CLASSIFICATION OF LUNG AND COLON CANCER TISSUES USING HYBRID CONVOLUTIONAL NEURAL NETWORKS

Chilyatun Nisa'[1)], Nanik Suciati[2)], dan Anny Yuniarti[3)]
[1, 2,3)] Departemen Teknik Informatika, Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia 60111
e-mail: nchilyatun@gmail.com[1)], nanik@if.its.ac.id[2)], anny@if.ac.id[3)]

ABSTRACT

Colon and lung cancers are two highly lethal kinds of cancer which can often coexist and pose a new challenge for accurate diagnosis. While research often concentrates on detecting a single cancer in a specific organ, this study proposes an innovative machine-learning approach to identify both colon and lung cancers. The objective is to create a hybrid machine learning classification model to enhance diagnostic precision. The LC25000 dataset comprises 25,000 color histopathological image samples of lung and colon cell tissues, indicating the presence or absence of cancer (adenocarcinoma). Image features are extracted using the pre-trained VGG-16 model. The cancer type is identified through three machine learning classification algorithms: Stochastic Gradient Descent (SGD), Random Forest (RF), and K-Nearest Neighbor (KNN). The model's evaluation employed a 10-fold cross-validation technique, with CNN-SGD exhibiting the highest performance based on evaluation metrics. On a scale of 0 to 100, it scored 99.8 for Area Under Curve (AUC) and 98.88 for Classification Accuracy (CA). CNN-RF, a model with performance closely following CNN-SGD, demonstrates training times 58.3 seconds faster than CNN-SGD. Meanwhile, CNN-KNN ranks last among the models evaluated in this study based on its F1, recall, AUC, and CA scores.

***Keywords**: convolutional neural networks, histopathological images, hybrid model, LC25000, machine learning*

KLASIFIKASI JARINGAN KANKER PARU-PARU DAN USUS BESAR MENGGUNAKAN *HYBRID CONVOLUTIONAL NEURAL NETWORKS*

Chilyatun Nisa'[1)], Nanik Suciati[2)], dan Anny Yuniarti[3)]
[1, 2,3)] Department of Informatics Engineering, Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia 60111
e-mail: nchilyatun@gmail.com[1)], nanik@if.its.ac.id[2)], anny@if.ac.id[3)]

ABSTRAK

Kanker usus besar dan paru-paru adalah dua jenis kanker yang sangat mematikan yang seringkali terjadi bersamaan dan menimbulkan tantangan baru dalam diagnosis yang akurat. Meskipun penelitian sering kali berkonsentrasi pada pendeteksian satu kanker pada organ tertentu, penelitian ini mengusulkan pendekatan pembelajaran mesin yang inovatif untuk mengidentifikasi kanker usus besar dan paru-paru. Tujuannya adalah untuk membuat model klasifikasi pembelajaran mesin hibrid untuk meningkatkan presisi diagnostik. Dataset LC25000 terdiri dari 25.000 sampel gambar histopatologi berwarna jaringan sel paru-paru dan usus besar, yang menunjukkan ada tidaknya kanker (adenokarsinoma). Fitur gambar diekstraksi menggunakan model VGG-16 terlatih. Jenis kanker diidentifikasi melalui tiga algoritma klasifikasi pembelajaran mesin: Stochastic Gradient Descent (SGD), Random Forest (RF), dan K-Nearest Neighbor (KNN). Evaluasi model menggunakan teknik validasi silang 10 kali lipat, dengan CNN-SGD menunjukkan kinerja tertinggi berdasarkan metrik evaluasi. Pada skala 0 hingga 100, skornya 99,8 untuk Area Under Curve (AUC) dan 98,88 untuk Classification Accuracy (CA). CNN-RF, model dengan performa yang mirip dengan CNN-SGD, menunjukkan waktu pelatihan 58,3 detik lebih cepat dibandingkan CNN-SGD. Sementara itu, CNN-KNN menempati peringkat terakhir di antara model-model yang dievaluasi dalam penelitian ini berdasarkan skor F1, recall, AUC, dan CA.

***Kata Kunci**: jaringan saraf konvolusional, citra histopatologi, model hibrida, LC25000, pembelajaran mesin*

I. INTRODUCTION

Cancer ranks as the second most prevalent reason for mortality worldwide. This disease is characterized by uncontrolled growth of abnormal cells and the ability to invade and spread to other cells and tissues of the body. In some cases, the cancers that can develop concurrently are in the lung and colon areas, classified as prevalent forms of malignancies [1]. In Indonesia, there were 396,941 new cases and 234,511 deaths from cancer, with lung cancer ranking fourth at 8.8% and colon cancer ranking sixth at 8.6% of total cases [2]. Certain individuals may experience both conditions, with lung cancer manifesting as a secondary primary cancer among

those diagnosed with colon cancer. Despite the stronger association of lung cancer linked to upper aerodigestive cancer, the connection to colon cancer must not be disregarded. In the field of healthcare, cancer identification in its early stages relies on manual examination and a diagnostic system [3]. Enhancements can be achieved through the implementation of image processing methods. Nonetheless, numerous scientists have enhanced this detection approach through the utilization of image analysis methodologies like Chest X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Sputum Cytology [3].

In this research, we introduce a computer-assisted diagnostic system that combines a hybrid Convolutional Neural Network (CNN) with traditional machine learning algorithms, including Stochastic Gradient Descent (SGD), Random Forest (RF), and K-Nearest Neighbor (KNN). The CNN model is used to extract features from image data, while traditional machine learning algorithms are employed to classify cancer tissues within the inputted data. This hybrid approach can address the shortcomings of each algorithm and enhance their performance [4]. To handle image data, especially for the task of feature extraction, a neural network algorithm that demonstrates significant performance is the CNN [5]. KNN serves as an example of an ideal algorithm for non-linear data and is capable of handling multi-class cases [6]. On the other hand, SGD is efficient in terms of speed and is well-suited for large datasets [7]. RF is an ensemble algorithm that is well-suited for addressing overfitting due to its robustness [8]. For training and evaluating these hybrid models, we employ the LC25000 image dataset, which comprises twenty-five thousand histopathological image color samples [9]. These samples correspond to either lung or colon tissues, indicating the presence or absence of cancer (specifically adenocarcinoma).

In contrast to the research undertaken by [10], our supplementary contribution involves the classification of colon data from the LC25000 dataset and the utilization of other hybrid CNN methods. Furthermore, we address gaps in their study, such as the AUC score, which is better than accuracy for measuring model performance [11], as well as the lift plot and calibration curve for comprehensive model evaluation. These aspects, which did not previously exist, form a significant part of our contributions.

II. Related Works

Some scientists have conducted research related to colon and lung cancer classification with CNN or hybrid approaches, leveraging LC25000 image datasets. The following are examples of well-executed research in this domain. In research conducted by Masud, Mehedi et al. [13], they are proposed a model for automatic diagnosis of histopathological patterns in colorectal and pulmonary tissues using colored image data based on deep learning. The process involved sharpening through unsharp masking, constructing a feature set of images using Fourier and wavelet methods, and subsequently feeding the extracted features into a CNN model. The obtained results demonstrated that proposed model is able to identify effectively cancerous tissues with 96.33% accuracy score and precision, recall, & F1 scores all above is 96%. A study that utilized a pre-trained CNN-based model to identify lung and colon cancer using histopathology images was carried out by [14]. This research used different trained CNN models, including VGG16, ResNet50, and NASNetMobile trained on the LC25000 dataset, which respectively obtained accuracy on test data of 98%, 96%, and 97%.There is another research related to medical images that have been carried out by [15] involved abnormal detection and classification of blood cells using Faster R-CNN and Graph CNN, the experiment compared the classification performance of pretrained CNN models (Resnet-101 and VGG-16) as well as GCN models (Resnet-101 + GCN and VGG-16 + GCN). The best results were obtained with the GCN model using the VGG16 architecture (VGG16 + GCN), achieving an accuracy of 95%.

Ahmad, Vasta, et al. [12] conducted a groundbreaking research endeavor focused on the automated detection of lung cancer in histopathological tissue samples. Employing a single handcrafted convolutional neural network (CNN) with five convolutional blocks, each containing two to three layers, and incorporating max pooling layers in every block, they achieved remarkable results. Their classification model surpassed the accuracy metrics of a previous study [13], boasting an impressive 99% accuracy and a notable 96.8% in averages for precision, recall, and F1-score. However, recent advancements in the field by Jehangir, Basra et al. [10] introduced an innovative approach, utilizing ensemble/hybrid machine learning models for lung cancer detection. Leveraging fifteen thousand histopathological digital tissue images from the LC25000 dataset, they applied a CNN to extract image features, subsequently subjecting them to various algorithms including Support Vector Machine, Random Forest, and XG Boost. This comprehensive approach demonstrated superior performance, surpassing the aforementioned study [12], with outstanding achievements such as the highest accuracy at 99.07%, an average recall of 100%, an average F1-score of 98.7%, and an average precision of 98.71%. The continual evolution of methodologies in lung cancer detection underscores the dynamic nature of medical research and its potential to significantly impact diagnostic accuracy and patient outcomes.

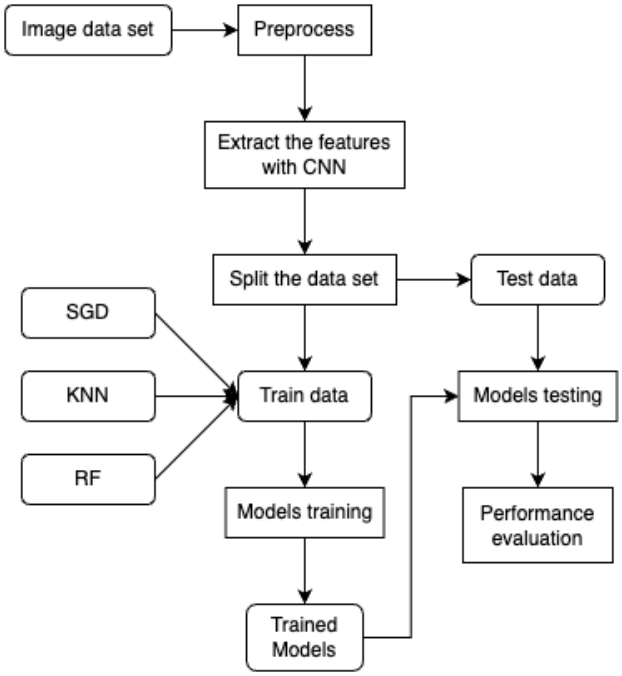


Fig 1. Illustration of the experimental process.

III. METHODOLOGY

This section describes the materials used and the methods applied in this research. Includes tools, datasets, algorithms, and model evaluations. The experiment process of this research is illustrated in Figure 1 below. Before the features extracted with CNN the data set must be preprocessed first, if its complete then we directly extract the features. Process continuous to splitting the data set into train data and test data. We train the classifiers on a train data, when the training is finished, we test it into a test data using cross validation. Then, we perform evaluation for model's performance by using the confusion matrix.

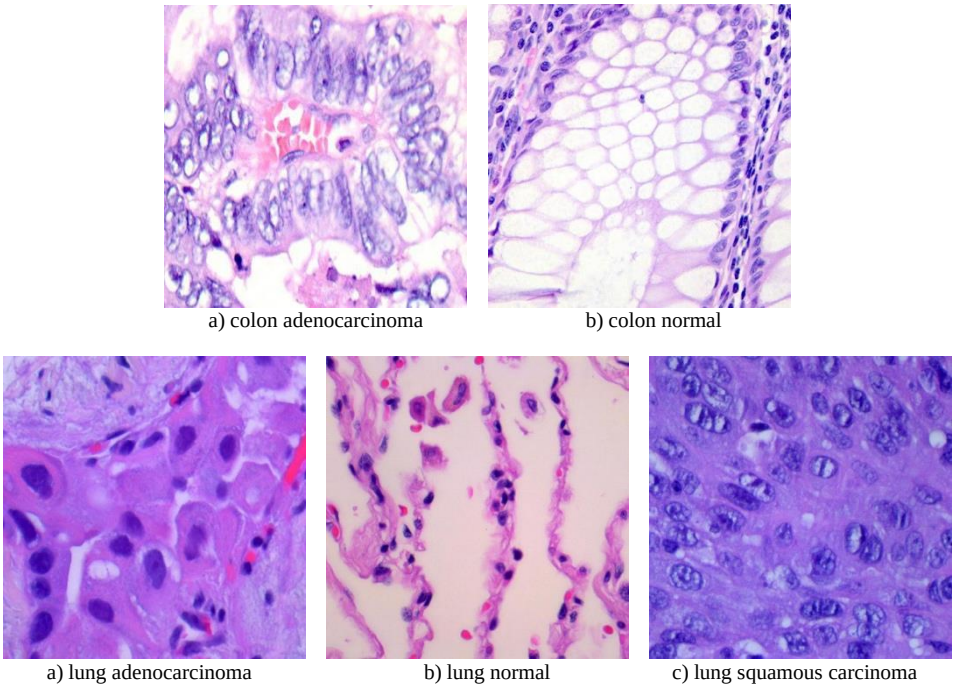
We have used Intel Core i7-9700F 3.00GHz (8 CPUs) as a with 32 GB of RAM. Every classifier was trained locally on Windows 10 Pro OS on Visual Studio Code using IPyNB kernel as IDE, Torch Vision or PyTorch for CNN framework, and scikit-learn for the ML algorithms. Additionally, the system employed an NVIDIA GeForce RTX 2080 Ti GPU for enhanced computational performance during the training process.

A. Data Set

This study uses the LC25000 dataset, a public dataset containing 25000 colored images of lung and colon tissue where cancer (adenocarcinoma) is not indicated or indicated. This dataset was created by A. Andrew Borkowski et al [9]. They collected images from the pathology department at James A. Haley Veterans' Hospital, Florida. It contains five directories, each category represented by one directory. Forty percent of LC25000 contains colon tissue data. There are 10000 colon tissue images from 25000 total images, shown in Figure 2. It has only two categories of it: benign (n) and adenocarcinoma (aca). Sixty percent rest of the LC25000 dataset (15000 images) contains of lung tissue data. It only has three categories of it: squamous carcinoma (scc), adenocarcinoma (aca), and benign (n). Every image in the LC25000 dataset possesses an identical resolution of 768x768 pixels. Additionally, the dataset's creators have augmented the images through random rotation, horizontal flipping, and vertical flipping, aimed at increasing the dataset's volume of data.

B. Preprocess

Before subjecting the image data to classification by machine learning models, a preprocessing step is indispensable for all images [16]. This initial phase commences with resizing the dimensions of each image from 768x768 pixels to 200x200 pixels. Following this, a subtle random brightness adjustment of 0.05% is implemented to refine identification accuracy, a refinement from the previously used 0.5% [17]. Consequently, the images undergo transformation into tensors, whereby every pixel in each image is translated into a numerical value ranging from 0 to 255 for each of the red, green, and blue channels. These values are subsequently normalized using a standard deviation and mean of 0.5. This meticulous preprocessing ensures that the image data is appropriately formatted and prepared for effective utilization by the machine learning models, thereby facilitating optimal performance and accuracy in subsequent classification tasks.



a) colon adenocarcinoma b) colon normal

a) lung adenocarcinoma b) lung normal c) lung squamous carcinoma

Fig 2. The LC25000 dataset samples

C. CNN for Feature Extraction

CNN is a machine learning (ML) algorithm that employs a feedforward neural network to extract and learn features of image using multiple layers [18]. The extracted features of images are from convolutional and pooling layers of CNN itself. Where a collection of image filters to perform the convolution operation on the given data is convolutional layers. The pooling layer serves as a layer for reducing dimensionality and determines the threshold [19]. There are many variations architecture of CNN, one of the examples is VGG16. Simonyan and Zisserman initially proposed VGG16 to introduce the improvements of AlexNet, a CNN model developed by Google [20]. It has been trained on 14 million images from the ImageNet database with 1000 labeled classes [21]. Like other CNN models which has the convolutional and pooling layer, VGG16 also can be used for extracting the image without pass it into fully connected layer. In this research we utilize pretrained VGG16 model to extract the image features and then pass it to traditional machine learning algorithms without the backpropagation step.

D. Machine Learning

Classification is one of the machine learning tasks. With this mechanism we can detect the label of some input, like a medical data, based on the previous input that have been understood what the label is by the algorithm [22]. In our case we use the colon and lung tissues data to be "learned" by the classification algorithm in order to make it capable to perform lung and colon cancer classification. Input for this algorithm is extracted feature from VGG16 CNN model. We choose three ML algorithm to perform classification task, they are K-Nearest Neighbor, Stochastic Gradient Descent, and Random Forest. KNN is an instance-based learning technique that used to classify a sample of data by measuring its proximity to neighboring data points that belong to a set of pre-labeled classes. The distance measurement usually using Euclidean distance, like the equation (1) below, but other methods are also possible. Where x_{in} is train data and x_{jn} is input data that want to be predict the class is. The predicted class is determined based on the popularity vote regarding its distance from other neighboring data points belonging to the neighboring class [6]. Due to its simplicity, easily handle multiclass data, and ideal for non-linear data we choose it to train the classification algorithm.

$$\Delta(x_i x_j) = \sqrt{\sum_{i=1}^{n} \left(|x_{in} - x_{jn}|\right)^2} \quad (1)$$

SGD is an ML technique that works to optimize the baseline classification model in order to find the best parameters throughout the gradient changes with mathematical calculations as in equation (2), where γ shows the model gain and ω shows the model weight that must be close to optimal. The estimated value z is chosen randomly and given a weight ω_t for each iteration t. This process helps optimize the values produced in each iteration as long as the values chosen are based on the ground truth distribution. The convergence of the SGD model assumes a gradual decrease in the value of the model gain γ, which cannot be too fast or too slow. The optimal value for convergence

can be achieved when γ_t approaches the value t-1, causing a decrease in the error rate at the same speed [23]. So, it is just an optimization technique that usually fits with linear ML models, not a specific family of ML models. SGD has proven effective in addressing extensive and sparse machine learning challenges frequently encountered in tasks like image and natural language processing [24]. We choose it due to its good performance in handling large data and in terms of computation speed. In scikit-learn implementation, it's used for several baseline ML models like SVM and Logistic Regression (LR). In this research, we use LR for our SGD baseline, so the model is equivalent to LR, which is fitted via SGD [24].

$$\omega_t + 1 = \omega_t - \gamma \frac{1}{n} \sum_{i=1}^{n} \nabla \omega Q(z_i \omega_t) \quad (2)$$

RF can be considered as extension version of the decision tree algorithm. It builds multiple decision trees and combines their outputs to improve accuracy and robustness, addressing some of the limitations of a single decision tree [25]. In RF, the final feature has an importance calculated as the average of all trees. The number of feature importance values in each tree is calculated and then divided by the total number of trees as in equation (3) below. Where RFi_i is the importance value of feature i calculated from all trees in the RF model. Then $normfi_{ij}$ denotes the normalized feature importance for i in the j-th tree, and T as the total number of trees. Scientists classify it as a group of bagging ensemble method. It has proven successful on wide array of classification and regression case [25]. There are many ways to combines the output from each tree that generated by RF algorithm. The easy way to do it is majority voting from tree outputs, but in the scikit-learn implementation is not work that way. It combines classifiers by averaging their probabilistic prediction (output), instead of letting each classifier vote for a single class at the end [24] due to it gives a more balanced and precise result by considering the confidence levels of each classifier's prediction.

$$RFfi_i = \frac{\sum_{j \in alltrees} normfi_{ij}}{T} \quad (3)$$

E. Model Evaluation

In this research we use two base methods to evaluate the performance of our models, which are k-fold Cross Validation and Confusion Matrix. k-Fold Cross Validation is employed to assess the stability of evaluation results for a model across multiple folds of data, each derived from a single dataset [26]. This helps measure the model's performance more reliably and reduces the risk of overfitting or underfitting on specific data. For the example if model A have accuracy score relatively same among all folds, then we can say that model A is stable if it's not then we should retrain model A with different parameter in order to make the accuracy score relatively same among all folds. After completing the process with k-folds Cross Validation, then we move the stable model to be evaluated using confusion matrix. A confusion matrix is a tabular tool utilized to evaluate the effectiveness of a classifier by comparing the model's predictions with the actual values of the tested data [27]. It has four cells, which are:

- True-Positive (TP): the number of samples that were predicted to be true and were true;
- False-Positive (FP): the number of samples that were correctly predicted, but turned out to be wrong;
- False-Negative (FN): the number of samples that were predicted incorrectly, but turned out to be correct
- True-Negative (TN): the number of samples that are predicted wrongly and are indeed wrong.

From four points above, we can evaluate our machine learning models with several metrics like recall, accuracy, F1, and precision score [27]. Where accuracy is the ratio between the number of correct predictions and the total number of predictions. Precision is to measure of correctness that is achieved in true prediction. Recall is a measure of actual observations which are predicted correctly. Then F1 is the harmonic mean of precision and recall. To strengthen our model evaluation, we add one more metric that proven more reliable than accuracy to evaluate the model in healthy diagnosed system which is ROC-AUC metric [24]. The AUC-ROC curve assesses classification performance across different threshold settings. The ROC is a probability curve, and the AUC signifies the level of distinction. It indicates the model's ability to differentiate between classes [28].

IV. RESULTS AND DISCUSSION

The results consist of evaluation metric tables gained from the models' confusion matrices for both colon and lung images, along with the ROC curve. Based on Figure 3, we can see that CNN-KNN has the lowest AUC score which is 0.9134. The highest AUC score achieved by CNN-SGD with 0.9975 and then AUC score for CNN-RF is not significant if compared with CNN-SGD the different is relatively small, only 0.0036 point.

The insights extracted from Figure 4, illustrated in Table I through the confusion matrix, reveal key performance metrics. The matrix assigns '0' to colon adenocarcinoma (aca), '1' to colon normal (n), '2' to lung adenocarcinoma (aca), '3' to lung normal (n), and '4' to lung squamous cell carcinoma (scc) categories. Examining Table I, the CNN-KNN model delivered remarkable results with an AUC score of 0.94 and an accuracy score of 0.946. Furthermore, the model exhibited impressive average precision, recall, and F1-score values of 0.83, 0.77, and 0.774, respectively.

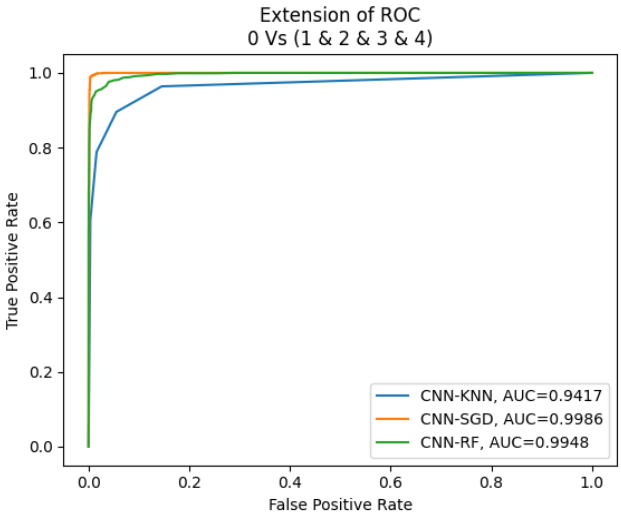


Fig. 3. Receiver operating characteristic curve for each model

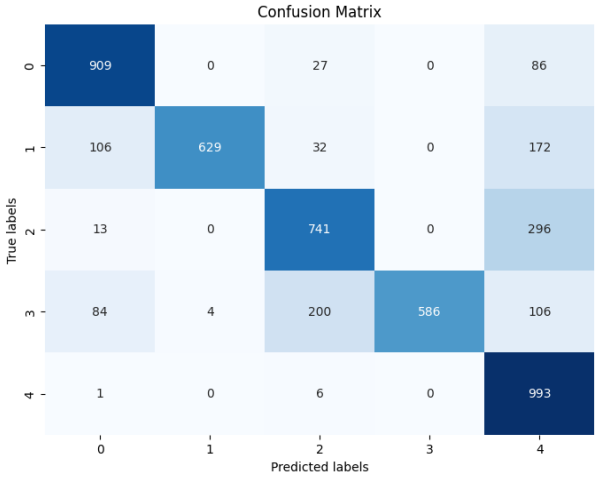


Fig. 4. CNN-KNN's confusion matrix

Noteworthy is the swift training time of the CNN-KNN model, clocking in at under 1 second, precisely at 0.863 seconds, owing to its streamlined training mechanism.

Illustrated in Figure 5 is the below confusion matrix, showcasing the exceptional accuracy of CNN-SGD's predicted values, accurately aligning with over 960 data points within each category. Beyond the diagonal line, CNN-SGD demonstrated a mere 87 misclassifications while astutely categorizing 4904 data points. The results in Table II further emphasize CNN-SGD's prowess, boasting an impressive AUC score of 0.998 and an elevated accuracy score of 0.98. Additionally, the model exhibited exceptionally high average precision, recall, and F1-score values, reaching 0.982, 0.984, and 0.984, respectively, surpassing the performance of the CNN-KNN model. It is noteworthy, however, that the training time for CNN-SGD is considerably longer, spanning 219 seconds to complete the training process, a significant contrast to the training time of CNN-KNN.

Delving into the nuanced details unveiled in Figure 6 is confusion matrix, the robust performance of CNN-RF becomes unmistakably apparent, presenting an exceptional feat with a remarkably low misclassification rate of merely 290 instances beyond the diagonal line. Elevating its prowess, CNN-RF not only showcases efficiency but also proficiency by achieving a remarkable accuracy in predicting over 920 data points within each category. This

TABLE I
EVALUATION METRICS SCORE OF CNN-KNN

Class	Precision	Recall	F1-Score	Accuracy	AUC
colon_aca	0.82	0.89	0.85		
colon_normal	0.99	0.67	0.8		
lung_aca	0.74	0.71	0.72	0.76	0.94
lung_normal	1	0.6	0.75		
lung_scc	0.6	0.99	0.75		

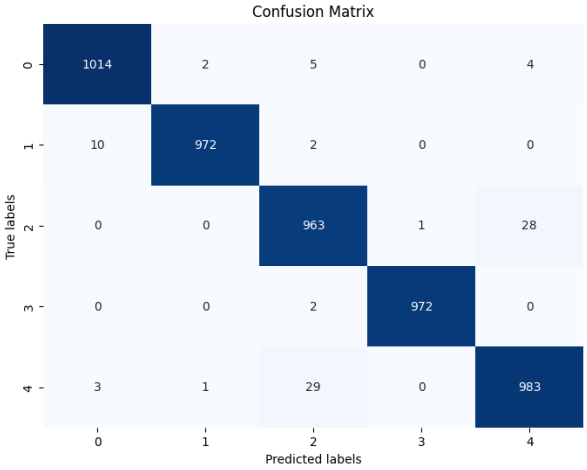


Fig. 5. Fig. CNN-SGD's confusion matrix

TABLE II
EVALUATION METRICS SCORE OF CNN-SGD

Class	Precision	Recall	F1-Score	Accuracy	AUC
colon_aca	0.99	0.99	0.99	0.98	0.99
colon_normal	1	0.99	0.99		
lung_aca	0.96	0.97	0.97		
lung_normal	1	1	1		
lung_scc	0.97	0.97	0.97		

not only underscores its aptitude for precise categorization but also highlights its efficacy in managing a diverse array of instances. What further accentuates the standout performance of CNN-RF is the meticulous alignment of difference values when compared to CNN-SGD. This alignment accentuates the consistent and reliable predictive capabilities of CNN-RF across various categories, with a noteworthy exception in category '2,' representing lung adenocarcinoma. Even in this exception, CNN-RF maintains a commendable level of predictive accuracy, solidifying its overall reliability and robustness in the realm of cancer tissue classification.

Presented in Table III, the CNN-RF model demonstrates a commendable AUC score of 0.994, coupled with a high accuracy score of 0.946. Additionally, the model exhibits elevated average precision, recall, and F1-score values, reaching 0.94, 0.94, and 0.942, respectively. These results not only surpass those of the CNN-KNN model but also underscore the superior performance of CNN-RF across multiple evaluation metrics. Despite achieving outstanding evaluation scores, CNN-RF distinguishes itself further by boasting a faster training time than CNN-SGD, completing the full training process in a mere 160.7 seconds. This represents a significant contrast with CNN-SGD, which takes almost 1 minute longer, precisely 58.3 seconds. The efficiency demonstrated by CNN-RF in both evaluation scores and training time adds to its appeal as a robust and time-effective machine learning model.

All the hybrid CNN models showcased remarkable performance with AUC scores surpassing 0.94, as detailed in Table I up to Table III. Among these models, CNN-SGD stands out as the top performer across diverse metrics,

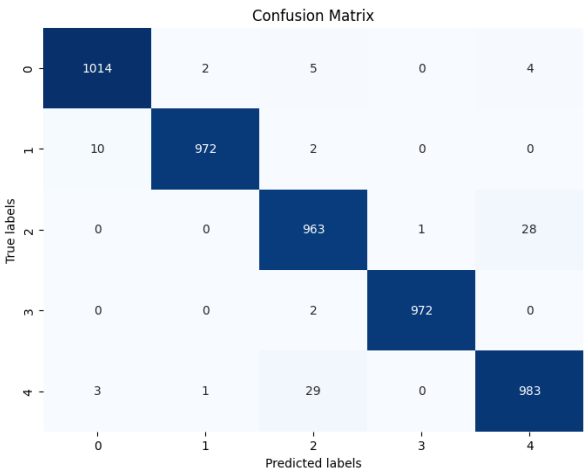


Fig. 6. CNN-RF's confusion matrix

TABLE III
EVALUATION METRICS SCORE OF CNN-SGD

Class	Precision	Recall	F1-Score	Accuracy	AUC
colon_aca	0.96	0.94	0.95	0.94	0.99
colon_normal	0.97	0.98	0.97		
lung_aca	0.89	0.89	0.89		
lung_normal	0.96	1	0.98		
lung_scc	0.94	0.91	0.92		

including accuracy, AUC score, average F1, average recall, and average precision, achieving exemplary scores of 0.98, 0.998, 0.984, 0.984, and 0.982, respectively. Despite CNN-SGD's superior results in evaluation metrics, the fastest machine learning model proposed in this study is CNN-KNN, boasting an impressive training time of less than 1 second. While CNN-RF's evaluation scores compete admirably with CNN-SGD, it gains an advantage in terms of training time, completing the process in a swift 58.3 seconds, almost 1 minute less than CNN-SGD. This nuanced analysis underscores the nuanced strengths of each model, with CNN-SGD excelling in performance metrics and CNN-KNN demonstrating unparalleled speed, while CNN-RF strikes a balance between competitive evaluation scores and efficient training time.

The experiments conducted in this research have yielded compelling evidence of enhanced model performance, notably achieving an impressive accuracy of 99.86%. This significant advancement stands in stark contrast to a prior study [14], wherein the classification model relied solely on a pre-trained VGG16 architecture and achieved an accuracy of 98%. The substantial improvement in accuracy observed in our research underscores the efficacy of the methodologies employed, suggesting that the novel approach or techniques implemented have contributed to a more robust and precise model.

V. CONCLUSIONS

The fusion of CNN as a feature extractor with traditional ML algorithms has yielded commendable results in the classification of colon and lung cancer tissues. Notably, the CNN-SGD model emerges as the frontrunner, surpassing its counterparts based on the comprehensive evaluation metrics provided. CNN-RF exhibits a distinct advantage in terms of training time when compared with CNN-SGD, further broadening the spectrum of efficient model alternatives. Conversely, while CNN-KNN boasts the shortest training time, its results fall considerably short of the robust performance achieved by CNN-RF. This analysis unveils promising avenues for improvement, particularly through the exploration of diverse cancer tissues from various organs. Enhancements can be sought through the incorporation of alternative hybrid learning methods, architectural refinements, and meticulous tuning of hyperparameters. Additionally, the establishment of a vector database for constructing the image data pipeline stands out as a pivotal initiative, promising to elevate the modeling experience for scientists, particularly when translating this expertise into industrial applications. The quest for continuous improvement remains a dynamic journey, fostering innovation and refinement in the landscape of cancer tissue classification.

REFERENCES

[1] Kurishima, K., Miyazaki, K., Watanabe, H., Shiozawa, T., Ishikawa, H., Satoh, H., & Hizawa, N. (2018). Lung cancer patients with synchronous colon cancer. 8(1), 137–140. https://doi.org/10.3892/mco.2017.1471

[2] Cancer Today (2020). Global Cancer Observatory (GLOBOCAN). Diakses pada Juni 12 2023, from http://gco.iarc.fr/today/home

[3] Sasikala, S., Bharathi, M., & Sowmiya, B. (2019). Lung Cancer Detection and Classification Using Deep CNN, International Journal of Innovative Technology and Exploring Engineering (IJITEE).

[4] Psichogios, D., dan Ungar, L., 1992. "A hybrid neural network-first principles approach to process modeling". AIChE Journal 38, 10:1499- 1511.

[5] Simie, E., & Kaur, M. (2019). Lung cancer detection using Convolutional Neural Network (CNN). International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIT).

[6] A. Joby, "K Nearest Neighbor (KNN): The Most Used ML Algorithm." https://learn.g2.com/k-nearest-neighbor (accessed Aug. 26, 2023).

[7] R. Bhatia, "How Stochastic Gradient Descent Is Solving Optimisation Problems In Deep Learning," Analytics India Magazine, Sep. 21, 2018. https://analyticsindiamag.com/how-stochastic-gradient-descent-is-solving-optimisation-problems-in-deep-learning/ (accessed Aug. 26, 2023).

[8] N. Donges, "What Is Random Forest? A Complete Guide | Built In," *builtin*, Mar. 14, 2023. https://builtin.com/data-science/random-forest-algorithm (accessed Aug. 26, 2023).

[9] Borkowski, A.A. et al. (2019). Lung and Colon Cancer Histopathological Image Dataset (LC25000). ArXiv,abs/1912.12142.

[10] B. Jehangir, S. R. Nayak and S. Shandilya, "Lung Cancer Detection using Ensemble of Machine Learning Models," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2022, pp. 411-415, doi: 10.1109/Confluence52989.2022.9734212.

[11] C. X. Ling, J. Huang, and H. Zhang, 'AUC: A Statistically Consistent and More Discriminating Measure than Accuracy', in Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico, 2003, pp. 519–524.

[12] V. Anand, K. S. Gill and S. Gupta, "Multi-class Classification of Colon and Lung Cancer using Deep Convolution Neural Network," 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2023, pp. 447-451, doi: 10.1109/ICSCSS57650.2023.10169254.

[13] M. Masud, N. Sikder, A.-A. Nahid, A. K. Bairagi, and M. A. AlZain, “A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning-Based Classification Framework,” Sensors, vol. 21, no. 3, p. 748, Jan. 2021, doi: 10.3390/s21030748.

[14] Garg, S., & Garg, S. (2021). Prediction of lung and colon cancer through analysis of histopathological images by utilizing Pre-trained CNN models with visualization of class activation and saliency maps. Proceedings of the 2020 3rd Artificial Intelligence and Cloud Computing Conference, 38–45. https://doi.org/10.1145/3442536.3442543

[15] Bramantya, B. A., Fatichah, C., & Suciati, N. (2022). DETECTION AND CLASSIFICATION OF RED BLOOD CELLS ABNORMALITY USING FASTER R-CNN AND GRAPH CONVOLUTIONAL NETWORKS. JUTI: Jurnal Ilmiah Teknologi Informasi, 20 Number 1, 33–44. https://doi.org/http://dx.doi.org/10.12962/j24068535.v19i3.a1118

[16] Pal, K.K., & Sudeep, K.S. (2016). Preprocessing for image classification by convolutional neural networks. 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 1778-1781.

[17] Crispell, D.E., Biris, O., Crosswhite, N., Byrne, J., & Mundy, J.L. (2017). Dataset Augmentation for Pose and Lighting Invariant Face Recognition. ArXiv, abs/1704.04326.

[18] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” nature, vol. 521, no. 7553, pp. 436–444, 2015.

[19] T. M. Navamani, ‘Chapter 7 - Efficient Deep Learning Approaches for Health Informatics’, in Deep Learning and Parallel Computing Environment for Bioengineering Systems, A. K. Sangaiah, Ed. Academic Press, 2019, pp. 123–137.

[20] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009, pp. 248–255.

[22] J. Brownlee, “4 Types of Classification Tasks in Machine Learning”, Machine Learning Mastery, Apr. 08, 2020. https://machinelearningmastery.com/types-of-classification-in-machine-learning/

[23] R. Roy, “ML | Stochastic Gradient Descent (SGD),” GeeksforGeeks, Feb. 15, 2019. https://www.geeksforgeeks.org/ml-stochastic-gradient-descent-sgd/ (accessed Aug. 30, 2023).

[24] Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O, Blondel, M, Prettenhofer, P, Weiss, R, Dubourg, V, Vanderplas, J, Passos, A, Cournapeau, D, Brucher, M, Perrot, M, Duchesnay, E. "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research 2011; 12:2825–2830.

[25] J. Brownlee, “How to Develop a Random Forest Ensemble in Python”, Machine Learning Mastery, Apr. 20, 2020. https://machinelearningmastery.com/random-forest-ensemble-in-python/

[26] J. Brownlee, “A Gentle Introduction to k-fold Cross-Validation”, Machine Learning Mastery, Aug. 23, 2020. https://machinelearningmastery.com/k-fold-cross-validation/

[27] A. Suresh, “What is a confusion matrix?,” Analytics Vidhya, Jun. 22, 2021. https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5 (accessed Aug. 30, 2023).

[28] Narkhede, Sarang. “Understanding AUC - ROC Curve.” Medium, 15 June 2021, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.