, the histopathological images are obtained from the Kaggle source. Next, the image may undergo preprocessing using the soft switch Weiner filter (SSWF). The area of interest was segmented using the hierarchical seed polarity transform (HSPT). The biomarker linked with Non-Hodgkin lymphoma is categorised using the stacked probabilistic attention neural network (SPANN) based on the segmented output. The whole experiment was conducted using a histopathologic cancer dataset from Kaggle under python environment. The proposed strategy outperformed the current state-of-the-art alternatives by obtaining high range of accuracy(95%), precision(95%), recall(95%) and F score (92%).

Key words: Non-hodgkin lymphoma, Bio marker, deep learning, soft switch weiner filter, hierarchical seed polarity transform , stacked probabilistic attention neural network

1. Introduction.

B-non-Hodgkin lymphomas (B-NHLs) are a subgroup of B-cell lymphomas that often display characteristics resembling the first phases of normal B-cell maturation. Flow cytometry, immunohistochemistry (IHC), immunoglobulin clonality assessment, fluorescence in situ hybridization (FISH), and next-generation DNA sequencing may be employed together with standard cytogenetics to more precisely classify these cancers. Protein expression is assessed by doing immunohistochemical staining on tissue sections placed on glass slides. This data is then used to guide clinical decision-making in many diagnostic scenarios, including cancer classification, detection of remaining illness, and identification of mutations. Standard brightfield chromogenic Immunohistochemistry staining, when performed at a high-throughput level, has limitations such as a restricted range of variation and images that have a significant overlap between the chromogen and the stain. This requires the use of specialised digital techniques to separate and deconvolve the stains as a preprocessing step for advanced research and commercial quantification algorithms used in Immunohistochemistry. Additional research is necessary to find dependable biomarkers for NHL. Despite thorough hyper-parameter tuning on a case-by-case basis or the laborious and error-prone manual tagging of many markers linked with NHL, colour separation remains suboptimal in areas with significant chromogen overlap. Multiplex immunofluorescence (mpIF) staining is more effective than brightfield immunohistochemistry (IHC) staining because it enables the analysis of multiple markers either separately (without the need for stain deconvolution) or together (as a composite). This leads to enhanced co-localization, standardised staining, objective scoring, and determination of thresholds for all marker values, particularly in regions with low expression that are challenging to evaluate using IHC staining. A new meta-analysis suggests that deep learning has the potential to replace the labor-intensive manual detection methods presently employed for gene expression profiling or immunohistochemically

[*] Department of Computer Science, Auxilium College (Autonomous), Vellore, & Periyar University, Salem, Tamil Nadu, India (sathiya.siva5@gmail.com)

[†] Department of Computer Science,Government Arts and Science College, Komarapalayam, Tamil Nadu, India (mdgomathi@gmail.com)

stained photographs. These methods are costly and limited due to the paucity of multiplex immunofluorescence (mpIF) testing. By using computational tools, which provide several benefits, we have a unique chance to improve the prognosis of the most lethal illnesses. Although co-registered high-dimensional imaging of the same tissue samples can offer crucial reference data for the superimposed brightfield IHC channels, current deep learning methods depend exclusively on unreliable manual annotations, which suffer from unclear cell boundaries, overlapping cells, and difficulties in assessing low-expression regions. Our method utilises a unique deep learning technique, using a stacked probabilistic attention neural network, to achieve more precise categorization of biomarker cells with enhanced gene specificity. Using a single registered IHC and training data from the same slides enables this.

Our method utilises a unique deep learning technique, using a stacked probabilistic attention neural network, to achieve more precise categorization of biomarker cells with enhanced gene specificity. Using a single registered IHC and training data from the same slides enables this. A trained stacked probabilistic attention neural network can effectively detect NHL biomarkers using just an immunohistochemistry (IHC) picture as input.

This study aims to accomplish the following objectives.

- In-order to get the precise output soft switch weiner filter based preprocessing was used.
- To separate the region of interest from the image hierarchical seed polarity transform was used.
- For classifying the cancer associated biomarker stacked probabilistic attention neural network model was implemented.

The rest of the study is laid out as follows. We shall review the current research in this field in the second section. The statement of the issue is given in Section 3. Our methodology's outline may be found in Section 4. Section 5 describes our approach's implementation and assessment. The conclusion section of our analysis is in Section 6.

2. Related works. Lymphomas are malignancies that originate in certain cells of the immune system. They are categorised into two primary groups: Hodgkin lymphomas (HL) and non-Hodgkin lymphomas (NHL. HL and NHL vary in their growth patterns and microscopic appearance. The early identification of these diseases is essential because of its considerable influence on treatment results. Some of the strategies shown here have potential as future versions of NHL prediction systems.

The main objective of [1] was to emphasise the need of including predictive biomarkers. Initially, artificial intelligence (AI) was used to the data obtained from a specific dataset (GSE10846) including the gene expression profiles of 414 patients. A combination of machine learning and predictive analytics models, including the C5.0 algorithm, logistic regression, Bayesian Network, discriminant analysis, random trees, tree-AS, and Chi-square Automatic Inference, was employed to decrease the number of dimensions in the investigation of a potential relationship between overall survival and other clinicopathological variables.

The author of [2] conducts a morphologic analysis of histological sections from 209 patients with DLBCL, together with clinical and cytogenetic data. We used tissue microarrays (TMAs) made from three identical core slices to perform staining for CD10, BCL6, MUM1, BCL2, and MYC using H&E and immunohistochemical stains. The pathologists have assigned labels to the tissue microarrays (TMAs) indicating the regions of interest (ROIs) that specifically identify tissue samples that test positive for diffuse large B-cell lymphoma (DLBCL). We used a deep learning model to detect specific areas of interest (ROIs), isolate and classify all cancer cell nuclei inside those ROIs, and quantify various geometric properties for each nucleus. Gene expression analysis has shown its utility in predicting the success of DLBCL therapy [[3], [4]]. The author of [4] suggests a novel approach to enhance the selection of optimal disease targets for a multilayer biomedical network by using PPI data that is annotated with stable information from OMIM diseases and GO biological processes. The author presents enough evidence to substantiate the efficacy of the RecRWR approach.

The author of [5] uses two approaches, namely MIDER (Mutual Information Distance and Entropy Reduction) and PLSNET (Partial least square based feature selection), to analyse data and establish the topology of a Gene Regulatory Network (GRN) by computational means. Gene expression analysis were used to demonstrate both methodologies in the context of inflammatory bowel disease (IBD), pancreatic ductal adenocarcinoma (PDAC), and acute myeloid leukaemia (AML). All the genes that regulate these three pathways have been identified. The UGT1A gene family was shown to have a critical role in regulating inflammatory bowel illness in the dataset. Similarly, the SULF1 and THBS2 genes were discovered as important factors in the pancreatic cancer dataset. Furthermore, they demonstrate that combining the results of the MIDER and PLSNET methods may result in a more precise ensemble-based strategy for inferring the topology of the gene regulatory network from data. Furthermore, an approximate estimate for the sample size of upcoming validation tests was established. They proposed an analytical approach that may identify potential regulator genes for validation testing and determine the required sample size for these studies. The objective of the suggested augmented ensemble learning approach in [6] is to improve the speed and accuracy of medical diagnosis. This model has been used to investigate a diverse array of ailments, including Alzheimer's, pancreatic, brain, and breast malignancies. The results indicate that the proposed model surpasses the existing techniques in terms of both accuracy and latency.

The author of [7] provides a concise overview of the current state of research and clinical use of MRI biomarkers in cancer therapy. This article provides a comprehensive discussion of MRI biomarkers, including the method of collecting and preprocessing MRI data, as well as the use of machine learning techniques. It concludes with an overview of the many types of biomarkers and their clinical utility in various cancer types.

A method for categorising solid lung cancer that has been treated before, based on the detection of anaplastic lymphoma kinase (ALK) gene rearrangement, was established in [8]. Scientists at [9] aimed to develop a deep learning system capable of directly predicting the immunohistochemistry (IHC) phenotype using whole-slide images (WSIs). This would enable more precise subtyping of lung cancer using resected and biopsied tissues. The objective of the study [10] was to provide an automated method for quantifying CMYC. In order to determine the proportion of cancer cells that express CMYC, researchers use attention-based multiple instance learning. This method involves analysing tissue microarray cores that have been evaluated by a pathologist.

The author of [11] selected the expression of the Ki-67 protein as a molecular information proxy. The researchers proposed a deep convolutional network model to predict the presence of Ki-67 positive cells using H&E stained slides. The researchers gathered images of cells that were labelled as either negative or positive for Ki-67, along with pictures of the surrounding tissue and the microscope plate. These images were then used to train the algorithm. Slides that have been stained with haematoxylin and eosin may be analysed for follicular lymphoma (FL) using an innovative deep-learning algorithm. The programme's accuracy is determined by a confidence estimate level set in a previous study [12]. A Bayesian neural network (BNN) was trained, tested, and scored using whole-slide images of lymph nodes exhibiting FL or follicular hyperplasia.

The researcher in [13] used deep learning techniques to develop a software application capable of detecting the MYC rearrangement in digital histology slides of diffuse large B-cell lymphoma. Slides stained with hematoxylin and eosin (H&E) were used for the purpose of instructing and evaluating medical students and professors from a total of 11 distinct institutions.

The author of [14] created a multitask deep learning system named DeepLIIF to address the challenges of stain deconvolution/separation, cell segmentation, and quantitative single-cell IHC scoring simultaneously. This paper presents a new dataset that combines co-registered immunohistochemistry (IHC) and multiplex immunofluorescence (mpIF) staining on the same slides. We use this dataset to convert affordable IHC slides into more informative but costly mpIF images. Additionally, we utilise this dataset to provide the required reference information for the overlaid brightfield IHC channels. The author has devised a gene expression test [15] that can differentiate between the seven most prevalent subtypes of B-cell NHL. This study uses ligation-dependent reverse transcription polymerase chain reaction (RT-PCR) and next-generation sequencing to investigate the expression of more than 130 genetic markers. The main objective of the method was to restore microenvironmental indicators of gene expression linked to B-NHL cells. We used a random forest methodology for classification, which we trained and validated using a dataset of more than 400 cases exhibiting diverse histology. The therapeutic effectiveness of the treatment was shown by the restoration of cell-of-origin signatures and the normalisation of MYC and BCL2 expression levels in high-grade lymphomas. Additionally, the treatment successfully prevented major misclassification in low-grade lymphomas. Therefore, this highly accurate pan-B-NHL predictor, which allows for a methodical assessment of several diagnostic and prognostic indicators, may be suggested as a supplementary tool to conventional histology in guiding patient management and enhancing patient classification for pharmacological trials.

The author in [16] explores the capacity of machine learning (ML) techniques to enhance the Cox Proportional Hazard (CoxPH) model. The authors thoroughly analyse the flaws in the most recent version of the CoxPH model and then provide a diverse array of remedies, including both established and innovative approaches. The accuracy of the models is evaluated using two metrics: the Brier score and the concordance index. Ultimately, drawing on our discoveries, they provide a series of recommendations on how practitioners might effectively capitalise on the latest advancements in AI.

The paper [17] outlines a method for subtyping NHLs by combining transfer learning (TL) with principal component analysis (PCA).When implemented on disorganised data, the scalable approach described in [18]—a Neural network—produces dependable results.

The author of [19] developed a MUltiple SUV Threshold (MUST)-segmenter to identify tumours on PET scans. This method involves placing seed points and then extending them into areas.The study investigated the integration of clinical, molecular genotype, and radiomics characteristics in predicting the prognosis of individuals with aggressive B-cell lymphoma [20]. We used fluorescent in situ hybridization to examine gene rearrangements of MYC, BCL2, and BCL6.

3. Problem statement. As of from the literature survey the primary scientific challenge in oncology is to identify the tumours or genes responsible for cancer and their mutational interactions with other organ systems in the body. The primary challenge in analysing this data is its unstructured and varied morality. The data are sourced from several places, resulting in the following issues:

1. The data annotations that are not sequential throughout a wider array of patients.
2. The annotation fails to provide any insights into the therapeutic actions necessary to enhance data quality.
3. To expedite drug discovery for therapeutic development.
4. To expedite drug discovery for therapeutic development.
5. Timely selection of appropriate medication is essential due to the absence of longitudinal data about the survival duration of cancer patients within the community.
6. Fragmentation of clinical data across organisations, incompatibility of data standards, and lack of system interoperability result from inadequate methods for the diffusion of innovation.
7. The intelligent and effective storing of extensive gene expression or image data is very challenging.
8. Even a skilled scientist finds it hard to manually evaluate the data, since the motivation for using learning technologies, transferring, and obtaining vast amounts of data is very time-consuming. Consequently, researchers have recently used AI-based deep learning methods for precise prediction.

The use of histological evaluation of tissue sections at different levels of magnification has supplanted the reliance on morphological characteristics seen by haematoxylin and eosin (H&E) staining as the primary method for a pathologist to suspect the presence of lymphoma. Machine learning has gained popularity in cancer research due to its ability to extract complex information from medical pictures. Multiple radiomic characteristics are derived from pictures; nonetheless, machine learning necessitates appropriate parameters, therefore demanding meticulous feature selection.

Nevertheless, machine learning still encounters some challenges, including:

1. The precision of the model is influenced by the calibre of the photos used throughout the training process. The accuracy of the results may be compromised when using low-resolution photographs. Several variables, including as the scanner's precision, the uniformity of slide fabrication, and the quality of the stain, might potentially affect the picture.
2. The extensive variety of diseases, tissues, cells, and antibodies that are accessible suggests that it may be difficult to establish a direct correlation between morphological and molecular data. We are now concentrating on one specific connection, but more effort is needed to apply our technique more broadly.
3. The determination of whether portions of an H&E-stained image include positive or negative cells can only be made by referring to the matching IHC-stained picture. Despite using IHC staining, accurately determining the level of positivity of a cell in an H&E stained image remains challenging, hence impeding precise inference of the model.

Deep learning (DL) has been a powerful technology in the last decade since it can directly extract characteristics from photos. The area of computer vision has advanced as a consequence. DL models need vast amounts of input data because to the intricate nature of the underlying layers. When training a highly complex network with a limited dataset, the probability of overfitting is much higher. Techniques like as data augmentation,

Table 2.1: Comparative performance analysis

Disease	Ref.	Methodology	Remarks	Drawbacks
HL and NHL	[23]	IF and Machine learning	This technique facilitates the concurrent observation of various lymphoma cells, promotes computational learning and identification, and assists in discovering therapies and enhancing the knowledge of lymphoma.	High error rate
	[24]	Supervised machine	By using these several techniques together, they create a robust and intelligent computational instrument. This tool assists physicians in comprehending the potential impact of Hodgkin's lymphoma on individuals.	Low range of accuracy
	[27]	PET-CT and Ann Arbor	PET-CT scans and the Ann Arbour staging system are essential instruments for detecting and staging lymphoma, facilitating treatment choices. They provide precise staging for certain lymphoma types, categorising patients into phases and informing successful treatment approaches.	Not a cost effective one
HL and NHL	[26]	EACCED machine learning	SEER's cause-specific death categorisation is a valuable prognostic instrument; nevertheless, its efficacy is contingent upon data quality and needs continuous development and validation.	Conventional algorithm makes the process a time consuming one
LM and NLM	[29]	Digital image analysis	Digital image analysis and deep learning methodologies are transforming lymphoma diagnosis by automating histological investigation, discerning intricate patterns, and minimising subjectivity, hence enhancing patient outcomes.	High training rate
T-cell and B- cell Lymphomas	[28]	AI models using CNN	AI models are used for detecting DLBCL and addressing obstacles in imaging, data gathering, and privacy, demonstrating excellent diagnostic accuracy and the possibility for improved patient outcomes.	High training rate and cost expensive
	[25]	J48	The research used the J48 machine learning algorithm and the WEKA platform to construct diagnostic algorithms, which may enhance the accuracy of lymphoma categorisation, underscoring the significance of dependable tools in medical research.	Conventional algorithm makes the process a time consuming one
HL and NHL	[26]	EACCEED machine learning	SEER's cause-specific death categorisation is a valuable prognostic instrument; nevertheless, its efficacy is contingent upon data quality and needs continuous development and validation.	Conventional algorithm makes the process a time consuming one
LM and NLM	[29]	AI using CNN	Digital image analysis and deep learning methodologies are transforming lymphoma diagnosis by automating histological investigation, discerning intricate patterns, and minimising subjectivity, hence enhancing patient outcomes.	Time consuming process
T-cell and B- cell Lymphomas	[28]	J48 algorithm	AI models are used for the diagnosis of DLBCL, addressing obstacles in imaging, data collecting, and privacy, while achieving high diagnostic accuracy and the promise for improved patient outcomes.	Highly expensive and need hardware support
	[25]		The research used the J48 machine learning algorithm and the WEKA platform to construct diagnostic algorithms, possibly enhancing the accuracy of lymphoma categorisation and underscoring the need of dependable tools in medical research.	Unable to analyze the drawback because the result range was not given properly

transfer learning, and cross-validation may be used to address issues such as overfitting and insufficient data sets. Utilising cross-validation to forecast model uncertainty is a prevalent practice within the AI safety field. Moreover, it is crucial to interpret DL-based findings in order to provide comprehensible results for human assessment. This is essential for evaluating the safety of AI and expediting the integration of DL in practical medical contexts. The determination of positive or negative cells in portions of an H&E-stained picture can only be made by referring to the matching IHC-stained image. Despite using IHC staining, accurately determining the level of positivity of a cell in an H&E stained picture remains challenging, hence impeding more precise inference of the model.

Hence here in order to overcome all the existing issues we implement the deep learning based stacked probabilistic attention neural network for the prediction of NHL.

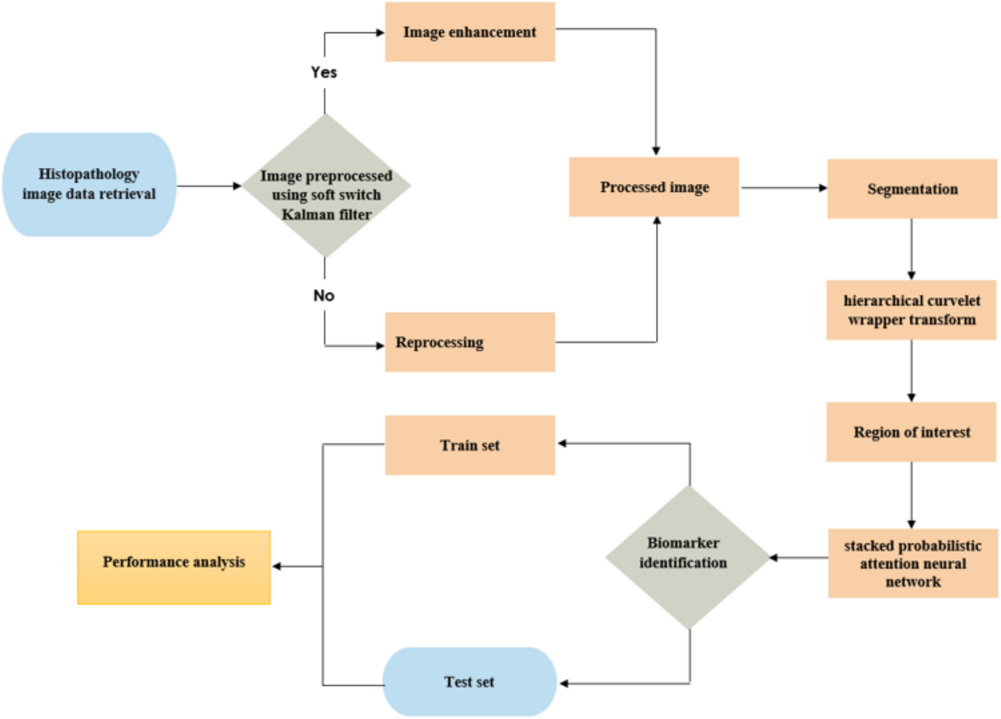


Image enhancement
Yes
Histopathology image data retrieval
Image preprocessed using soft switch Kalman filter
Processed image
Segmentation
No
Reprocessing
hierarchical curvelet wrapper transform
Region of interest
Train set
Performance analysis
Biomarker identification
stacked probabilistic attention neural network
Test set

Fig. 4.1: Schematic representation of the suggested methodology

4. Proposed methodology. Major advances in image processing and learning methodologies have not yet removed significant barriers to the development of quantitative imaging biomarkers for use in medical decision making.The recommended strategy for predicting the NHL-related biomarkers is shown in Figure 4.1.

4.1. Dataset. The dataset was extracted from the Kaggle https://www.kaggle.com/datasets/sparxi/ihc-images. It contains 28.8k IHC images for biomarker prediction.

4.2. Preprocessing. Images that have degraded and are noisy are filtered before being restored using various SSEF methods. The equivalent mathematical expression would be:

$$h(O,z) = f(O,z) * u(0,z) + n(O,z) \tag{4.1}$$

$$h(O,z) = R[h(O,z)] \tag{4.2}$$

The variables in the equation are defined as follows: f(O,z) represents the original input picture, u(O,z) represents the degradation function, ”*” denotes the error function, n(O,z) represents the noise (usually Gaussian noise), g(O,z) represents the deteriorated output image, and h(O,z) represents the final degraded output image after the application of procedure R. By using noise reduction filters that use nonlinear spatial domains, like the one seen in this example, it becomes feasible to reconstruct denoised images from noisy source images. Here are few methods to improve the quality of your photographs: The first step in the noise-reduction filter involves creating a mask matrix of dimensions nm. The mask pixel value and mask pixel size are used in the mask matrix to calculate the new pixel value for the degraded image. The filter assigns the value of each pixel to be equal to the element in the middle of the mask's matrix. This method has the capability to eliminate irregularities without compromising the quality of the image. The proposed filter use the average and standard deviation of the pixel values in the mask matrix.

$$\mu = \frac{1}{OM} \sum_{n,m \in \eta} a(0,m) \tag{4.3}$$

$$\sigma^2 = \frac{1}{oM} \sum_{0,m \in o} a^2(o,m) - \mu^2 \tag{4.4}$$

The mask's neighbourhood area has a size of nm., σ^2 is the variance of the Gaussian noise in the image, and $a(n,m)$ is the representation of each pixel in the mask. Next, the SSEF filter is created for the updated pixels using the expected values, which are represented as $b_w(o,m)$.

$$c_w(o,m) = \mu + \frac{\sigma^2 - v^2}{\sigma^2}.(a(0,m) - \mu) \tag{4.5}$$

wherev^2 is the mask matrix's noise variance setting when using the SSEF filter.
Now, the imputed pixel values are given by

$$c_i^{imp} = \sum_{j=1}^{k} w_j z_j, \ \ i = 1, .., m \tag{4.6}$$

After error removal and imputing the pixel values the error free images are obtained.

4.3. Segmentation. The HSPT segmentation algorithm may be fed the processed picture. We'll refer to the areas in S_i that contain the first seeds, or $B_1, B_2, , B_i$. $(\bar{O}, \bar{D}_b, \bar{D}_r)$ to show how the sum of all S_i seed pixels breaks down into O, D_b, andD_r. In this section, we outline our suggested method of segmentation.

(1) Choose your seeds automatically.
(2) Give each seed area a label.

The seed pixel, first, has to share a lot of characteristics with its surrounding pixels. Second, in order to construct the predicted area, at least one seed must be created. Third, it's important to keep seeds for various locations apart.

The following formula is used to calculate the degree of similarity between a given pixel and its neighbours. The dispersion measures of the Y, C_b, and C_r Using the, the components of a 3 × 3

$$\sigma_Y = \sqrt{\frac{1}{9} \sum_{i=1}^{9} (O_i - \bar{O})^2}, \tag{4.7}$$

Where O can be Y, D_b, or D_r, the mean value $\bar{Y} = \frac{1}{9} \sum_{i=1}^{9} x_i$. Standard deviation, on the whole, is

$$\sigma = \sigma_K + \sigma_{D_b} + \sigma_{D_\tau} \tag{4.8}$$

To get the standard deviation back inside the range $[0, 1]$, we,

$$\sigma_M = \frac{\sigma}{\sigma_{\max}}, \tag{4.9}$$

where $\sigma_{\max}$ is the image's greatest standard deviation. We may define a pixel's resemblance to its neighbours as

$$H = 1 - \sigma_M \tag{4.10}$$

The first requirement for the potential seed pixel is derived from the degree of similarity as follows.

The threshold similarity of a seed pixel candidate must be greater than 1.

The second step is to determine the $Y D_b D_r$ distances (relative Euclidean distances) between a pixel and its immediate neighbours.

$$d_i = \frac{\sqrt{(O - O_i)^2 + (D_b - D_{bi})^2 + (D_r - D_{r_1})^2}}{\sqrt{O^2 + D_b^2 + D_r^2}} \tag{4.11}$$

where $i = 1, 2, ...8$.

We determine the greatest possible separation between each pixel and its neighbours as,

$$d_{\max} = \max_{i=1}^{8}(d_i) \tag{4.12}$$

Create a list T of all the areas that are close by, then sort them by decreasing distance.

Remove the first point (p), even if T is not empty, and check to see whether any of its four neighbor's are empty. If all of p's labelled neighbor's have the same label, then p ought to get that label as well. If p's labelled neighbor's have different labels, p should be put in the area that is closest to it. The region's mean is then adjusted, and T is then expanded to include p's unclassified neighbor's in decreasing order of distance.

Using this method, we may get the fractional Euclidean distance, di, between pixel i and its neighbours.

$$d_i = \frac{\sqrt{(Z_i - \bar{Z})^2} + (D_{b_i} - D)^2 + (D_D - \bar{D}_r)^2}{\sqrt{Z_i + D_{b_i}^2 + D_{r_i}^2}} \tag{4.13}$$

where $(\bar{O}, \bar{D}, \bar{D}_r)$ are the medians of the distributions of Y, D_b, and D_r in the region. Pixel with the shortest distance value, p, is selected as the best one. If several neighbouring pixels have the same minimum value, we choose the one that best characterises the bigger of the two adjacent areas.

The red pixels indicate seeds, green pixels represent pixels in a sorted list T, white pixels represent the pixels with the shortest distance to the seed areas, white pixels are linked to the surrounding red region, and black pixels are added to which causes a recalculation of the mean of the new region and the distances between the new region and its neighbours. Once there is nowhere left where the distance is less than the criteria, we stop. What the distance between two points is in Euclidean space.

When discussing the colour differences between these regions, we use the labels R_i and R_j.

$$d_i = \frac{\sqrt{(\bar{O}_i - \bar{O}_j)^2 + (\bar{C}_D - \bar{C}_D)^2 + (\bar{D}_{r_i} - \bar{D}_{r_j})^2}}{\sqrt{O^2 + D_b^2 + D_r^2}} \tag{4.14}$$

After repeating the process the ROI can be separated.

4.4. Biomarker Identification. SPANN uses a direct influence on network structure data to identify unlabeled nodes by transmitting their labels across transfer and sink nodes. Equation 4.15 defines undirected graphs.

$$\zeta = (G, \epsilon) \tag{4.15}$$

where $G = G_1, G_2, , G_n$ represent nodes, $\epsilon = \epsilon_1, \epsilon_2, , \epsilon_n$ represent edges. Matrix adjacency ζ, A' may be calculated to determine whether two nodes are related

$$B_{ij}^{'} = \begin{cases} \alpha, & G_i \neq G_j \\ 1, & G_i = G_j \\ e^{K||G_i - G_j||}, & otherwise. \end{cases} \tag{4.16}$$

where, $\alpha_{mhsa_{i}j}$ value of 0.2 in the studies, demonstrating multi-head self-guided attention determines neighbour node weights.

We provide multi-tiered SPANN topologies. This data allows real-time adjacency matrix adjustments.

$$B^{'(r)} \leftarrow B(B^{'(r-1)} + \alpha V^{(r-1)} h^{(r)T}) B^T + \beta l \tag{4.17}$$

where $I^{(r-1)}$ represents biomarker-associated aspects of the $(r-1)$"th" layer's output; indicates the coefficient of correlation. $h^{(r)}$ indicates represents biomarker-associated aspects of the $(r-1)$"th" layer's output that is the coefficient of correlation

$$l^{(r)} = \delta(BX^r) \tag{4.18}$$

where δ signifies the softplus (.) activation function is engaged, and $\bar{B}$ may take the values specified by Equation (4.19).

$$\begin{cases} 1 + E^{-\frac{1}{2}} B' E^{-\frac{1}{2}} \tilde{\rightarrow} E^{\frac{1}{2}} \tilde{A} E^{-\frac{1}{2}} \\ \tilde{E_{ij}} = \sum_j \tilde{A_{ij}} \end{cases} \tag{4.19}$$

I is the identity matrix.

We use the stack attention module to reduce superfluous subspace pixel blocks and create the same pixel block with the completely connect pixel cut to carry different numbers of subpixel block nodes depending on size. To divide subpixel blocks optimally, compute the global and local property information gain. Equation 4.20 shows subpixel block formation.

$$\beta_{t,x} = \lambda \beta_{t,x}^{global} + \gamma \beta_{t,x}^{local}, \;\; \gamma = \frac{1}{2} - \lambda \tag{4.20}$$

where γ, λ indicates weighing factor $\beta_{t,x}^{global}$ and $\beta_{t,x}^{local}$ signals global or region attention $\beta_{t,x}^{global}$ and $\beta_{t,x}^{local}$ as:

$$\begin{cases} \beta_{t,x}^{global} = \frac{\exp(score(V_t, \bar{V_x}))}{\sum_{Y=1}^{T_Y} \exp(score(V_t, \bar{V_x}'))} \\ \beta_{t,x}^{local} = \frac{\exp(score(h_t, \bar{h_x}))}{\sum_{T_x \delta(v_p^T tanV(w_p v_t)) - E}^{T_x \delta(v_p^T tanV(w_p v_t)) + E} \exp(score(V_t, V_x'))} exp\left(- \frac{Y - T_Y . \delta(v_p^T tanV(W_p V_t))}{8E^2} \right) \end{cases} \tag{4.21}$$

where δ activates a function.

M distinct attention-directed adjacency matrices need M tightly connected layers. We alter each layer's calculation as stated below (for the $l"th"$ matrix $\bar{A}^t$) in equation 22 .

$$V_{ti}^{l} = \delta(\sum_{j=1}^{n} \bar{B}_{ij}^{t} W_t^{(l)} Z_j^{(l)} + b_t^{(l)}) \tag{4.22}$$

Focused adjacencies $\bar{A}^t.Z_j^{(l)}$, where t=1,...,M, and t selects the bias term and weight matrix related to $\bar{A}^t$.

$$Z_j^{(l)} = [X_j, V_j^{(l)}, ..., V_j^{(l)}] \tag{4.23}$$

L is the number of closely coupled layer sublayers. The stack's primary purpose is to split the super pixel block of flawlessly connected pixels into a smaller subspace to create an adjacency matrix. Since a connectionless edge's weight is set to 0, pruning is necessary. Cutting the completely connected super pixel block might destroy part-relevant information. Thus, we designed a self-attention guidance module to redistribute edge weight to the trimmed subspace pixel block, stressing graph node relationships and interactions, establishing a more dependable multi-scale graph structure, and addressing the problem.

The self-guided attention module transforms the multi-scale subspace pixel block into a totally linked graph using multi-head self-attention. While attention guidance builds an adjacency matrix A', edge weights are enhanced. Each A' represents a totally connected graph, and entry Aij' denotes the degree of connection between nodes i and j. Attention to build node relations allows the self-attention machine to record interactions between any two places in a single sequence. Equation 4.4 calculates A'.

$$\alpha_{SPANNs_ij} = \frac{exp(LeakyReLu(\bar{a}^T [w\bar{v_i} || w\bar{v_j}]))}{\sum_{k \in N_i} exp(LeakyReLu(\bar{a}^T [w\bar{v_i} || w\bar{v_k}]))} \tag{4.24}$$

where T represents the matrix transpose and w the node weights. Node i's neighbours, denoted by N_i, are i, LeakyReLu(.) indicates activate function.

For example, we may join the results of the recommended network, which has a fully-connected layer for the masked function, as.

$$FCN_{g_{st}} = SoftMax([HCN_{g_s0}, ...HCN_{g_st}] W_{fcn}) \tag{4.25}$$

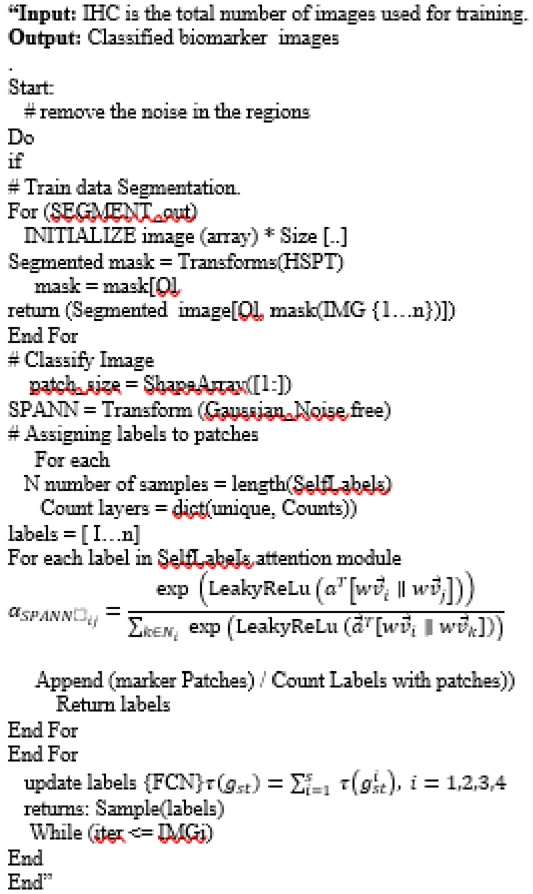
where $st = 0, 1, 2, 3, W_fcn$ represents layer weights when all nodes are linked.

$$\zeta(V_{st}) = \sum_{i=1}^{s} \zeta(V_{st}^{i}), \quad i = 1, 2, 3, 4 \tag{4.26}$$

where $\zeta(V_st)$ i denotes the cross-entropy error loss, which is a measure of how far a network's predictions deviate from the labels used to build the training set.

Algorithm 1 illustrates the implementation procedure of the SPANN.

Algorithm 1 SPANN



```
"Input: IHC is the total number of images used for training.
Output: Classified biomarker  images

.
Start:
  # remove the noise in the regions
Do
if
# Train data Segmentation.
For (SEGMENT_out)
  INITIALIZE image (array) * Size [..]
Segmented mask = Transforms(HSPT)
    mask = mask[O],
return (Segmented  image[O], mask(IMG {1...n})])
End For
# Classify Image
   patch_size = ShapeArray([1:])
SPANN = Transform (Gaussian_Noise free)
# Assigning labels to patches
    For each
  N number of samples = length(SelfLabels)
     Count layers = dict(unique, Counts))
labels = [ I...n]
For each label in SelfLabels attention module
$\alpha_{SPANN\square_{ij}} = \frac{\exp\left(\text{LeakyReLu}\left(a^T[w\vec{v}_i \parallel w\vec{v}_j]\right)\right)}{\sum_{k\in N_i} \exp\left(\text{LeakyReLu}\left(\vec{a}^T[w\vec{v}_i \parallel w\vec{v}_k]\right)\right)}$

    Append (marker Patches) / Count Labels with patches))
       Return labels
End For
End For
  update labels {FCN}$\tau(g_{st}) = \sum_{i=1}^{s} \tau(g_{st}^{i}), i = 1,2,3,4$
  returns: Sample(labels)
   While (iter <= IMGi)
End
End"
```

5. Performance analysis. Here, we provide empirical data that substantiates the efficacy of the suggested analytical methodology. In general, the tests were carried out in a Python environment. The parameters of the proposed solution for biomarker prediction are computed, and the system's efficiency is compared to that of current techniques.

Figure 5.1 depicts a visual representation of the sample input acquired from the Kaggle database.

The objective of preprocessing is to optimise the efficiency of the classifier by determining the most valuable set of features. In this scenario, the Gaussian error in the picture may be repaired, as seen in Figure 5.2.

The objective of HSPT picture segmentation is to assign a categorical label to every individual pixel. We are using pixel-level predictions to identify the Region of Interest (ROI) within the selected parts of the image. Figure 5.3 displays the segmented output.

SPANN was used to examine the IHC protein markers included in the datasets under scrutiny. Every point is an immunohistochemical (IHC) picture of a marker found in the NHL. Figure 5.4 displays illustrative

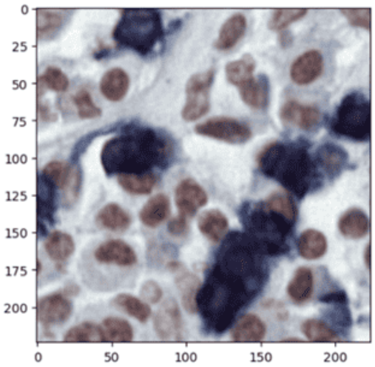


Fig. 5.1: Sample input

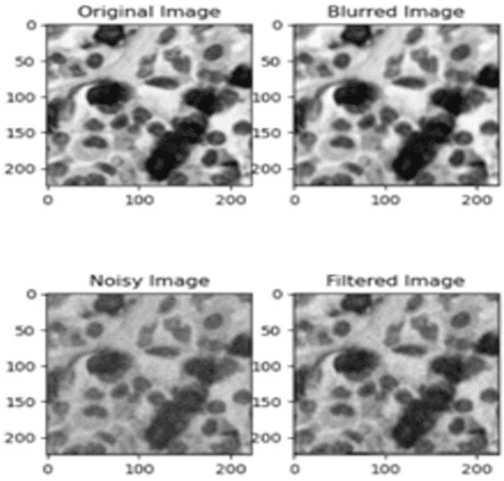


Fig. 5.2: Processed output

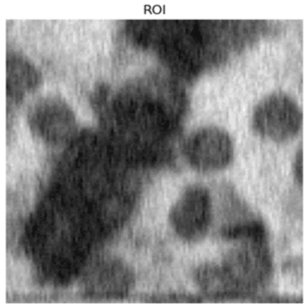


Fig. 5.3: Segmented output

examples of each marker picked at random. Figure 5.4 shows that the proposed technique can effectively separate and categorize the diverse group of testing sets spanning three distinct IHC markers.

Training accuracy and efficiency depend on epochs. Too few epochs may not provide the model enough

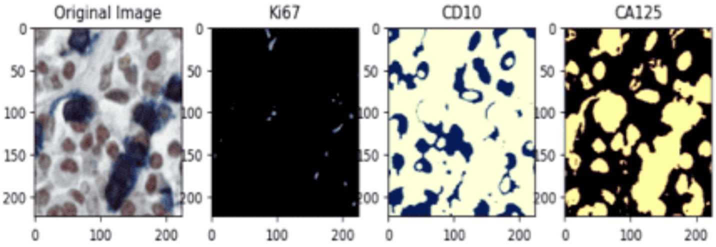


Fig. 5.4: Output classification

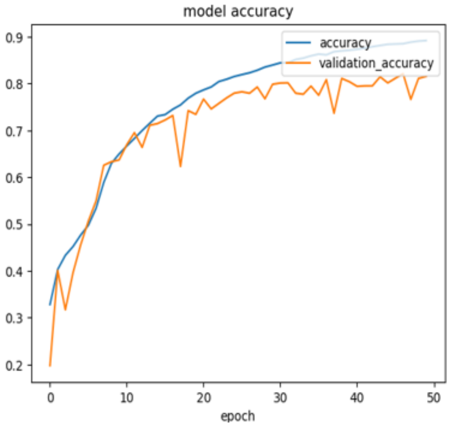


Fig. 5.5: Epoch vs. Accuracy

time to comprehend the data structure. Increase epoch size to 0.95 percent to increase accuracy. Similar binary classification solutions are assessed using confusion matrices. The confusion matrix evaluates categorization solutions by comparing predictions to reality. Displays false negatives, accurate forecasts, and incorrect predictions. Confusion matrix-based classifier assessment metrics may be constructed from this data. SPANN and learning-based models are evaluated using accuracy, recall, precision, F-measure, and AUC.

Accuracy. This heuristic performance metric predicts accuracy. Equation (5.1) calculates score by dividing total occurrences by accurate guesses.

$$Recall = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

Recall. It is sometimes referred to as sensitivity. Equation (5.2), which, when solved, gives the percentage of outcomes that were properly predicted when the result was positive, may be used to calculate this metric.

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

Precision. It is the ratio of accurately anticipated positive occurrences to the total number forecasted. Its formula can solve (29).

$$Precision = \frac{TP}{TP + FP} \tag{5.3}$$

F-measure. When class sizes are uneven, it is a common performance measure. This measure averages accuracy and recall scores, as stated in Equation (5.4).

$$F - Measure = 2 \times \frac{Prec \times Rec}{Prec + Rec} \tag{5.4}$$

Table 5.1: Performance analysis

S.no	Performance metrics	Performance ratio(%)
1	Accuracy rate	95%
2	Precision rate	95%
3	Recall rate	95%
4	F1 score rate	92%
5	AUC rate	99%

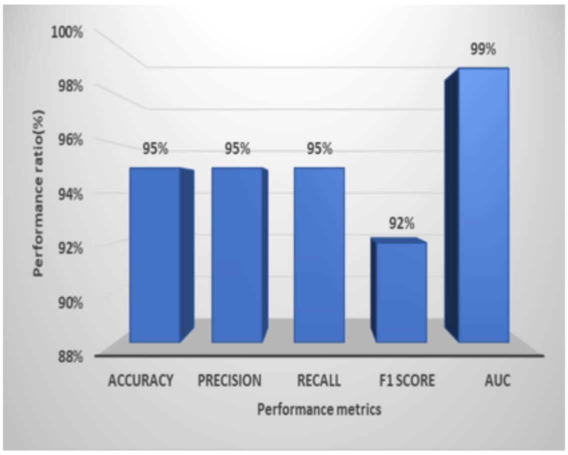


Fig. 5.6: Performance analysis of the suggested methodology

Area under the curve. Measures model categorisation accuracy. Equation (20) estimates the area under the receiver operating characteristic (ROC) curve for test evaluation.

$$AUC = \int TruepositivityR, d(FalPR) \tag{5.5}$$

TPR and FPR. Integrating the TPR with regard to the FPR yields the AUC score from the area under the ROC curve, which demonstrates the connection between these ratios.

There are multiple matches for performance evaluation methodology evaluation, including performance evaluation methods and performance evaluation process . Here we are evaluating our suggested mechanism with accuracy, precision, recall and F score. Table 5.1 and figure 5.6 show the methodology's performance. Comparing the proposed technique to known mechanisms helps assess its efficacy[22,11,21].

Divide the total of all true positives and negatives by the sum to get a classifier's accuracy. The suggested approach is 95 percentage more accurate than traditional practises (see Figure 5.7).

To obtain the precision for a given class, we divide the number of true positives by the classifier bias towards this class (number of times that the classifier has predicted the class). Figure 5.8 shows that HSPT and SPANN (95 percentage) outperform other biomarker prediction techniques.

Based on Figure 5.9, the suggested HSPT and SPANN technique has a recall of up to 95 percentage, which is far higher than existing approaches.

Commonly used as an evaluation metric in binary and multi-class classification , the F1 score integrates precision and recall into a single metric to gain a better understanding of model performance The proposed

Table 5.2: Comparative performance analysis

Methodology	Accuracy	Precision	**AUC**	F1	Recall
LR [22]	0.869	0.871	0.887	0.803	0.762
Adaboost [22]	0.808	0.714	0.806	0.722	0.747
Decision Tree [22]	0.812	0.790	0.806	0.708	0.665
Boost [22]	0.842	0.822	0.875	0.759	0.720
SVM [22]	0.849	0.932	0.890	0.747	0.63
Resnet 18 [11]	0.93	0.95	-	0.937	0.937
Proposed	**0.95**	**0.95**	**0.99**	**0.95**	**0.95**

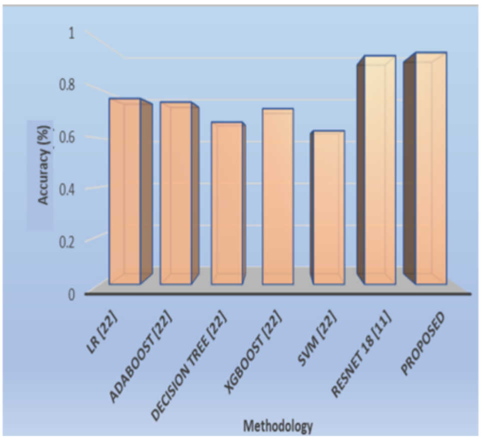


Fig. 5.7: Accuracy percentile analysis

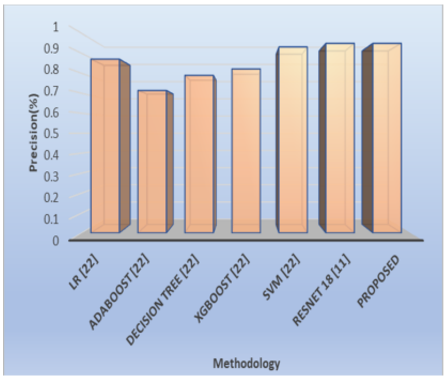


Fig. 5.8: Precision percentile analysis

HSPT and SPANN approach has a high rate of F1 score (95 percentage) compared to the current mechanisms, as can be shown in Figure 5.10.

Figure 5.11 depicts the two-dimensional ROC curve. The x-axis indicates positive rates, the y-axis shows true positive rates, and the threshold ranges from 0 to 1 (higher right to lower left). All threshold classification

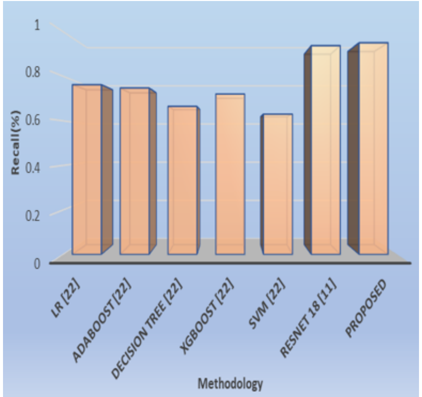


Fig. 5.9: Recall percentile analysis

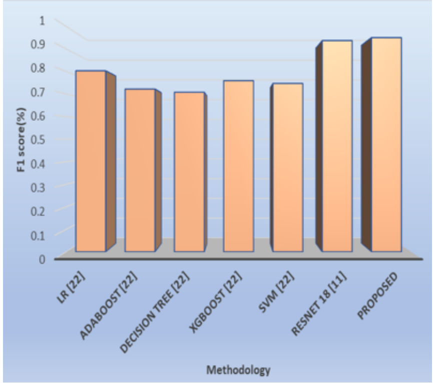


Fig. 5.10: F1 score percentile analysis

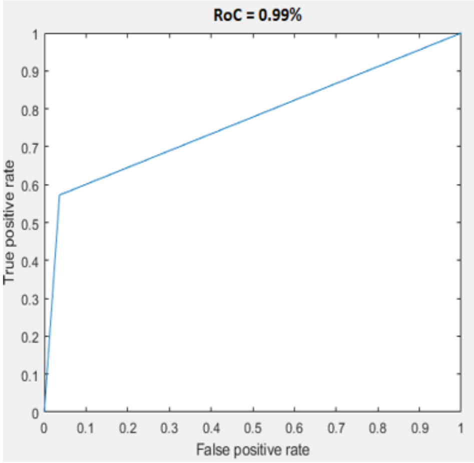
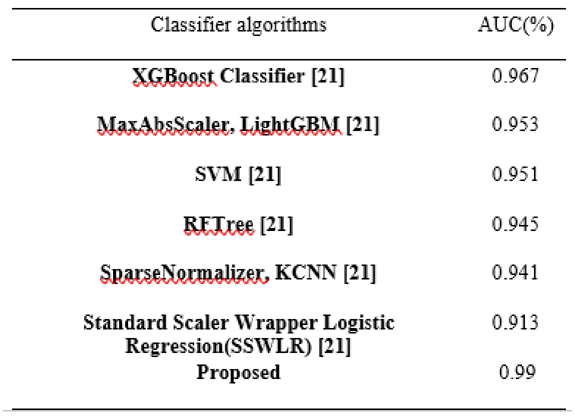


Fig. 5.11: ROC analysis

Table 5.3: AUC analysis



Classifier algorithms	AUC(%)
XGBoost Classifier [21]	0.967
MaxAbsScaler, LightGBM [21]	0.953
SVM [21]	0.951
RFTree [21]	0.945
SparseNormalizer, KCNN [21]	0.941
Standard Scaler Wrapper Logistic Regression(SSWLR) [21]	0.913
Proposed	0.99

results are graphed. An AUC of 99 percentage implies that the classifier is completely accurate.

Table 5.3 of the AUC performance measure comparison demonstrates that DL distinguishes biomarkers well. Our HSPT and SPANN models outperform gold standard methods. The table proves the suggested model is better than its competitors. The suggested model outperforms Tables 5.2 and 5.3. Compared to previous biomarker prediction mechanisms, the recommended technique yields satisfactory results".

6. Conclusion. Slides stained for Ki-67, CD 10, and CA125 were analysed to see whether they might be used to predict outcomes for NHL patients. To anticipate biomarker expression from H& E stained pictures without the need for IHC labelling, we developed an HSPT and SPANN model. Our findings demonstrate the close relationship between morphological and molecular data by demonstrating that histological pictures of tissue and cell morphologies have underlying molecular origins. Once this connection has been discovered, the abundance of a target protein may be predicted among the samples using a deep learning-based technique. The proposed strategy here significantly beat the state-of-the-art biomarker prediction mechanisms, by as much as 95 percentage. The following are where our future efforts will be concentrated. We need to enlarge our sample size to get more accurate results. By training the model on the new data, its resilience and generalization abilities will increase. To further generalize our findings, we recommend further trials on samples including a variety of tissues and stains; also, there is a suggestion for optimizing the model. Semi-supervised learning, for instance, may be used to reduce the burden of annotation. While our work demonstrated the capability of tumour histology to forecast pCR using DL methodologies and introduced a unique biomarker that serves as a more efficacious predictor than sTILs or subtype, it remains subject to certain limitations. This study used a restricted number of patients retrospectively for training and validation; hence, future research should aim for prospective multicenter investigations.

Acknowledgment. I would like to express profound gratitude to my guide for this invaluable support, encouragement, supervision and useful suggestions throughout this study. Her moral support and continuous guidance enabled me to complete my work successfully.

Author's Contributions. *Sivaranjini N.:* Designed, analysis and acquisition of data. *Gomathi M.:* Reviewed and Organized the study.

REFERENCES

[1] J. Carreras, Y. Y. Kikuti, M. Miyaoka, S. Hiraiwa, S. Tomita, H. Ikoma, et al., "A combination of multilayer perceptron, radial basis function artificial neural networks and machine learning image segmentation for the dimension reduction and the prognosis assessment of diffuse large B-cell lymphoma," AI, vol. 2, pp. 106-134, 2021.

[2] D. Vrabac, A. Smit, R. Rojansky, Y. Natkunam, R. H. Advani, A. Y. Ng, et al., "DLBCL-Morph: Morphological features computed using deep learning for an annotated digital DLBCL image set," Scientific Data, vol. 8, p. 135, 2021.

[3] I. S. Lossos, ”Diffuse large B cell lymphoma: from gene expression profiling to prediction of outcome,” Biology of blood and marrow transplantation, vol. 14, pp. 108-111, 2008.

[4] J. Perdiz Arrais and J. L. Oliveira, ”RecRWR: a recursive random walk method for improved identification of diseases,” BioMed Research International, vol. 2015, 2015.

[5] F. Aziz, A. Acharjee, J. A. Williams, D. Russ, L. Bravo-Merodio, and G. V. Gkoutos, ”Biomarker prioritisation and power estimation using ensemble gene regulatory network inference,” International Journal of Molecular Sciences, vol. 21, p. 7886, 2020.

[6] K. Vaishali, S. Shambharkar, R. K. Somkunwar, and R. R. Kolte, ”Augmented Ensemble Learning Model for Biomarkers Prioritization to Enhance Disease Identification Efficiency,” in 2023 6th International Conference on Information Systems and Computer Networks (ISCON), 2023, pp. 1-7.

[7] R. Hajjo, D. A. Sabbah, S. K. Bardaweel, and A. Tropsha, ”Identification of tumor-specific MRI biomarkers using machine learning (ML),” Diagnostics, vol. 11, p. 742, 2021.

[8] Musthafa, M. M., TR, M., V, V. K., & Guluwadi, S. (2024). Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification. BMC Medical Imaging, 24(1), 201

[9] Y. Chen, H. Yang, Z. Cheng, L. Chen, S. Peng, J. Wang, et al., ”A whole-slide image (WSI)-based immunohistochemical feature prediction system improves the subtyping of lung cancer,” Lung Cancer, vol. 165, pp. 18-27, 2022.

[10] T. E. Tavolara, M. K. K. Niazi, D. Jaye, C. Flowers, L. Cooper, and M. N. Gurcan, ”Deep learning to predict the proportion of positive cells in CMYC-stained tissue microarrays of diffuse large B-cell lymphoma,” in Medical Imaging 2023: Digital and Computational Pathology, 2023, pp. 12-16.

[11] Y. Liu, X. Li, A. Zheng, X. Zhu, S. Liu, M. Hu, et al., ”Predict Ki-67 positive cells in H& E-stained images using deep learning independently from IHC-stained images,” Frontiers in Molecular Biosciences, vol. 7, p. 183, 2020.

[12] C. Syrykh, A. Abreu, N. Amara, A. Siegfried, V. Maisongrosse, F. X. Frenois, et al., ”Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning,” NPJ digital medicine, vol. 3, p. 63, 2020.

[13] Chakravarthy, S., Nagarajan, B., Kumar, V. V., Mahesh, T. R., Sivakami, R., & Annand, J. R. (2024). Breast tumor classification with enhanced transfer learning features and selection using chaotic map-based optimization. International Journal of Computational Intelligence Systems, 17(1), 18.

[14] P. Ghahremani, Y. Li, A. Kaufman, R. Vanguri, N. Greenwald, M. Angelo, et al., ”Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification,” Nature machine intelligence, vol. 4, pp. 401-412, 2022.

[15] V. Bobée, F. Drieux, V. Marchand, V. Sater, L. Veresezan, J.-M. Picquenot, et al., ”Combining gene expression profiling and machine learning to diagnose B-cell non-Hodgkin lymphoma,” Blood Cancer Journal, vol. 10, p. 59, 2020.

[16] C. Beaulac, J. S. Rosenthal, Q. Pei, D. Friedman, S. Wolden, and D. Hodgson, ”An evaluation of machine learning techniques to predict the outcome of children treated for Hodgkin-Lymphoma on the AHOD0031 trial,” Applied Artificial Intelligence, vol. 34, pp. 1100-1114, 2020.

[17] J. Zhang, W. Cui, X. Guo, B. Wang, and Z. Wang, ”Classification of digital pathological images of non-Hodgkin's lymphoma subtypes based on the fusion of transfer learning and principal component analysis,” Medical Physics, vol. 47, pp. 4241-4253, 2020.

[18] J. Carreras and R. Hamoudi, ”Artificial neural network analysis of gene expression data predicted non-hodgkin lymphoma subtypes with high accuracy,” Machine Learning and Knowledge Extraction, vol. 3, pp. 720-739, 2021.

[19] Ahmed, S. T., Sivakami, R., Mahesh, T. R., Khan, S. B., Mashat, A., & Almusharraf, A. (2024). PrEGAN: Privacy Enhanced Clinical EMR Generation: Leveraging GAN Model for Customer De-Identification. IEEE Transactions on Consumer Electronics..

[20] J. J. Eertink, G. J. Zwezerijnen, S. E. Wiegers, S. Pieplenbosch, M. E. Chamuleau, P. J. Lugtenburg, et al., ”Baseline radiomics features and MYC rearrangement status predict progression in aggressive B-cell lymphoma,” Blood Advances, vol. 7, pp. 214-223, 2023.

[21] García, R., Hussain, A., Chen, W., Wilson, K., & Koduru, P. (2022). An artificial intelligence system applied to recurrent cytogenetic aberrations and genetic progression scores predicts MYC rearrangements in large B-cell lymphoma. EJHaem, 3(3), 707-721.

[22] Hao, P., Deng, B. Y., Huang, C. T., Xu, J., Zhou, F., Liu, Z. X., ... & Xu, Y. K. (2022). Predicting anaplastic lymphoma kinase rearrangement status in patients with non-small cell lung cancer using a machine learning algorithm that combines clinical features and CT images. Frontiers in Oncology, 12, 5627.

[23] Bharanidharan, N., Chakravarthy, S. S., Venkatesan, V. K., Abbas, M., Mahesh, T. R., Mohan, E., & Venkatesan, K. (2024). Local entropy based remora optimization and sparse autoencoders for cancer diagnosis through microarray gene expression analysis. IEEE Access..

[24] Thakur, A., Gupta, M., Sinha, D. K., Mishra, K. K., Venkatesan, V. K., & Guluwadi, S. (2024). Transformative breast Cancer diagnosis using CNNs with optimized ReduceLROnPlateau and Early stopping Enhancements. International Journal of Computational Intelligence Systems, 17(1), 14..

[25] Mahesh, T. R., Vinoth Kumar, V., Vivek, V., Karthick Raghunath, K. M., & Sindhu Madhuri, G. (2024). Early predictive model for breast cancer classification using blended ensemble learning. International Journal of System Assurance Engineering and Management, 15(1), 188-197..

[26] Z. L. . Huan Wang , ”Using Machine Learning to Expand the Ann Arbor Staging System for Hodgkin and Non-Hodgkin Lymphoma,” BioMedInformatics, p. 12, 2023.

[27] Bruce D. Cheson et al, ”Recommendations for Initial Evaluation, Staging, and Response Assessment of Hodgkin and Non-Hodgkin Lymphoma: The Lugano Classification,” Journal Of Clinical Oncology, vol. 32, p. 10, 2024.

[28] Mahesh, T. R., Geman, O., Margala, M., & Guduri, M. (2023). The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification. Healthcare Analytics, 4, 100247..

[29] Andy N.D. Nguyen, Kareem A. Allam, ”Deep Learning for Digital Image Analysis with Whole Slide Imaging for Lymphoma Diagnosis: Challenges and Promises,” 21st Century Pathology, p. 10, 2022.

Edited by: Dhilip Kumar V
Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing
Received: Jun 16, 2024
Accepted: Oct 2, 2024