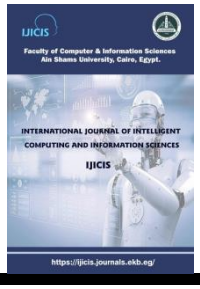
IJICIS, Vol.23, No.3, 73-82
DOI: 10.21608/ijicis.2023.211720.1275

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



Larg-scale Histopathological Colon Cancer Annotation Model Using Machine Learning Techniques

Esraa A.-R. Hamed*

Department of Basic Science,
Faculty Computer and Information Sciences, Ain Shams University,
Cairo,Egypt
Esraa.raouf@cis.asu.edu.eg

Mohammed A.-M. Salem

Department of Image and Vision Computing,
Media Engineering and Technology, GUC,
Cairo,Egypt
Mohammed.Salem@guc.edu.eg

Nagwa L. Badr

Department of Information Systems,
Faculty Computer and Information Sciences, Ain Shams University,
Cairo,Egypt
Nagwabadr@cis.asu.edu.eg

Mohamed F. Tolba

Department of Scientific Computing,
Faculty Computer and Information Sciences, Ain Shams University,
Cairo,Egypt
Fahmytolba@cis.asu.edu.eg

Received 2023-05-17; Revised 2023-08-01; Accepted 2023-08-01

Abstract: *Colon cancer ranks among the leading factors contributing to mortality and morbidity among adults. One of the main components in determining the kind of cancer is the histopathological diagnosis. This study presents the development of a computer-aided diagnosis system for adenocarcinomas of the colon using machine learning (ML) to analyze digital pathology images. A dataset of 10,000 images was gathered from the LC25000 collection, with 5000 images for each class. The Convolutional Neural Network with a Light Gradient Boosting Machine (CNN-LightGBM) with multiple threads was used as the classification model, and the system was evaluated against other ML algorithms. The reported diagnosis accuracy for colon cancer has achieved greater than 90%, outperforming the latest ML algorithms in disease classification accuracy. However, the accuracy was less than that for lung cancer classification based on this approach. This study demonstrates the potential for ML to improve the accuracy and efficiency of medical diagnosis and highlights the need for further research to improve the accuracy of colon cancer diagnosis.*

Keywords: *Colon cancer, Convolutional Neural Network, Deep Learning, Machine Learning, LightGBM*

1. Introduction

As per the World Health Organization (WHO), cancer stands as the second most prevalent cause of death worldwide. 9.4% of deaths are attributed to colorectal cancer [1]. Globally, the incidence of malignant

***Corresponding Author**: Esraa A.-R. Hamed

Basic Science Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

Email address: Esraa.raouf@cis.asu.edu.eg

tumors has been rising, which may be connected to population expansion. Malignancy can affect any age group depending on the histological type; however, it is frequently found in people who are over 50 to 60 [2]. According to predictions, cancer mortality might increase to 60% by 2035 [3]. The colon, the last section of our digestive system, may develop colon cancer if it contains malignant cells. Although not age-related, older people often get colon cancer. On the internal side of the colon, tiny polyps—clumps of cells that are not malignant (benign)—typically occur at the beginning. Some of these polyps may eventually develop into colon cancer [2].

In the majority of colon cancer cases, a tumor develops when normal cells in the colon or rectum lining grow out of control. Adenocarcinomas of the colon or rectum begin as epithelial cells in the lining of the large intestine. They later expand to other layers. Despite being less common subtypes of adenocarcinomas, signet ring cell adenocarcinomas and mucinous adenocarcinomas are aggressive and challenging to treat. The human body can change over time based on factors including gender, race, age, smoking habits, and socioeconomic status. However, if a person has a rare genetic disease, the alterations may occur quickly—within a few months [3].

Several applications, including language processing, image processing, and audio processing, may now be successfully integrated into Deep Learning. Since it serves as the foundation of ML, this has advanced significantly. Recent advancements [4] suggest that deep learning might be used to analyze medical images, including those from computed tomography, whole-slide imaging, and magnetic resonance imaging. Due to the availability of digital pathology datasets to the general public, researchers may now assess the viability of Deep Learning algorithms. This will increase the effectiveness and accuracy of histological inspection.

Our research's main objective is to determine whether features extracted from histopathological images can be used to distinguish colon cancer cells from healthy cells. This is done using the proposed CNN-LightGBM model. Furthermore, the attained outcomes were contrasted with alternative ML methods, including Support Vector Machines (SVM), K Nearest Neighbor (KNN), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost).

Section 2, presents a literature review on colon cancer classification, followed by an examination of available histopathology colon cancer datasets. Additionally, it outlines the data pre-processing steps applied to the selected dataset to prepare it for training with the proposed CNN model in Section 3. Section 4 explores into the CNN-LightGBM model for colon cancer histopathology detection and classification. The findings of our experiments are discussed in Section 5, where we compare the classification performance of the proposed CNN-LightGBM model with existing ML models. Finally, Section 6 concludes the work with a summary and presents recommended next steps.

2. Related Work

Artificial intelligence (AI) is the process of enhancing human intelligence in computer software to enable communication with computers that resemble humans. Recent developments have made AI, which is used in many computer vision disciplines, the most significant science of the twenty-first century.

This work utilizes the LC25000 histopathological imaging dataset of colon and lung cancer, which was published in 2020, to analyze the proposed technique. Throughout this section, we emphasize the contributions of numerous researchers who have employed this dataset to develop applications based on deep learning.

A multi-modal Sparse Representation-based Classification (mSRC) technique for diagnosing lung cancer was described by authors in [5]. Their investigation used needle biopsy samples to automatically segment

4372 cell nuclei regions for lung cancer diagnosis. Their approach has an average classification accuracy of 88.10%. A technique for classifying nodule malignancy based on Multi-crop CNN (MC-CNN) was described by authors in [6]. On the CT scan images they utilized, they did not apply any feature extraction or segmentation techniques. Instead, they just used their ML model, and their classification accuracy was 87.14%.
Authors in [7], extracted four distinct types of characteristics from a series of histological colon images and subsequently employed three different variants of Support Vector Machines (SVMs) for image classification. Instead of using a single-level classification, they used a multi-label classification to find several cancer kinds that were concealed in various image regions. A deep-learning technique that automatically detects polyps from colonoscopy video has been put out by authors in [8]. For classification, they utilized the renowned CNN-based architecture AlexNet, achieving a classification accuracy of 91.47%. In trials using the LC25000 and Colorectal Adenocarcinoma Gland (CRAG) datasets to train and categorize histological images, ResNet-50 (96.77%) showed the best sensitivity, followed by ResNet-30 (95.74%) and ResNet-18 (94.79%) [9]. A CNN model achieved accurate identification of lung cancer images, with training and validation accuracy reaching 96.11 percent and 97.2 percent, respectively, while utilizing cross-entropy as the loss function [10].

Among all the models developed in our study [11], the proposed CNN-LightGBM model demonstrates the highest accuracy while utilizing the fewest total parameters. This model comprises merely four convolutional layers, four maximum pooling layers, and one leaky layer. Arguably, the CNN-LightGBM model, employing multiple threads, outperforms most existing models, achieving an impressive accuracy of up to 99.6% in a mere three seconds and with the least number of parameters needed for lung tissue identification and classification.

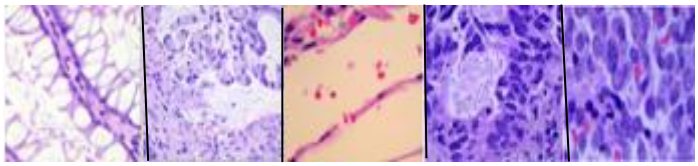
Additionally, in the context of colon cancer datasets, CNN-LightGBM exhibits superior performance compared to state-of-the-art ML methods, attaining an accuracy rate of 90%.

3. The Dataset

The LC25000 Lung and Colon Histopathological Image collection contains a total of 5000 images, encompassing various types of lung and colon cancer. This dataset has been validated for accuracy and compliance with HIPAA regulations [12]. Out of the 750 original images collected, 250 were assigned to each category, and all images have a resolution of 1024 x 768 pixels. The actual size of each image is 768 x 768 pixels. Figure 1 illustrates the five types present in the dataset: Lung Squamous Cell Carcinoma, Lung Adenocarcinoma, Benign Lung Tissue, Benign Colonic Tissue, and Colon Adenocarcinoma.

Colon adenocarcinomas, the most common type of colon cancer, account for over 95% of occurrences. They develop when an adenoma polyp forms within the large intestine and transforms into cancer. Lung adenocarcinomas, which constitute around 60% of all lung cancer cases, typically originate in the glandular cells on the lung's outer surface and progress to the alveoli. On the other hand, lung squamous cell carcinoma, the second most prevalent lung cancer type, initiates in the bronchi or airways of the lungs, making up approximately 30% of all cases [9].

For this study, a total of 10,000 histopathological images representing two types of colon tissue, namely Colon Adenocarcinoma and Benign Colonic Tissue, were gathered from the LC25000 collection.



(a) (b) (c) (d) (e)

Figure. 1: Samples of LC25000 dataset; (a) Colon Benign, (b) Colon Adenocarcinoma, (c) Lung Benign, (d) Lung Adenocarcinoma, (e) Lung Squamous Cell Carcinoma

4. The Proposed Architecture

In this section, we provide a comprehensive elucidation of the proposed CNN-LightGBM model with multiple threads, designed to categorize colon cancer images utilizing the LC25000 histopathological colon cancer dataset. The model architecture is divided into two steps, feature extraction and image classification, as seen in Figure 2.

During the preprocessing step, the model performs several operations, including converting images into BGr2RGB format, transforming them into a NumPy array, conducting feature scaling, and labeling the images. Subsequently, when images are fed into the proposed CNN feature extraction model, relevant characteristics are extracted from each image. The final stage of the proposed framework involves creating a LightGBM classifier with multiple threads, incorporating all the extracted features across four threads. This boosting method enables us to train our recommended network to categorize various types of colon histology images in a scalable and highly effective manner.

Finally, for the testing phase, the data is divided into benign and malignant colon cancer categories using the trained model parameters.

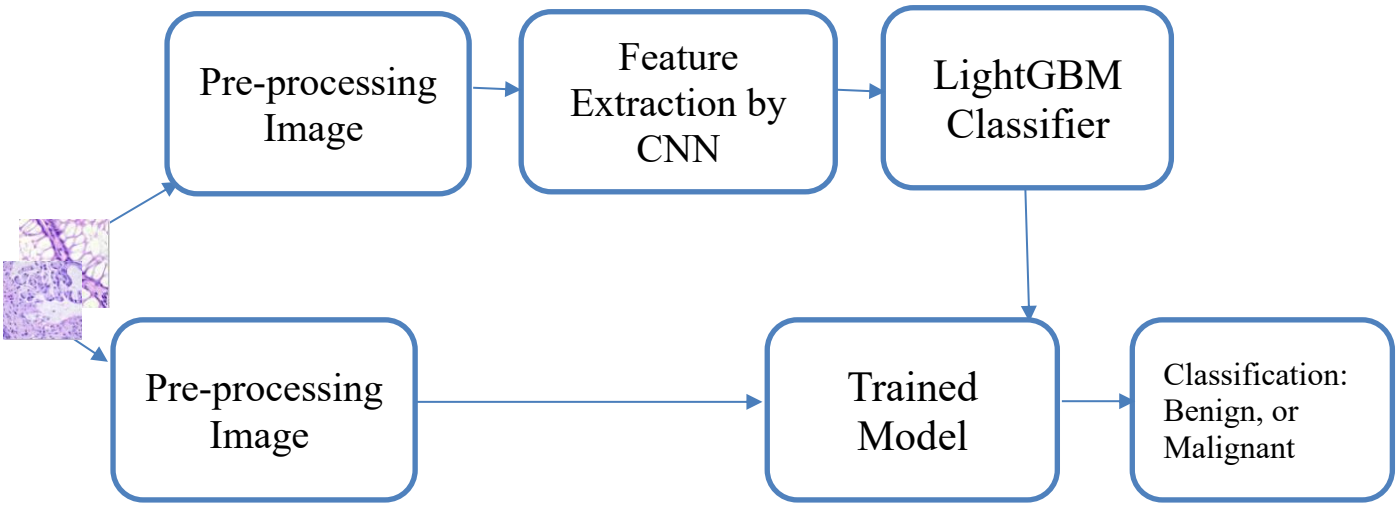


Figure. 2: The proposed Archticture

4.1 The Proposed CNN Feature Extraction Model

In this section, we present a detailed description of the proposed CNN feature extraction model. We carefully examined the pre-trained CNNs, such as AlexNet, VGG, etc., and created our own CNN model with fewer number of parameters to accelerate the processing time as we had previously identified [11].

Multiple layers make up a convolution neural network (CNN), which is used to extract features. As shown in Figure 3 the used model has four convolution layers (CL) that have kernel values of 32, 64, 128, and 256. The first CL layer used a kernel size of (11x11), whereas (3x3) was used for the second, third, and fourth layers. Following each convolutional layer, a max-pooling layer with a (2x2) kernel size is applied.

The feature vector was created for the Fully Connected (FC) layers by a Flatten Layer, following the final CL layer. Two fully connected layers were used with nodes 1024, and 512 neurons. Each convolution layer in the model uses ReLU as an activation function, while Softmax is used after the output layer. The dropout layer is applied with a rate of 0.4.

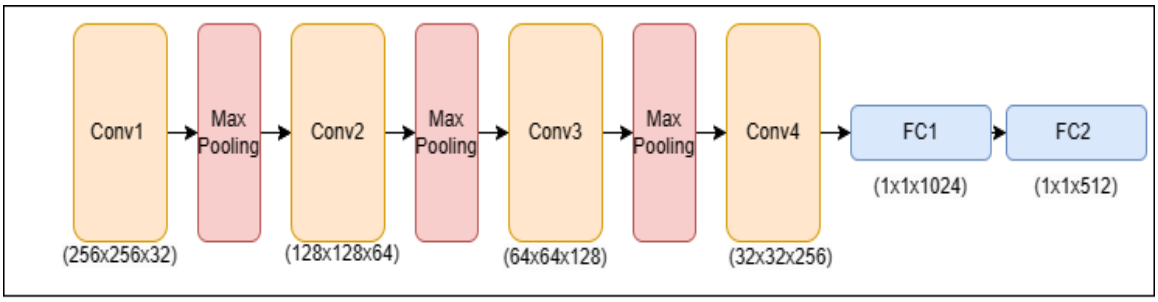


Figure. 3: The proposed CNN Architecture

4.2 The LightGBM Model

Microsoft released the distributed gradient boosting framework LightGBM, whose computation performance is ten times faster than the original GBDT approach. It only requires a third of memory. This enables researchers to apply the XGBoost approach more effectively. The LightGBM model incorporates two crucial enhancements: the leaf-wise technique with depth limitation and the histogram algorithm. The histogram algorithm creates a histogram by partitioning continuous data into K integers, where the discretized value is stored in the histogram as an index during traversal, facilitating the search for the optimal split point for the decision tree.
To optimize the LightGBM model's hyperparameters, a grid search approach is employed after the model has been trained using the gradient boosting method. The LightGBM model's primary hyperparameters are 200 weak regression trees, 50 leaves, 0.1 learning rate, and 2000 iterations. These are the hyperparameters optimized by grid search [13].

A strong regression tree is created by linearly combining M weak regression trees [14]. The following calculating formula:

$$F(x) = \sum_{m=1}^{M} f_m(x) \qquad (1)$$

where F(x) is the final output and f_m (x) is the output of the m^{th} weak regression tree.

Before constructing a decision tree, it is crucial to determine the optimal segmentation point. The conventional approach involves sorting feature values and enumerating accessible feature points, which can be memory-intensive and time-consuming. However, the LightGBM algorithm introduces an enhanced histogram approach. It selects division points among k values to split continuous eigenvalues into k intervals, leading to improved training time and space efficiency compared to the GBDT algorithm.

Additionally, the histogram approach incorporates regularization to counter overfitting, enhancing the decision tree's classification performance. To further reduce training data, the LightGBM technique employs a leaf-wise generation strategy, which can achieve greater loss reduction compared to the traditional level-wise method, even when growing the same leaf. Moreover, an additional parameter is employed to limit the depth of the decision tree and prevent overfitting. [15].

5 The Experimental Work and Results

The current study employed the CNN-LightGBM method to categorize images of benign colon tissue and adenocarcinoma of the colon using the LC25000 colon histology images dataset, which contains 5000 images in each class. The testing was conducted using Google Colab [16] and a Python 3 Google Compute Engine backend (GPU) with 12.68 GB of RAM and 78.19 GB of disk space.

5.1 The Evaluation Measures

The confusion matrix plot in Figure 4 was used to evaluate the model's performance, and equations (2), (3), (4), and (5) were used to compute the metrics accuracy, F1-Score, sensitivity, and specificity.

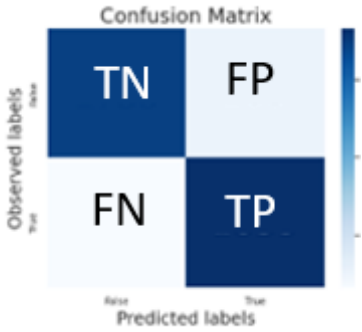


Figure. 4: The proposed Archticture

$$Accuracy = \frac{TP+TN}{TN+FP+FN+TP}, \quad (2)$$

$$F1 - Score = \frac{2\times TP}{(2\times TP)+FP+FN}, \quad (3)$$

$$Sensitivity = \frac{TP}{TP+FN}, \quad (4)$$

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

There are often only two groups I have, benign and malignant (binary categorization). Therefore, a binary classifier's confusion matrix evaluates the following four factors:

True Positives (TP): These are situations where the data point's expected and actual classes are both 1 (True).

True Negatives (TN): In these situations, both the actual class of the data item and the expected value are 0 (False).

False Positives (FP): These are cases where the data point's expected class, 1 (True), differs from its actual class, 0 (False).

False Negatives (FN): These are situations where a data point's expected value is 0 but its actual class is 1 (True).

5.2 Evaluation and discussion

In this study, the proposed CNN-based model was employed to classify the LC25000 Colon Histopathology Images as either benign or malignant colon cancer. Two separate trials were conducted using Google Colab [16], with each trial utilizing a different amount of data for the training and testing sets. The model was trained for 50 epochs in each trial.

In the first experiment (Exp.1), the proposed CNN-LightGBM method with multiple threads was evaluated using a batch size of 150, with 40% of the dataset allocated to the training set and 60% to the testing set. The research also compared this strategy with several ML models, such as KNN, SVM, RF, and XGBoost. The suggested method achieved a maximum accuracy of 89%, as indicated in Table 1.

Table 1 The accuracy of the proposed model compared with other ML models in Exp.1

Proposed CNN+ML	**Accuracy (%)**
KNN	78.7
SVM	84.3
RF	83.6
XGBoost	86.4
LightGBM multi thread	89

In the second experiment (Exp.2), batch size 150 was used, and the training set received 60% of the dataset while the testing set received 40%. The suggested CNN-LightGBM model with multiple threads was compared to other ML methods, including KNN, SVM, RF, and XGBoost. The proposed model, which is shown in Table 2, has a maximum accuracy of 90%.

Table 2 The accuracy of the proposed model compared with other ML models in Exp.2

Proposed CNN+ML	**Accuracy (%)**
KNN	*81.6*
SVM	*85.5*
RF	*86.7*
XGBoost	*88.2*
LightGBM multi threads	*90*

Figure 5 displays the accuracy chart for feature extraction and classification models. To assess classification performance, the suggested CNN-LightGBM multiple thread model was compared with various ML models, such as KNN, SVM, RF, XGBoost, and LightGBM with multiple threads. The results demonstrated that the suggested CNN-LightGBM model outperformed the other models, including the most advanced ones, in terms of both feature extraction and classification accuracy.

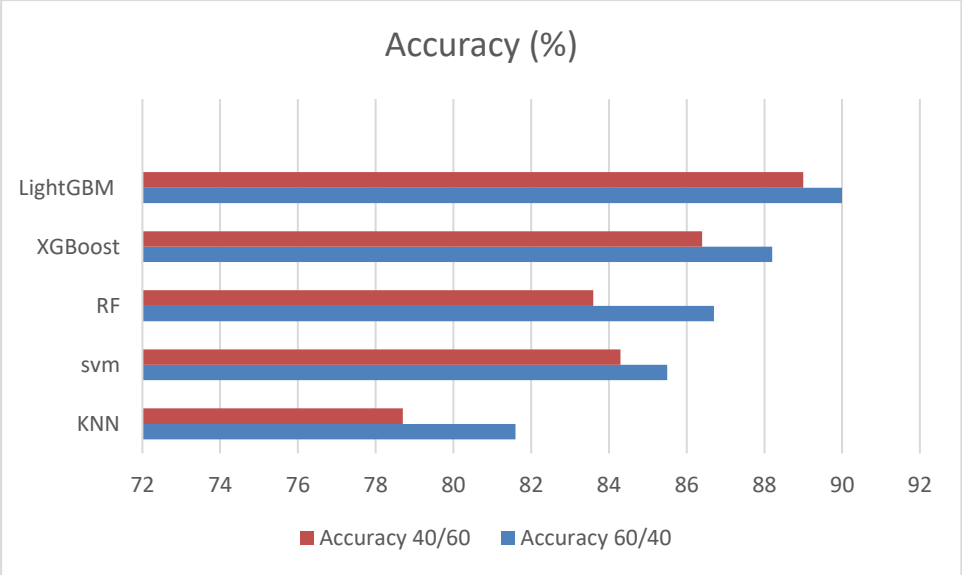


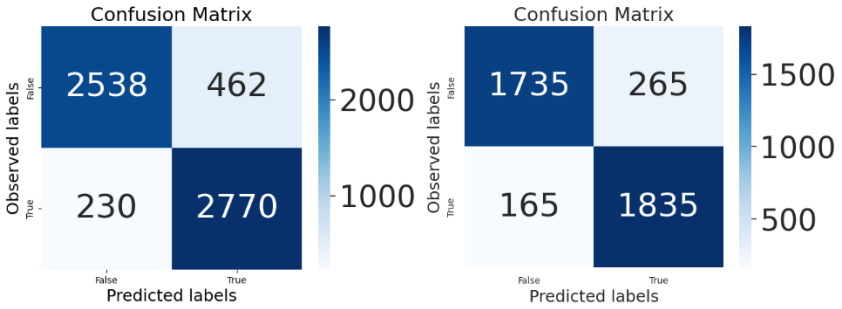
Figure. 5: The accuracy of the proposed CNN model compared with other ML models for classification

Table 3 Evaluation metrics of the proposed CNN-LightGBM model in two applied experiments

CNN+ML	Accuracy (%)	F1-Score (%)	Sensitivity (%)	Specificity (%)
Exp.1	89	89	93	85
Exp.2	90	90	92	87

We used a variety of evaluation metrics to assess the effectiveness of the proposed CNN-LightGBM multiple thread technique for histological colon classification. As seen in Table 3, the suggested approach attained accuracy, F1-Score, sensitivity, and specificity of 90%, 90%, 92%, and 87%, respectively, when using a training to testing dataset ratio of 60:40%. Furthermore, the proposed approach demonstrated a higher accuracy of 99.6% when applied to lung cancer histopathological images [11] compared to colon

cancer histopathological images.



(a) Exp1 (b) Exp2

Figure. 6: The confusion matrix display of the proposed model for each training to testing percentage datasets

In Figure 6, the confusion matrix for each trial is presented, utilizing the same dataset. This matrix allows for a comparison between the actual and predicted labels for the images in the test data categories. Evaluating the proposed CNN-LightGBM model for different training-to-test percentage datasets reveals that ML models, particularly the LightGBM model, exhibit effectiveness in categorizing various colon cancer subtypes. The trial results demonstrate the promise of this approach, even though there is scope for further refinement.

6 Conclusions and Future Work

Colon cancer is a major cause of death globally, and its histopathological diagnosis is crucial for guiding treatment decisions. Deep learning algorithms have been employed to aid pathologists and speed up the recognition of colon cancer. This paper focuses on the feature extraction and classification of colon cancer histopathology images through a novel CNN-LightGBM technique. The proposed technique is evaluated using the Colon Histopathological Imaging Dataset (LC25000), which contains 5000 histological images for every colon type.

The proposed CNN-LightGBM approach exhibits remarkable performance in classifying colon cancer datasets, surpassing state-of-the-art ML techniques with an accuracy rate of 90%. Notably, the CNN model achieved the lowest training parameter out of one million, while the LightGBM classifier outperformed other ML models. These findings suggest that the LightGBM tree model is simpler than other ensemble learning strategies and enhances efficiency by effectively combining multiple classifiers.

The experimental results demonstrate that the proposed technique surpasses comparable cancer diagnostic methods in terms of time efficiency. The computer-based identification approach can assist pathologists in diagnosing more patients with colon and lung cancer, with reduced effort, expenses, and hospitalization time. The CNN-LightGBM approach can serve as a benchmark for future research in histopathological cancer classification.

Moreover, when applied to lung cancer histopathological images, the proposed approach achieved a higher accuracy of 99.6% compared to its performance with colon cancer histopathological images. This observation suggests that detecting colon cancer is comparatively more challenging than detecting lung cancer. Future research efforts can concentrate on enhancing the suggested model's performance for colon

histopathological images and exploring its applicability in classifying histological images of other types of cancers.

References

1. Hyuna S. et al., "Global cancer statistics 2020: GLOBOCAN estimates of prevalence and death rates worldwide for 36 cancers in 185 countries," CA Cancer J Clin, vol. 71, no. 3, 209-249, 2021.
2. A. Necmi et al., "Analysis of cases of cancer from Dicle University Hospital ten decades' experience", 2018; 9; 102–106; Journal of Clinical and Analytical Medicine.
3. I. Soerjomataram et at.,"Global trends in cancer of the colon mortality: estimates to 2035." Journal of international cancer society,2019, 144(12), 2992-3000.
4. Khan, M. et al.,"a deep learning method for the automated diagnosis and multi-class categorization of Alzheimer's disease stages using resting-state fMRI and residual neural networks". Journal of medical systems, 2020, 44, 1-16.
5. Gao, Y. et al., "Multimodal sparse representation-based categorization for lung biopsy needle images", IEEE Transactions on Biomedical Engineering, 2013, 60(10), 2675–2685.
6. Yang, F. et al., "Multi-crop convolution neural networks for lung nodule cancer suspiciousness classification", Pattern Recognition, 2017, 61, 663-673.
7. Wei, J. et al., "Multi-label categorization of histological images for colon cancer", Microscopy Research and Technique, 2013, 76(12), 1266–1277.
8. Mokkapati, D. et al., "Automatic polyp recognition in colorectal videos." Image Processing, 2017, Vol. 10133, pp. 718–727.
9. Bokhari, S. et al., "The histological diagnosis of colonic cancer by using partial self-supervised learning", 2020, MedRxiv, *pp. 2020–2028*.
10. Thapa, H. et al., "Lung cancer recognition using CNN on histopathology images", Int. J. Comput. Trends Technol, 2020, 21-24.
11. E. A.-R. Hamed et al., “An Efficient Combination of Convolutional Neural Network and LightGBM Algorithm for Lung Cancer Histopathology Classification,” Diagnostics, vol. 13, no. 15, p. 2469, 2023.
12. Wilson, C. P. et al., "Lung and colon cancer histopathology image collection (lc25000)", 2019, ArXiv paper arXiv:1912.12142.
13. G. Ke et al., “Lightgbm: A highly efficient gradient boost decision tree,” Adv Neural Inf Process Syst, vol. 30, 2017.
14. H. Wang and coworkers, "LightGBM technique and the differential evolution algorithm-based multi-objective optimization design for DS-APMM," IEEE Transactions on Energy Conversion, vol. 36, no. 1, 2020, pp. 441-455.
15. 15. "A CNN-LSTM-LightGBM based short-term wind energy forecasting method based on attention mechanism," Energy Reports, vol. 8, pp. 437-443, 2022.
16. E. Bisong, Building machine learning and deep learning models on Google cloud platform. Springer, 2019.