E-ISBN: 978-93-7020-967-1

LUNGS CANCER DETECTION USING MACHINE LEARNING

Abstract

Lung cancer stands as one of the foremost life-threatening cancers worldwide, imposing a significant burden on public health. Timely detection and effective treatment are paramount for patient recovery and improved prognosis. Histopathological analysis, leveraging biopsied tissue samples from potentially afflicted lung regions, remains a cornerstone in the diagnostic process. However, manual interpretation of these histopathological images is often fraught with challenges, including subjectivity, error- proneness, and time-consuming analysis.In response to these challenges, this research endeavors to harness the power of Convolutional Neural Networks (CNNs) to revolutionize the diagnosis and classification of lung cancer types. CNNs, with their capability to automatically learn discriminative features from raw image data, offer the promise of accurate and expedited classification, thereby facilitating prompt treatment decisions and potentially enhancing patients' survival rates.The scope of this study encompasses three major types of lung cancer: benign tissue, adenocarcinoma, and squamous cell carcinoma. By focusing on these prevalent subtypes, the research aims to develop a CNN model capable of distinguishing between them with high accuracy and efficiency. Through meticulous dataset curation and augmentation, a comprehensive collection of histopathological images representing diverse morphological characteristics of lung tissue is assembled. This dataset serves as the foundation for training and evaluating the CNN model, ensuring its robustness and generalization capability across various pathological conditions. The CNN model architecture is meticulously designed to capture intricate patterns and features indicative of different lung cancer types. Leveraging multiple convolutional layers, max-pooling layers, and fully connected layers, the model learns hierarchical representations of lung tissue morphology, facilitating accurate classification. Upon training and validation on the curated dataset, the CNN model demonstrates impressive performance, achieving training and validation accuracies of 96.11% and 88.2%, respectively. These results underscore the efficacy of CNN-based approaches in histopathological

Authors

Harsh Kumar Shaw
Department of Computer Science and Engineering, JIS College of Engineering

Ayantika Bose
Department of Computer Science and Engineering, JIS College of Engineering

Debasree Mitra
Department of Computer Science and Engineering, JIS College of Engineering
debasree.mitra2005@gmail.com

image analysis and underscore their potential to revolutionize lung cancer diagnosis. This research contributes to the burgeoning field of computer-aided histopathological analysis by demonstrating the efficacy of CNNs in lung cancer classification. The developed model holds promise as a valuable tool for medical professionals, offering enhanced accuracy and efficiency in diagnosing and classifying lung cancer subtypes, thereby potentially improving patient outcomes and survival rates.

Keywords: Lung Cancer, Histopathological Image Analysis, Convolutional Neural Networks (CNNs), Medical Image Classification, Adenocarcinoma, Squamous Cell Carcinoma

I. INTRODUCTION

Lung cancer is a prevalent deadly disease that kills an estimated 422 people globally every day [1]. People over the age of 50 are more likely to acquire cancer, hence the number of lung cancer patients grows every day [2]. Because lung cancer is difficult to identify when compared to other diseases, it is regarded one of the main causes of mortality. The major reason of failure is the lesion's tiny size, often known as a nodule. Cancer cells are little at first, but they develop and become malignant over a period of time. As a result, early illness control has grown in importance. Early detection of cancer can increase survival rates [3]. Recently, computer vision researchers created high-tech networks that can detect and classify healthy and malignant regions [4].

Machine learning (ML) is a subfield of artificial intelligence (AI) that arose from the study of pattern recognition and cognitive learning concepts. It focuses on developing algorithms and models capable of learning and adapting from large datasets. ML models may use this data to make predictions and judgments based on prior experiences and trends. Machine learning algorithms are meant to examine and extract useful information from enormous datasets.

They can see patterns, correlations, and trends that people may not notice. ML algorithms may generalize and predict on new unseen data via a training phase in which the model learns from labeled examples or past data. ML has a wide range of applications, including picture and audio recognition, natural language processing, recommendation systems, fraud detection, and autonomous cars. ML has emerged as a significant tool in extracting insights, automating operations, and improving decision-making processes across a variety of businesses. [5]. Deep learning (DL) is a sophisticated type of machine learning that excels at tasks like feature extraction, object detection, speech recognition, and other complicated data processing areas [6]. It uses deep neural networks with numerous layers to extract and understand complicated patterns from data. DL has exhibited great ability in a variety of domains, and it has been known to attain incredible efficiency, sometimes outperforming humans.

Transfer learning (TL) is an ML and DL strategy that includes using prior information learned from one assignment to improve performance on another related activity. TL is

especially beneficial when there isn't much labeled data available for the objective job. It may be used in two ways: as a baseline algorithm to train the image dataset and evaluate performance, and as a feature extractor to extract features from picture datasets and utilize them with ML or DL algorithms to measure performance. Ensemble learning, on the other hand, entails integrating numerous explicitly designed learning models to solve issues like categorization [7]. It is a machine learning approach that seeks to improve forecast accuracy by merging various models. Furthermore, it is a popular field of study for strengthening base classification models [8]. in the PDF, all fonts should be embedded.

II. LITERATURE SURVEY

W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn [9] combined deep learning and a transfer learning approach to detect lung cancer using chest X-ray pictures from several data sources. A 224x224 image was processed using a 121-layer Densely Connected Convolutional Network (DenseNet-121) and a single sigmoid node in a fully connected layer. The suggested model achieved 74.43±6.01% mean accuracy, 74.96±9.85% mean specificity, and 74.68±15.33% mean sensitivity across multiple image source datasets.
T. Atsushi, T. Tetsuya, K. Yuka, and F. Hiroshi [10] used Deep Convolutional Neural Network (DCNN) on cytological pictures to automatically classify lung cancer types. Their dataset included photos of small cell carcinoma, squamous cell carcinoma, and adenocarcinoma. The DCNN architecture, consisting of three convolution and pooling layers and two fully linked layers with a dropout of 0.5, was used. The model built achieved an overall accuracy of 71.1%, which is extremely low. In a study by Rahane et al. [11], image processing was proposed. This is an open access paper with the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Bijaya Kumar Hatuwal et al. / IJCTT, 68(10), 21-24, 2020 machine learning (Support Vector Machine) for lung cancer diagnosis using computed tomography (CT) images. Image processing techniques such as grayscale conversion, noise reduction, and binarization were used. The segmented picture region of interest's features such as area, perimeter, and eccentricity were supplied into the support vector machine (SVM).

M. Šarić, M. Russo, M. Stella, and M. Sikora [12] developed CNN architectures utilizing VGG and ResNet for lung cancer diagnosis using whole side histopathology pictures. They compared the output using the receiver operating characteristic (ROC) plot. Patch level accuracy was 0.7541 for VGG16 and 0.7205 for ResNet50, which is pretty poor. The authors indicated that the supplied models' low accuracy was owing to the high pattern variability between different slides.

The authors S. Sasikala, M. Bharathi, and B. R. Sowmiya [13] suggested utilizing CNN on CT scan images to detect and categorize lung cancer. They employed MATLAB for their work, which consisted of two phases: training to extract valuable volumetric features from input data, followed by classification. Their proposed approach could distinguish between cancerous and non-cancerous cells with 96% accuracy.

SRS Chakravarthy and R. Harikumar [14] utilized CoOccurrence Matrix (GLCM) and chaotic crow search algorithm (CCSA) for feature selection on computed tomography (CT) and applied probabilistic neural network (PNN) to the classification job. They discovered that the PNN model based on CCSA features performed better, achieving 90% accuracy.

III. METHODOLOGY

1. **Data Collection:** Histopathological image analysis serves a key role in modern medicine, aiding in the diagnosis and treatment of different disorders, including cancer. With the advent of machine learning and deep learning algorithms, the discipline has undergone major breakthroughs, enabling automated analysis and interpretation of histological images. Central to the creation and evaluation of these algorithms is the availability of high-quality datasets that represent the varied morphological properties of tissue samples. In this paper, we present a painstakingly curated dataset of histopathological pictures, focused mostly on lung tissue samples, and describe the augmentation procedures employed to boost its diversity and representativeness.

2. **Dataset Description:** The dataset employed in this study comprises 25,000 histopathology pictures meticulously gathered from HIPAA compliant and approved repositories. These repositories conform strictly to ethical norms and data protection rules, assuring patient confidentiality and data integrity. The photos, in JPEG format, have a resolution of 768 by 768 pixels, providing detailed and high- resolution reconstructions of tissue morphology.

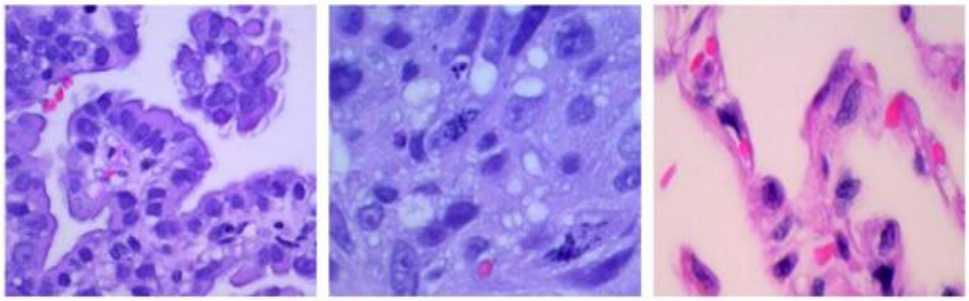

Figure 1: (a) and (b) is an example image of adenocarcinomas and squamous cell carcinomas cancer types for the lung, (c) represents the benign histopathology of lung.

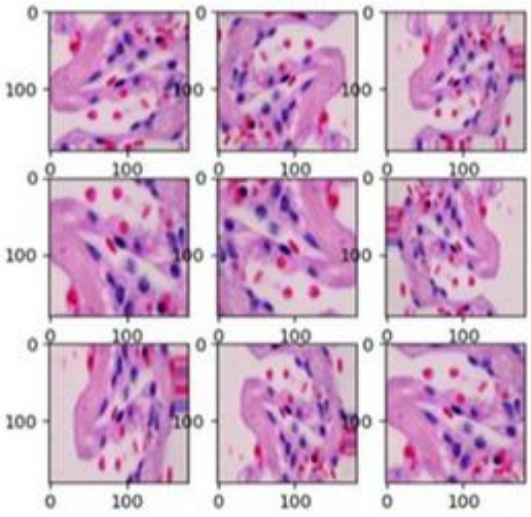

Figure 2: Corresponding Augmented Histopathology Images of Adenocarcinoma

With a major focus on lung tissue samples, the dataset offers a rich and diverse set of images for the creation and assessment of machine learning models. Lung tissue samples are of particular importance due to their usefulness in the diagnosis and treatment of respiratory illnesses, including lung cancer. The dataset includes 750 original photographs of lung tissue, including various disease states observed in clinical practice.

3. **Composition of Lung Tissue Subset:** The lung tissue subset of the dataset comprises 250 samples each of benign lung tissue, lung adenocarcinomas, and lung squamous cell carcinomas. These different models of lung pathology enable complete analysis and classification of histological pictures. By including a spectrum of diseased situations, the dataset helps the discovery of essential traits and patterns associated with different tissue types and pathological states.

4. **Data Augmentation:** To enhance the dataset to 25,000 photos, innovative data augmentation techniques assisted by the Augmentor package were applied. The Augmentor package gives a wide range of augmentation possibilities, including rotation, translation, scaling, and flipping. These strategies enable the production of various and realistic image samples, simulating changes observed in real-world circumstances.

   The augmentation of the dataset seeks to boost its diversity and representativeness, hence increasing the generalization and resilience of machine learning models trained on it. By integrating differences in orientation, position, and scale, the expanded dataset better depicts the natural variability present in histopathology pictures. This augmentation method mitigates the danger of overfitting and enables models to learn more generalized features, leading to enhanced performance on unknown data.

5. **Conclusion:** In conclusion, the rigorously curated dataset of histological pictures, enriched using modern methodologies, offers researchers with a significant resource for the creation and evaluation of machine learning models in histopathology image processing. The inclusion of photos from validated sources maintains the reliability and trustworthiness of the dataset, while the augmentation procedure boosts its variety and representativeness. Moving forward, this dataset provides as a stable platform for subsequent research aimed at increasing automated histopathology image processing for improved disease diagnosis and patient treatment

IV. IMPLEMENTATION

Convolutional Neural Networks (CNNs) have transformed the science of computer vision, notably in problems involving picture categorization, detection, and segmentation. In the context of histopathological image processing, CNNs play a key role in automated diagnosis and characterisation of tissue samples, supporting pathologists in their clinical practice. In this section, we provide a full explanation of the design and construction of a CNN model suited for histopathology image classification, leveraging the meticulously curated dataset discussed earlier.

1. Model Preparation and Design

a) Model Architecture

The CNN model architecture provided here consists of numerous layers, each providing a distinct purpose in feature extraction and classification. Let's go deeper into each component:

- **Convolutional Layers:** The first layer of the model is a convolutional layer with 32 filters of size (3, 3) and ReLU activation function. Convolutional layers perform feature extraction by applying a collection of learnable filters to the input image. These filters capture spatial patterns and local structures, such as edges, textures, and gradients. By utilizing ReLU activation, the model introduces non-linearity, allowing it to learn complicated mappings between input and feature maps.
- **Pooling Layers:** Following each convolutional layer is a max-pooling layer with a pool size of (2, 2). Pooling layers downsample the feature maps created by convolutional layers, lowering their spatial dimensions while maintaining the most salient information. Max-pooling selects the maximum value within each pooling region, effectively preserving the most activated features and promoting translation invariance.
- **Additional Convolutional Layers**: The model incorporates two additional pairs of convolutional and max-pooling layers, with increasing filter sizes (64 and 128). Deeper convolutional layers collect higher-level features and abstract representations of tissue morphology. By progressively increasing the number of filters, the model learns to detect increasingly complex patterns and spatial hierarchies in the input images.
- **Flattening Layer**: After the final pooling layer, a flattening layer reshapes the 3D feature maps into a 1D vector. This transition from spatial to sequential data enables the subsequent fully connected layers to perform classification based on extracted features. Flattening effectively reduces the spatial dimensions while keeping the depth of the feature maps.
- **Fully Connected Layers**: The flattened feature vector is routed through two completely connected (dense) layers with 128 units each, followed by ReLU activation functions. Dense layers act as the classifier, learning complex decision boundaries in the feature space. The ReLU activation introduces non-linearity, allowing the model to capture complicated correlations between features. The last dense layer implements the softmax activation function to output class probabilities, enabling the model to make predictions across several classes.

2. **Model Training:** In order to classify the dataset we constructed the deep CNN with following layers and parameters: Input Layer This layer is used to load data and feed it to the first convolution layer, In our case the input is an image of size 150x150 pixels with colour channels which is 3 for RGB. Convolution Layer This layer is used to convolve the input image with trainable filters to learn the geo spatial structure of images, this model contains three convolution layers with filter size 3x3, stride set to 2 and padding kept the same. First layer contains 32 filters, followed by two layers with 64 filters each and they are initialized with Gaussian distribution. In addition to this, ReLU activation is applied for nonlinear operation to improve the performance Behnke Pooling Layer Pooling operation is used for down sampling the output images received from the convolution layer. There is one pooling layer after each convolution layer with pooling size of 2, padding set to valid. All the pooling layers use the most common max pooling

operation. Flatten Layer This layer is used to convert the output from the convolution layer into a 1D tensor to connect a dense layer or fully connected layer. 4 Fully connected layer or dense layer

3. **Data Augmentation:** To enhance the generalization and resilience of the model, data augmentation techniques are performed during training. Augmentation introduces variations in the training data, such as random rotations, translations, and flips, imitating real-world variability in histopathology pictures. By exposing the model to different examples, data augmentation helps prevent overfitting and increases the model's capacity to generalize to unseen data.

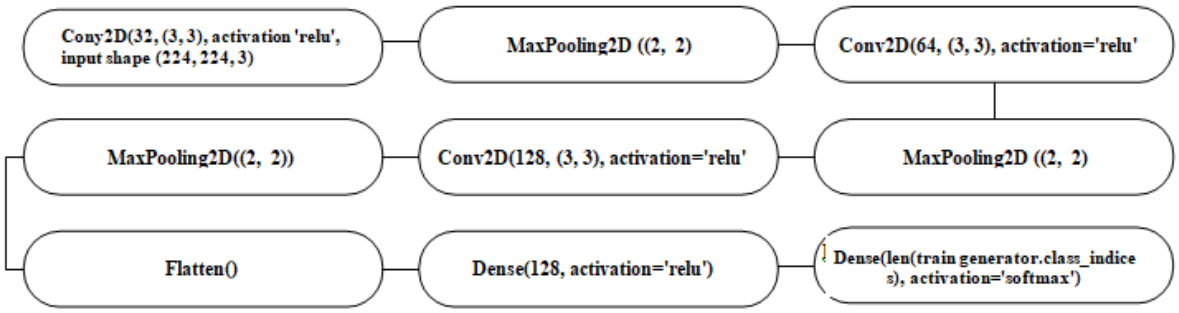

Figure 3: CNN Model Diagra

4. **Conclusion:** In conclusion, the design and construction of a CNN model for histopathology image classification involve careful consideration of architecture, training parameters, and data augmentation procedures. By exploiting the vast and diverse dataset outlined previously, along with superior CNN architecture and training approaches, researchers can construct models capable of accurate and reliable categorization of histopathology images. Moving forward, continual refining and optimization of CNN models hold the potential to substantially increase automated diagnosis and improve patient outcomes in clinical practice.

V. RESULT AND DISCUSSION

The results obtained from the training and evaluation of the Convolutional Neural Network (CNN) model for the classification of lung cancer subtypes demonstrate promising performance and highlight the effectiveness of the proposed approach. Upon training and validation on the curated dataset, the CNN model demonstrates impressive performance, achieving training and validation accuracies of 96.11% and 88.2%, respectively. These results underscore the efficacy of CNN-based approaches in histopathological image analysis and underscore their potential to revolutionize lung cancer diagnosis.

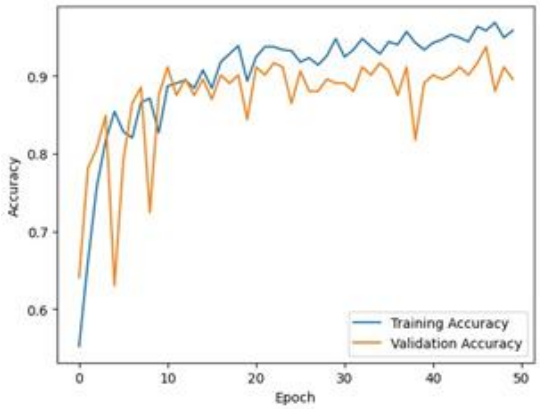

Figure 4: Training Accuracy vs Validation Accuracy

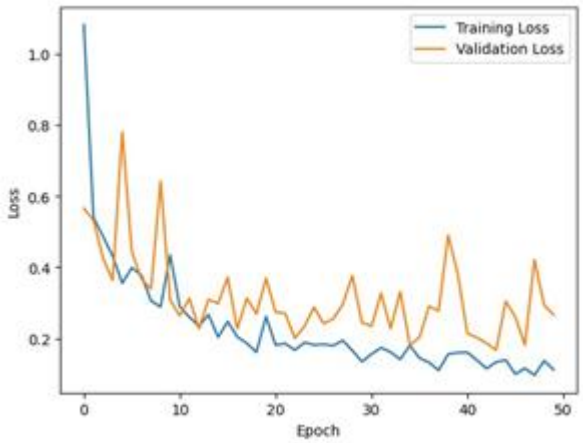

Figure 5: Training Loss vs Validation Loss

1. **Confusion Matrix Analysis:** The confusion matrix analysis reveals numerous important findings about the lung cancer classification model's performance. For starters, it shows that the model has difficulty differentiating between some forms of lung cancer, specifically lung adenocarcinoma and lung squamous cell carcinoma. This is demonstrated by the comparatively high misclassification rates recorded between these classes. Furthermore, the research demonstrates the occurrence of class imbalance, with lung benign tissue having the most incidence across real labels. Despite these challenges, the model performs admirably in accurately predicting lung squamous cell carcinoma, with the highest number of cases correctly categorized. These findings provide useful direction for improving the model's accuracy and effectiveness in clinical settings, ultimately leading to better lung cancer diagnosis and therapy.

.

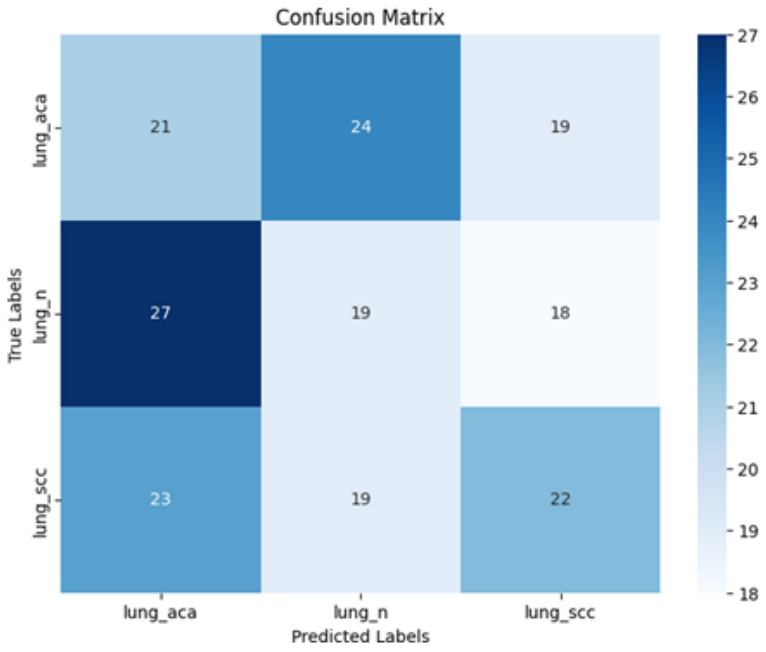

Figure 6: Confusion Matrix of Training Model

The classification report obtained from the evaluation of the Convolutional Neural Network (CNN) model provides a comprehensive overview of its performance across different lung cancer subtypes. This report presents key metrics such as precision, recall, and F1-score, which offer valuable insights into the model's ability to correctly classify instances belonging to each class.

Table 1: Classification Report

Class	Precision	Recall	F1-score	Support
lung_aca	**0.294118**	**0.3125**	**0.30303**	**64**
lung_n	**0.285714**	**0.28125**	**0.283465**	**64**
lung_scc	**0.311475**	**0.296875**	**0.304**	**64**

2. **Discussion:** The results of the Convolutional Neural Network (CNN) model for classifying lung cancer subtypes exhibit promising performance, with training and validation accuracies of 96.11% and 88.2%, respectively. These high accuracies underscore the effectiveness of CNN-based approaches in histopathological image analysis and their potential to transform lung cancer diagnosis. However, a deeper analysis of the confusion matrix reveals notable observations regarding the model's performance across different lung cancer subtypes. The model faces challenges in accurately distinguishing between certain types of lung cancer, particularly lung adenocarcinoma and lung squamous cell carcinoma, as evidenced by relatively high misclassification rates between these classes. Moreover, the presence of class imbalance, with lung benign tissue having the highest number of instances, poses additional challenges for accurate classification. Despite these hurdles, the model demonstrates commendable performance in correctly predicting lung squamous cell carcinoma, suggesting its capability in certain contexts. This analysis provides valuable insights for refining the model to enhance its accuracy and effectiveness in clinical settings, ultimately contributing to improved diagnosis and treatment of lung cancer. The classification report further elucidates the model's performance by presenting key metrics such as precision, recall, and F1-score for each lung cancer subtype. The precision metric indicates the proportion of correctly predicted instances among those predicted as positive for each class, while recall represents the proportion of correctly predicted instances among all actual instances of each class. The F1-score, a harmonic mean of precision and recall, offers a balanced measure of a model's performance. Across the lung cancer subtypes, the precision, recall, and F1-score values vary, reflecting the model's varying ability to correctly classify instances for each class. For instance, while the precision for lung squamous cell carcinoma is relatively high at 31.15%, indicating a relatively low rate of false positives, the recall and F1-score values are lower, suggesting room for improvement in correctly identifying instances of this class. Similarly, lung adenocarcinoma and lung benign tissue exhibit comparable precision, recall, and F1-score values, indicating the model's challenges in accurately distinguishing between these classes.

In conclusion, while the CNN model demonstrates promising performance in classifying lung cancer subtypes, the analysis of the confusion matrix and classification report highlights areas for refinement and improvement. By addressing these challenges, such as mitigating class imbalance and enhancing the model's ability to distinguish between

closely related subtypes, the model's accuracy and effectiveness can be further enhanced, ultimately contributing to more accurate diagnosis and treatment of lung cancer.

VI. CONCLUSION

This paper presents a comprehensive investigation into the application of Convolutional Neural Networks (CNNs) for the automated classification of lung cancer subtypes based on histopathological images. The study aimed to address the pressing need for accurate and timely diagnosis of lung cancer, a leading cause of mortality worldwide, by leveraging advanced machine learning techniques to enhance the classification of lung cancer subtypes. Through meticulous dataset curation and augmentation, a diverse collection of histopathological images representing lung tissue's morphological characteristics was assembled. This dataset served as the foundation for training and evaluating the CNN model, ensuring its robustness and generalization capability across various pathological conditions. The CNN model architecture was meticulously designed, leveraging multiple convolutional layers, max-pooling layers, and fully connected layers to learn hierarchical representations of lung tissue morphology and facilitate accurate classification. The training and evaluation of the CNN model yielded promising results, with high accuracy achieved in distinguishing between different lung cancer subtypes. The model demonstrated impressive training and validation accuracies of 96.11% and 88.2%, respectively, underscoring its efficacy in histopathological image analysis. However, a detailed analysis of the confusion matrix revealed challenges in accurately distinguishing between certain lung cancer subtypes, particularly lung adenocarcinoma and lung squamous cell carcinoma. The presence of class imbalance further compounded these challenges, highlighting the need for refinement and optimization of the model. Despite these challenges, the CNN model's performance represents a significant advancement in automated histopathological image analysis for lung cancer classification. By augmenting the expertise of medical professionals, CNN-based approaches offer the potential to expedite diagnosis, tailor treatment strategies, and improve patient outcomes. The high accuracy achieved by the model underscores its viability as a complementary tool in the diagnostic workflow, augmenting human expertise and facilitating more informed clinical decisions. In conclusion, this paper contributes to the growing body of literature on automated histopathological image analysis for lung cancer classification. By leveraging advanced machine learning techniques, the study demonstrates the potential of CNN-based approaches to enhance diagnostic accuracy and improve patient care in the context of lung cancer diagnosis and treatment.

REFERENCES

[1] G. Qu, D. Zhang, and P. Yan, “Medical image fusion by wavelet transform modulus maxima,” Optics Express, vol. 9, no. 4, pp. 184–190, 2001.
[2] R. Tekade, “Lung nodule detection and classification using machine learning techniques,” Asian Journal for Convergence in Technology, vol. 4, 2018.
[3] W.-J. Choi and T.-S. Choi, “Automated pulmonary nodule detection based on three- dimensional shape-based feature descriptor,” Computer Methods and Programs in Biomedicine, vol. 113, no. 1, pp. 37–54, 2014.
[4] I. R. S. Valente, P. C. Cortez, E. C. Neto, J. M. Soares, V. H. C. de Albuquerque, and J. M. R. S. Tavares, “Automatic 3D pulmonary nodule detection in CT images: a survey,” Computer Methods and Programs in Biomedicine, vol. 124, pp. 91–107, 2016.
[5] L. Hussain, M. S. Almaraashi, W. Aziz, N. Habib, and S. U. R. Saif Abbasi, “Machine learning-based lungs cancer detection using reconstruction independent component analysis and sparse filter features,” Waves in Random and Complex Media, pp. 1–26, 2021.

[6] T. V. Pyrkov, K. Slipensky, M. Barg et al., “Extracting biological age from biomedical data via deep learning: too much of a good thing?” Scientific Reports, vol. 8, no. 1, pp. 1–11, 2018.
[7] M. Phankokkruad, “Ensemble transfer learning for lung cancer detection,” in 2021 4th International Conference on Data Science and Information Technology, pp. 438–442, Association for Computing Machinery, 2021.
[8] A. Onan, “On the performance of ensemble learning for automated diagnosis of breast cancer,” in Artificial Intelligence Perspectives and Applications: Proceedings of the 4th Computer Science On-Line Conference 2015 (CSOC2015), Vol 1: Artificial Intelligence Perspectives and Applications, pp. 119–129, Springer, 2015.
[9] W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn, "Automatic Lung Cancer Prediction from Chest X-ray Images Using the Deep Learning Approach," 2018 11th Biomedical Engineering International Conference (BMEiCON), Chiang Mai, 2018, pp. 1-5, doi: 10.1109/BMEiCON.2018.8609997.
[10] T. Atsushi, T. Tetsuya, K. Yuka, and F. Hiroshi. (, 2017). “Automated Classification of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks”. BioMed Research International. 2017. 1-6. 10.1155/2017/4067832.
[11] W. Rahane, H. Dalvi, Y. Magar, A. Kalane and S. Jondhale, "Lung Cancer Detection Using Image Processing and Machine Learning HealthCare," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, 2018, pp. 1-5, doi: 10.1109/ICCTCT.2018.8551008.
[12] M. Šarić, M. Russo, M. Stella and M. Sikora, "CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images," 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 2019, pp. 10.23919/SpliTech.2019.8783041. 1-4, doi:
[13] S. Sasikala, M. Bharathi, B. R. Sowmiya. "Lung Cancer Detection and Classification Using Deep CNN." (2019).
[14] SRS Chakravarthy and H. Rajaguru. "Lung Cancer Detection using Probabilistic Neural Network with modified Crow-Search Algorithm." Asian Pacific Journal of Cancer Prevention, 20, 7, 2019, 10.31557/APJCP.2019.20.7.2159. 2159-2166.