Detection of Lung Cancer by Using Hybrid Classification Algorithm of CT Scan Images

Shazia Shamas, Ishu Sharma, Surya Narayan Panda

Abstract— **A powerful hybrid AI-driven classification method with the goal of revolutionizing the diagnosis of lung cancer using CT scan pictures. In the process, integrating various machine learning algorithms such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree and Artificial Neural Network techniques, this novel algorithm aims to significantly enhance accuracy in lung cancer identification. The investigation extensively utilizes the Lung Image Database, comprising meticulously documented CT scans meticulously annotated for precise nodule locations, facilitating in-depth analysis for lung cancer detection. Throughout the rigorous evaluation process, the ANN consistently showcases outstanding precision, recall, F-measure, and accuracy metrics, significantly surpassing other tested models such as DT, KNN, and SVM. Across both dataset distributions (70:30 and 80:20), the ANN showcased exceptional metrics: recall (0.9240 / 0.9486), precision (0.9240 / 0.9486), F-measure, and accuracy (93.15% / 95.50%), surpassing Decision Trees, K-Nearest Neighbors, and SVM. Among the models examined, the ANN consistently stood out, demonstrating unparalleled performance in accurately identifying and delineating lung cancer instances.**

Index Terms— **Lung Cancer Analysis, Image Processing, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Artificial Neural Network**

I. Introduction

One of the most prevalent and lethal types of cancer in the world is lung cancer. If lung cancer is identified early, the likelihood of a suitable outcome can be significantly increased [1]. Using CT scan imaging, machine-learning algorithms were created to help interpret these images, which have become a potent tool for the timely identification of pulmonary cancer. Machine learning can be used to amend lung cancer detection's precision and effectiveness. Utilizing vast collections of clinical and medical imaging data, these algorithms identify patterns and predict lung cancer risks. The algorithms may reduce the need for invasive diagnostic procedures through the early detection of lung cancer. [2, 3]. Figure 1 shows the major tasks of digital image processing [1].

Manuscript received February 29, 2024; revised December 3, 2024.

Shazia Shamas is a PhD candidate of Chitkara University Institute of Engineering, Chitkara University, Rajpura, Punjab, India (email: shaziabhat6115@gmail.com)

Ishu Sharma is an Associate Professor of Department of Computer Science and Engineering, Chandigarh Engineering College, Chandigarh Group of Colleges Jhanjeri, Mohali, Punjab, India (corresponding author to provide phone: +91-701-541-8009; e-mail: ishu.sharma001@gmail.com)

Surya Narayan Panda is a Professor of Chitkara University Institute of Engineering, Chitkara University, Rajpura, Punjab, India (email: snpanda@chitkara.edu.in)

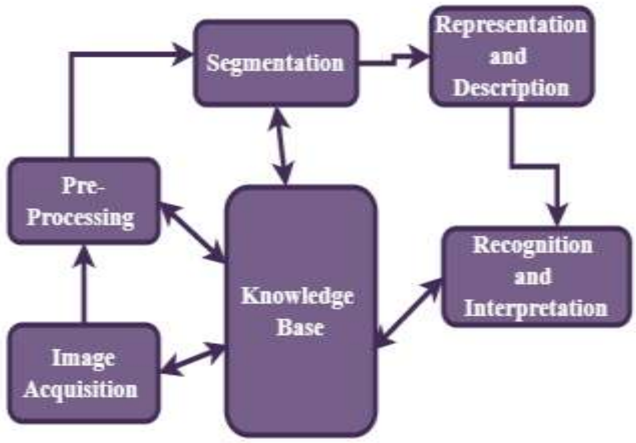

Fig. 1. Primary Objectives in Computational Image Manipulation.

The identification and diagnosis of lung cancer. Medical practitioners are able to spot suspected tumors or other abnormalities with detailed images that CT scans of the lungs may create [4]. The CT pictures can be improved by image processing methods, making it simpler to see these abnormalities and monitor how they change over time. Texture analysis algorithms, for instance, might find changes in the texture or density of lung tissue that can be signs of cancer [5, 6]. Segmentation algorithms, meanwhile, can isolate the lung tissue and spot probable tumors or nodules. These image-processing methods could help diagnose lung cancer more accurately and quickly, which might result in earlier detection and better treatment. Further study and improvement are required to improve these algorithms' performance and guarantee their reliable incorporation into clinical practice, as the correctness of these algorithms depends on the caliber of the CT scans and the skill of the medical specialists analyzing them [7,8].

To enhance the accuracy and resilience of pulmonary cancer identification. The paper uses hybrid classification algorithms for CT scan images to incorporate different machine learning algorithms. These algorithms often combine feature extraction and classification approaches, with the classification model using features retrieved from the CT scans as inputs. For feature extraction, a hybrid approach is used i.e. scale invariant feature transform and particle swarm optimization. The selected features using SIFT and PSO are used as the input feature set for classification into normal and abnormal samples. The classifiers evaluated here are decision tree, k-nearest neighbor, support vector machine and artificial neural networks. These machine learning classifiers were used to distinguish normal and abnormal samples. The performance of the proposed work methodology is computed in terms of precision, recall, f-measure and accuracy by which the two classes get detected. The comparative analysis of average performance values of different classifiers shows that ANN outperforms the other classifiers.

II. Literature Review

This section contains a thorough analysis of the pertinent literature. A variety of algorithms have been investigated by numerous researchers with the goal of identifying lung cancer. The level of exploration and research into these algorithms, meanwhile, has been rather constrained.

Arulmurugan and Ananda kumar [9] suggested a system that combines the classification power of an ANN with wavelet function descriptors. The system initially applies wavelet transformation to the incoming data. After that, the data is processed, and statistical features including energy, contrast, entropy, and autocorrelation are retrieved. The neural network classifier uses these measured attributes as input parameters. For training purposes, the neural network classifier is set up to make use of both feed-forward and feedback propagation network architectures. There are several training programs used, including Trained, Traingda, Traingdm, and Traingdx. In terms of accuracy, the feed-forward classification falls short of the feedback propagation network. With a precision rate of 92.6%, specificity of 100%, flexibility of 91.2%, and a mean square error of 0.978, the classification system produced promising results. These results show that wavelet function descriptors and ANN classification work well together for the task at hand [9]. Kumar et al. [10] assessed how well five optimization methods performed when it came to extracting tumors from lung pictures. Among the techniques that were tested were GCPSO, inertia-weighted particle swarm optimization, and particle swarm optimization (PSO). When the study examined the efficiency of median, adaptive median, and average filters in the pre-processing stage, the adaptive median filter was shown to be the best appropriate for medical CT images. Additionally, the researchers employed adaptive histogram equalization to enhance picture contrast. The better-quality pre-processed photos were then subjected to the four optimization methods. The useful outcomes were verified using twenty MATLAB lung imaging samples. The results of the investigation demonstrated that, at 95.89%, the GCPSO algorithm had the highest tumour extraction accuracy [10].Sweetlin et al. [11] suggested the use of a computer-aided diagnostic (CAD) system to improve the consistency and accuracy of the interpretation of pulmonary TB images. The system employs a wrapper approach that combines a one-against-all SVM classifier with the cuckoo search optimization algorithm to determine which subset of characteristics is the best. The cuckoo search method was applied both with and without the entropy measure in order to choose the most relevant characteristics. The selected characteristics are then used to train the one-against-all SVM classifier. Out of the 98 features retrieved from the sick areas, 47 features are chosen using the entropy measure, resulting in an accuracy of 92.77%. With an accuracy of 91.89%, 51 characteristics are selected without the use of the entropy measure. These results imply that the features used for training have a major impact on the classifier's performance [11]. Lenin et al. [12] centered on integrating Fuzzy C-Means (FCM) with a variety of optimization approaches (ABC, Firefly, Cuckoo, and SA) for picture segmentation. For the purpose of segmenting abdominal CT images using FCM, a novel optimization method known as Crow Search (CS) Optimization was presented and contrasted with previous methods. The methods were built in MATLAB 2015a after the researchers carried out a comprehensive examination to choose suitable validation functions for clustering approaches. Furthermore, a Raspberry Pi B+ embedded board was used to implement the FCM-CS algorithm in hardware. In FCM-based picture segmentation, the CS optimization approach outperformed competing algorithms and demonstrated promising results. The work investigates hardware implementation on embedded boards and advances picture segmentation [12]. Jena et al. [13] aimed to improve the accuracy of early disease diagnosis through the use of various techniques. The researchers applied Gaussian and Wiener filters to a large volume of scanned images using the LIDC dataset to reduce noise. By merging adjacent pixels based on seed points, they used region growing segmentation to precisely identify the Region of Interest (ROI). Subsequently, relevant characteristics like area, perimeter, entropy, intensity, and statistical-based features were extracted from the segmented areas. A deep Gaussian mixture model in a region-based convolutional neural network (DGMM-RBCNN) was utilized to decrease dimensionality and offer a flexible and nonlinear model for describing the picture data. At each layer, Gaussian-based dimensionality reduction was used to prevent overly parameterized solutions. Metrics used in performance evaluation included Martin's correlation coefficient, F-measure, accuracy, sensitivity, and specificity. The model trained and tested the image samples in a MATLAB environment, and during the 18th epoch, it achieved an accuracy of around 87.79% [13].He et al. [14] examined an ANN algorithm model that was used to create a model for the identification of lung cancer. The model conducted a comparative experiment to verify its accuracy and used image segmentation algorithms to locate the lung cancer lesion area. Lung cancer was detected by the ANN-based approach with 94.6% accuracy, 95.7% sensitivity, and 93.5% specificity. Using a combination of image retrieval techniques and lung cancer image segmentation algorithms, the study successfully displayed the lesion area [14]. Guleria et al. [15] carried out a study in which they built prediction models for determining the class of hypothyroidism using a variety of machine learning-based methodologies. Among these techniques were naive Bayes, decision trees, random forests, and multiclass classifiers. They also used an ANN deep learning model, which is well known for its efficiency in handling text data. The authors compared the effectiveness of the various classifiers as well as their suggested model with earlier works. Their model achieved an accuracy of 93.8226%, which was higher than that of other studies in the field. With accuracy rates of 99.5758% and 99.3107%, respectively, and very low error rates of 0.0424 and 0.0689, the decision tree and random forest were found to perform best, according to the performance evaluation [15].Karthick et al [16] presented the S-FNGC method, which improves color picture segmentation. It employed shape priority and connection measure as a thresholding technique to separate items from the background. The approach employed a normalized fuzzy network cut measure based on the S membership function to tackle structural faults in color pictures. The S-FNGC algorithm provided details about the object's involvement in the image boundary by using a system of S fuzzy sets. It addressed difficulties like inaccurate segmentation and poor accuracy. A comparison of the S-FNGC approach with other methods, such as mask thresholding, Gabor filter, GA, and K-Means clustering algorithm, showed that it performed better while having the fewest misclassification errors and error rates, which improved color image segmentation [16].

Jian et al [17] aimed to increase retinal vascular segmentation accuracy for more accurate identification of cardiovascular disorders. For feature extraction in medical picture segmentation, a Dual-Branch encoder structure based on the U-Net model was included into the Dual-Branch-U Net framework. The framework utilized a parallel encoder with various convolutional modules to enhance the feature extraction process and generate richer semantic data. Instead of pooling, convolution operations were used for lower sampling to control the step size and enable effective information fusion. An attention module was implemented in the decoder stage to reduce image noise and filter out irrelevant features. When tested on the DRIVE and ARIA datasets, the suggested strategy outperformed five other cutting-edge methods for precisely segmenting retinal arteries [17].

Lee et al. [18] intended to address the shortcomings in particular medical data, which is frequently sparse and inconsistent. It recommended a preprocessing technique to increase deep learning models' ability to detect cardiovascular illnesses from CT scans automatically. During the preprocessing phase, the CT images were split into areas of interest and uninteresting regions using the Grabcut technique and K-means clustering. Three sets of data were used to assess the efficacy of deep learning: the original data, data treated just using K-means clustering, and data processed with both K-means clustering and the Grabcut technique. With IRB approval, the study made use of data from Korea's Soonchunhyang University Cheonan Hospital. Training using the VGG, Inception ResNet V2, and Resnet 50 models showed that Resnet 50 had the highest validation and testing accuracy. The proposed preprocessing strategy demonstrated a considerable improvement in deep learning model accuracy, ranging from 10% to 40%. All things considered, the study offered a helpful preprocessing method to address the limitations of specific medical data and improve the effectiveness of deep learning models [18].

III. Research Methodology

In this proposed scheme, the technology of image processing is to use to detect the lung cancer. Actually, this is done by using lung CT scans. Then the cancer detection task is accomplished into four stages so that the lung cancer can be localized and classified. All phases of lung cancer detection are summed up as.

1) Pre-processing: In this phase, lung CT scans image is used as input and Adaptive Intensity Adjustment and using limited contrast stretching techniques are used.
2) Segmentation: This second phase applied region-based segmentation approach for segmenting the similar and dissimilar portions in the CT scans. By means of k-means with cuckoo search algorithm.
3) Feature Extraction: Scale invariant feature transform with particle swarm optimization is applied so that the valuable features can be extracted from the lung CT scan image.
4) Classification: The final stage of classification implements different classifiers to identify the classifier that yields better outcomes among all and is more suitable for this research work . The ANN outer performs the other classifiers.

Several methods are employed to identify lung cancer in its early stages using image processing techniques. A method used to manipulate picture-based data and extract the most valuable information is called image processing. Signal processing and image processing are comparable in that they both take pictures as input, process them, and provide an enhanced image as the result [3]. Image processing is now advancing quickly and has taken over as the primary area of study for engineers. It consists of the three primary processes listed below:

1) Use image acquisition tools to import images.
2) Examine and handle the input photos, and
3) Output the enhanced image for classification purpose.

Two technologies are being used to process the images: digital image processing (DIP) and analogue image processing. For physical copies like printed pictures and photographs, analogue image processing is used. With the use of computers and software, digital processing of the picture is used to process digital images. The primary steps in digital image processing are information extraction, augmentation, pre-processing, and presentation [2]. Figure 2 illustrates the general procedures used in the digital picture processing. Figure 2 illustrates the general procedures used in the digital picture processing.

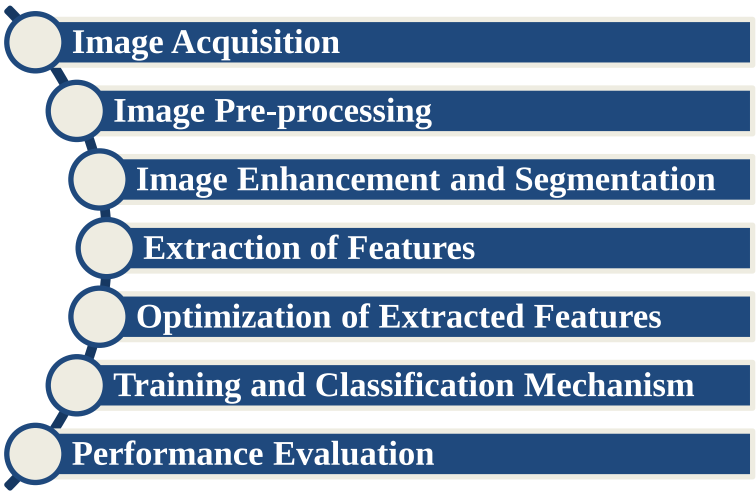

Fig. 2. Steps of Image Processing and Classification.

A. Simulation Parameters

The parameters used in the simulation analysis of the proposed work are discussed below.

Precision measures how well a model predicts the favorable outcomes. To put it another way, precision is the ratio of actual positive forecasts to all of the model's positive predictions. It quantifies the quantity of positive results from model. This is calculated by use of below mentioned formula.

$$Precision = \frac{TP}{TP+FP} \quad \textbf{(1)}$$

Recall also called sensitivity or true positive rate are employed in categorization tasks. It assesses a model's capacity to locate every pertinent instance within a dataset, paying particular attention to true positive predictions. It is also known as recall or true Positive rate calculator. It is the ratio of correctly identified outputs to the sum of actual correct values. This is calculated by use of below mentioned formula.

$$Recall = \frac{TP}{TP+FN} \quad \textbf{(2)}$$

The harmonic mean of recall and accuracy is used to calculate F-meaure value. Since the F-measure is a harmonic mean, recall and accuracy are factors that affect its value. When recall and accuracy are identical, it can reach a maximum value of 1.

$$F-measure = \frac{2*Precision*Recall}{(Precision+Recall)} \quad \textbf{(3)}$$

The most important indicator that is used for calculating the performance of the model is accuracy. It is actually calculating the fraction of prediction the model got right. The equation used for the calculation is stated below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \textbf{(4)}$$

$$Where, TP \rightarrow\ True\ Positive,\ TN \rightarrow\ True\ Negative,\ \ FP \rightarrow\ False\ Positive, \text{ and } FN \rightarrow False\ Negative.$$

The analysis of the model efficiency is based on the values of parameters discussed in the section.

IV. Results And Discussion

The segmentation and classification approaches created for lung cancer detection and prognosis are evaluated. The results obtained from the implemented segmentation approaches are used to refine the overall classification model. The best outcomes are used to assure the approach for segmentation stage and the outcomes are fed to the classification model. The performance analysis is discussed in detail, encompassing its performance metrics, predictive accuracy and execution time.

A. Precision Analysis

The Precision Analysis of Segmentation Techniques in Table I offers a detailed breakdown of the precision metrics derived from various segmentation methodologies employed in lung cancer detection. This table serves as a comprehensive comparison platform, allowing an in-depth assessment of each technique's precision performance in isolating lung cancer regions within medical imaging data.

This analysis strongly highlights the superior performance of k-means with csa, The Figure 3. Showcases the Average Precision Analysis of Segmentation Techniques, comparing the performance of segmentation methods like K-means, K-means with PSO, ABC, FFA, and K-means with CSA in identifying lung cancer regions within medical images across various sample sizes. The data illustrates that K-means with CSA consistently achieves higher average precision compared to other segmentation techniques across different sample sizes. Starting at 0.817 and reaching 0.907, K-means with CSA consistently outperforms other methods, demonstrating its effectiveness in accurately segmenting lung cancer regions. This analysis strongly highlights the superior performance of K-means with CSA in achieving higher average precision throughout different sample sizes, establishing it as a robust and reliable method for accurately identifying lung cancer regions within medical images. In Figure 3 the provided percentages represent the superior performance of K-means with CSA in lung cancer segmentation compared to other techniques. Demonstrating an 11.05% improvement over standard K-means, this method stands out for its significant advancements in accurately identifying and delineating lung cancer instances.

Table I. Precision Analysis Of Segmentation Techniques

Sample Images	K-Means	K-Means with PSO	K-means with ABC	K-means with FFA	K-means with CSA
10	0.789	0.835	0.865	0.867	0.878
20	0.793	0.838	0.874	0.871	0.879
40	0.8	0.847	0.877	0.876	0.886
60	0.804	0.854	0.879	0.88	0.89
80	0.808	0.857	0.885	0.887	0.894
100	0.809	0.859	0.889	0.888	0.9
150	0.812	0.868	0.89	0.895	0.902
200	0.816	0.876	0.896	0.898	0.91
250	0.823	0.879	0.899	0.903	0.912
300	0.825	0.884	0.9	0.908	0.916
350	0.832	0.885	0.905	0.916	0.922
400	0.833	0.894	0.907	0.921	0.927
450	0.839	0.899	0.908	0.927	0.935
500	0.844	0.901	0.911	0.935	0.939

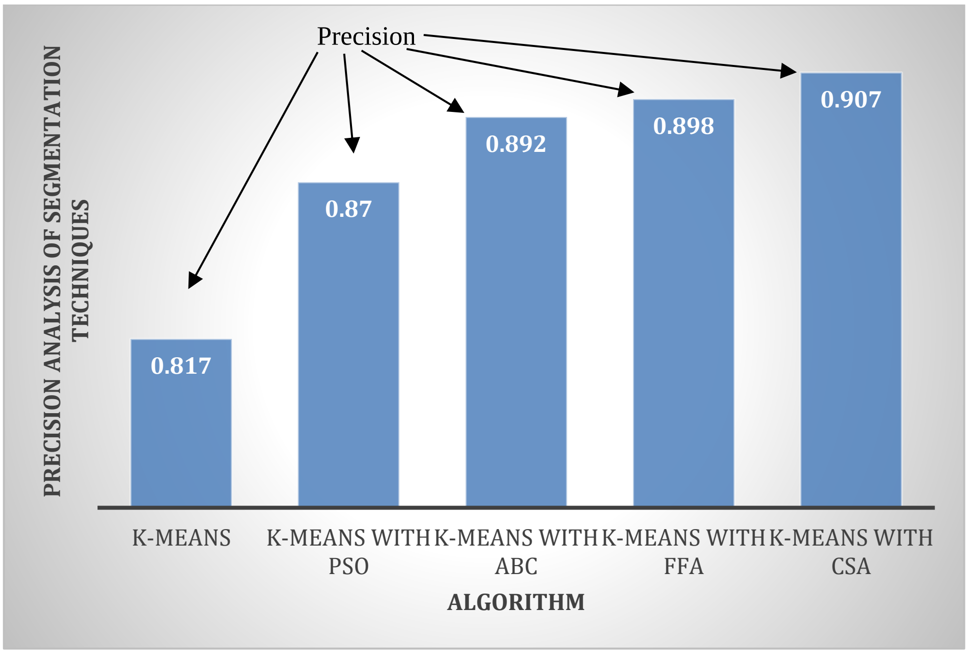

Fig. 3. Average precision analysis of segmentation techniques.

Additionally, compared to K-means with PSO, K-means with CSA showcases a 4.23% enhancement, highlighting its substantial superiority in refining segmentation accuracy. Similarly, when compared to K-means integrated with ABC and FFA, K-means with CSA exhibits improvements of 1.63% and 0.94%, respectively. These figures underscore the considerable lead of K-means with CSA, emphasizing its efficacy in enhancing precision and accuracy in lung cancer segmentation over these other specialized methodologies.

In Figure 4. The provided percentages represent the superior performance of K-means with CSA in lung cancer segmentation compared to other techniques. Demonstrating an 11.05% improvement over standard K-means, this method stands out for its significant advancements in accurately identifying and delineating lung cancer instances.
Additionally, compared to K-means with PSO, K-means with CSA showcases a 4.23% enhancement, highlighting its substantial superiority in refining segmentation accuracy. Similarly, when compared to K-means integrated with ABC and FFA, K-means with CSA exhibits improvements of 1.63% and 0.94%, respectively. These figures underscore the considerable lead of K-means with CSA, emphasizing its efficacy in enhancing precision and accuracy in lung cancer segmentation over these other specialized methodologies.

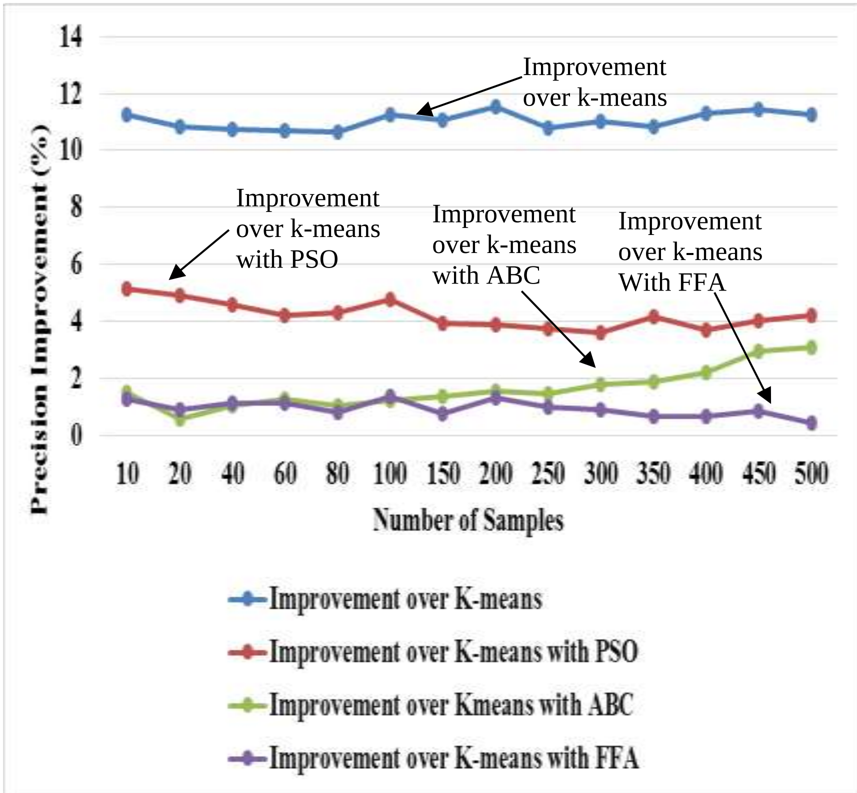

Fig. 4. Improvement Precision Analysis of Segmentation Techniques.

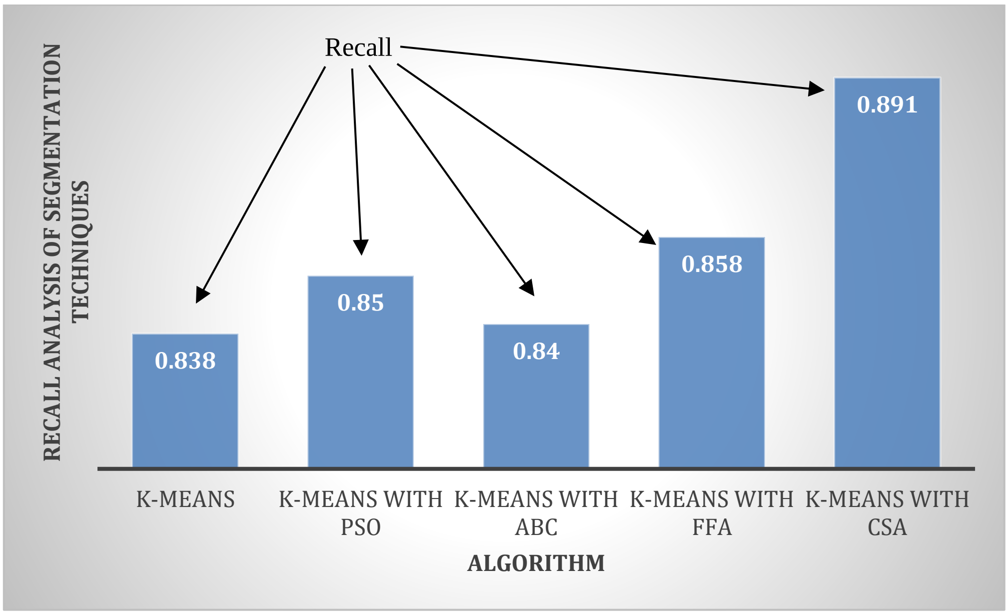

Fig. 5. Average Recall Analysis of Segmentation Techniques.

B. Recall Analysis

In Figure 5, which represents the Average Precision Analysis of Segmentation Techniques, K-means with CSA stands out with the highest precision value of 0.891 compared to other methodologies. This precision metric indicates the accuracy in identifying lung cancer instances within the dataset. In contrast, alternative techniques like K-means, K-means with PSO, K-means with ABC, and K-means with FFA show precision values ranging between 0.838 and 0.858. The substantial margin between K-means with CSA and other methods underscores its superior performance in precise lung cancer segmentation. Overall, the figure clearly demonstrates that K-means with CSA outperforms other segmentation techniques, showcasing its robustness and effectiveness in accurate lung cancer instance delineation. In Figure 6. the improvement percentages of K-means with CSA over alternative segmentation techniques, K-means, K-means with PSO, K-means with ABC, and K-means with FFA, demonstrate its superior performance in lung cancer segmentation. K-means with CSA showcases significant enhancements, exhibiting a 6.47% improvement over K-means, 4.88% over K-means with PSO, 6.17% over K-means with ABC, and 3.85% over K-means with FFA. These values emphasize the substantial lead of K-means with CSA, highlighting its robustness and effectiveness in refining the accuracy and precision of lung cancer instance identification compared to established segmentation techniques.

C. F-measure Analysis

In Figure 7, the Average F-measure Analysis of Segmentation Techniques provides insights into the accuracy of various methodologies in identifying lung cancer instances within the dataset. K-means with CSA notably stands out, showcasing the highest average F-measure of 0.899 among the compared segmentation techniques, K-means, K-means with PSO, K-means with ABC, and K-means with FFA. Comparatively, other techniques exhibit slightly lower average F-measure values: K-means with a score of 0.827, K-means with PSO at 0.86, K-means with ABC at 0.865, and K-means with FFA at 0.878. However, K-means with CSA's substantially higher average F-measure value signifies its superior accuracy in identifying lung cancer instances compared to these methods. This indicates that K-means with CSA is better suited for this task, offering enhanced accuracy in identifying and segmenting lung cancer instances compared to the other examined techniques.

.

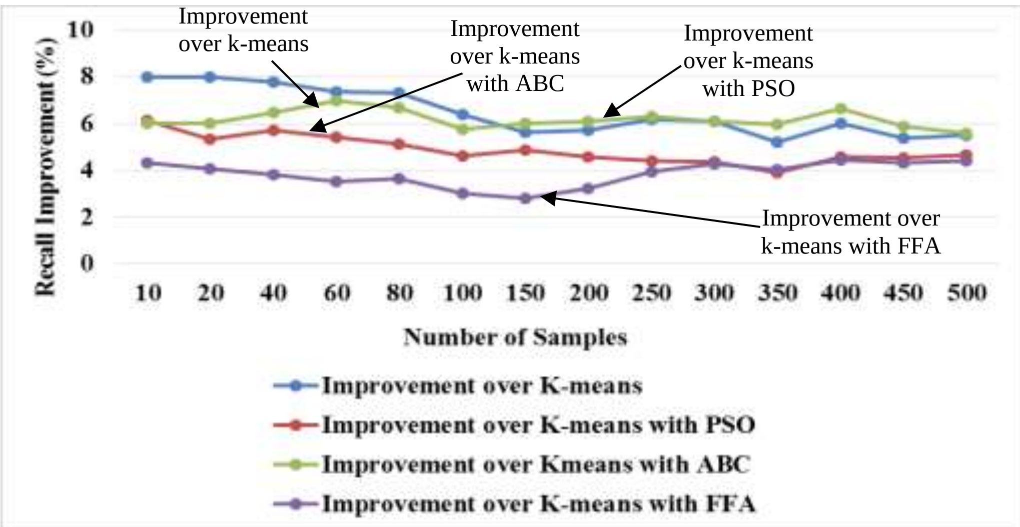

Fig. 6. Improvement Recall Analysis of Segmentation Techniques.

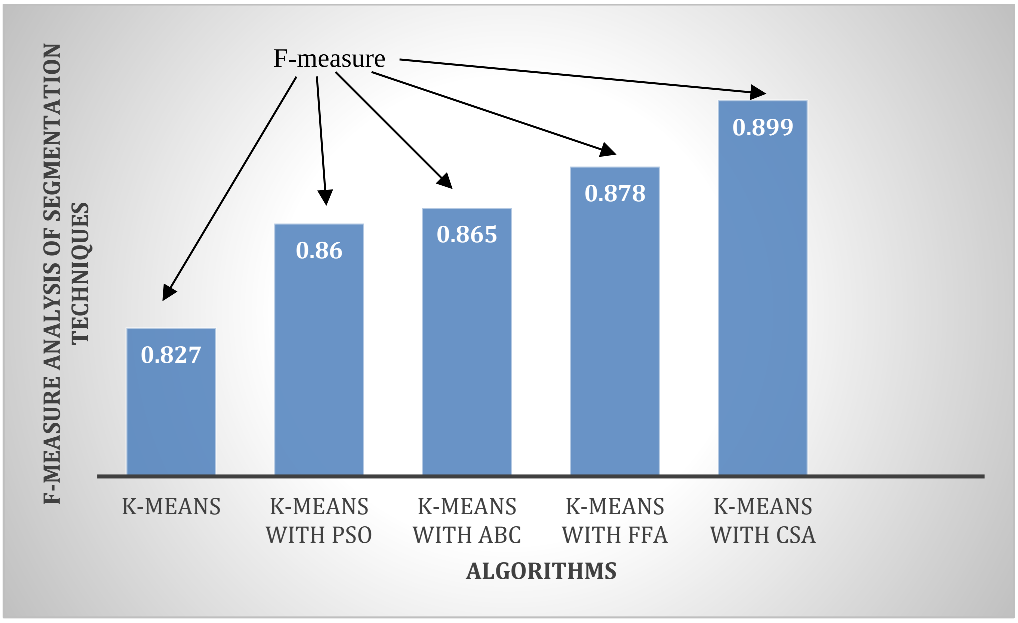

Fig. 7. Average F-measure Analysis of Segmentation Techniques.

Figure 8, depicting the Improvement F-measure Analysis of Segmentation Techniques, showcases the remarkable advancements achieved by K-means with CSA over alternative methodologies, K-means, K-means with PSO, K-means with ABC, and K-means with FFA in accurately identifying lung cancer instances within the dataset. The improvement values emphasize the substantial performance gaps favoring K-means with CSA, an impressive 8.74% enhancement over K-means, a notable 4.56% lead over K-means with PSO, a 3.92% improvement over K-means with ABC, and a 2.41% leap over K-means with FFA. These underscore the consistent superiority of K-means with CSA, highlighting its efficacy in refining accuracy and precision in lung cancer instance identification compared to other methodologies, positioning it as a frontrunner in this domain.

D. Accuracy Analysis

The Average Accuracy Analysis of Segmentation Techniques, depicted in Figure 9, presents a comparative evaluation of diverse methodologies for identifying lung cancer regions within a dataset. Among the techniques assessed, K-means with CSA emerges as the standout performer, showcasing significantly higher average accuracy compared to K-means, K-means with PSO, K-means with ABC, and even K-means with FFA. Starting at 81.268, K-means sets the baseline for accuracy, while K-means with CSA remarkably outperforms all other techniques, culminating in an accuracy of 93.83. This notable disparity in accuracy underscores the robustness and efficacy of K-means with CSA in precisely identifying lung cancer regions within medical images. The substantial gap in accuracy values emphasizes the method's superiority and its potential significance in medical diagnostics. This analysis underlines the pivotal role of segmentation techniques in medical imaging analysis, particularly in lung cancer identification. K-means with CSA showcases remarkable promise, offering substantially higher average accuracy compared to other methodologies considered in this evaluation.

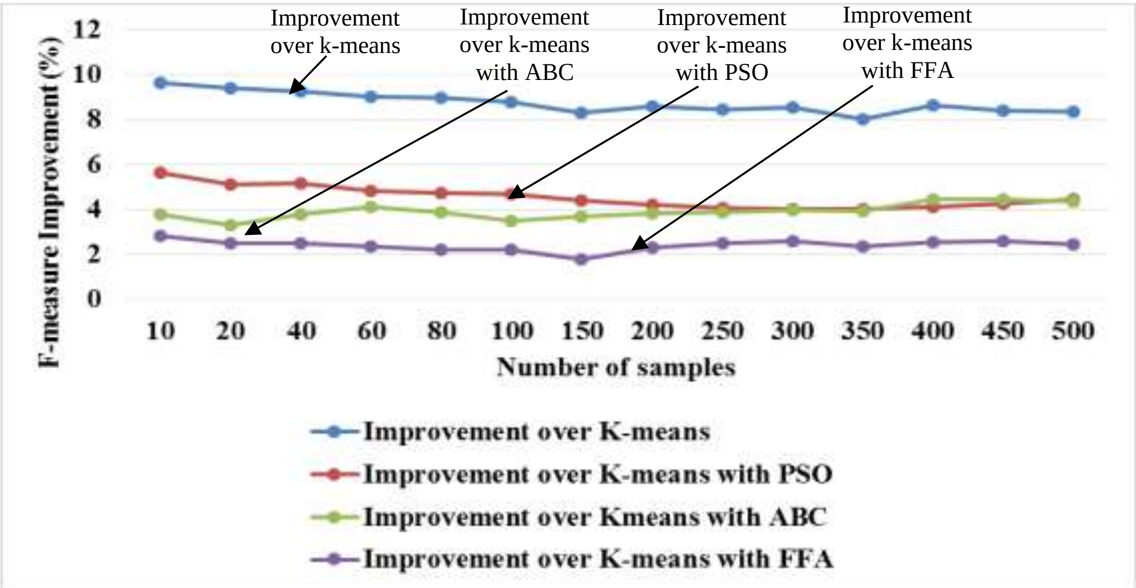

Fig. 8. Improvement F-measure Analysis of Segmentation Techniques.

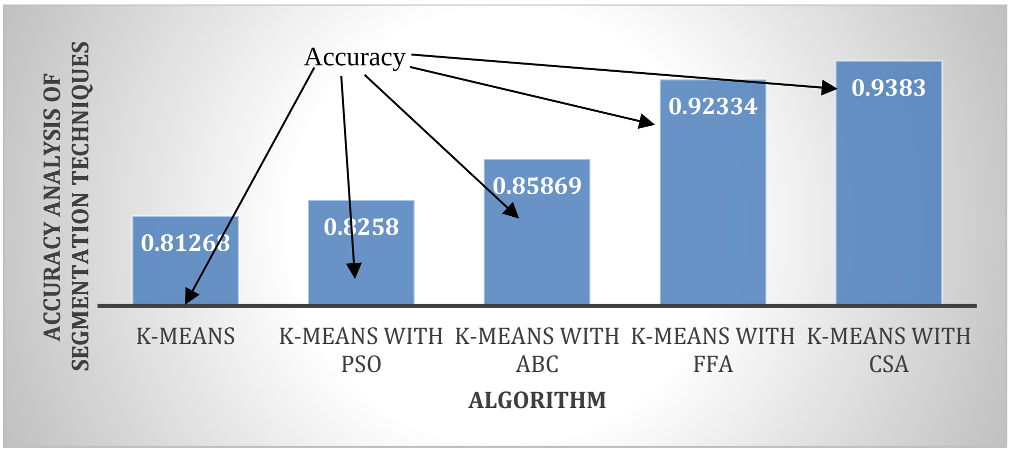

Fig. 9. Average Accuracy Analysis of Segmentation Techniques.

The Improvement Accuracy Analysis, illustrated in Figure 10, outlines the substantial performance enhancements of K-means with CSA over other segmentation techniques like K-means, K-means with PSO, K-means with ABC, and K-means with FFA in accurately identifying lung cancer regions within datasets. The improvement values vividly illustrate the significant strides made by K-means with CSA, showcasing an advancement of 15.47% over K-means, 13.63% over K-means with PSO, 9.29% over K-means with ABC, and 1.63% over K-means with FFA. These substantial margins emphasize the pronounced effectiveness of K-means with CSA in achieving higher accuracy compared to alternative methods, particularly in precisely delineating lung cancer regions in medical images. This clear distinction in improvement values reaffirms the superiority of K-means with CSA, indicating its ability to notably enhance accuracy in identifying lung cancer instances. This analysis underscores the potential of K-means with CSA as a promising approach for improving accuracy in lung cancer identification, thus offering considerable promise in advancing diagnostic precision and subsequent medical interventions.

E. Analysis of Machine Learning based Lung Cancer Classification

To present a comprehensive simulation analysis, the data is divided into two distributions using four classifiers.

70:30 Dataset distribution: In this, 70% of the data is used for the training and learning of the system, and 30% is used for the direct testing of the system. The performance analysis is further, summarized in the next sections.

80:20 Dataset distribution: In this dataset distribution, 80% of the data is used for the training of the system while 20% data is reserved for the testing and the performance is evaluated for four performance metrics, namely precision, recall, f-measure and accuracy.

F. Precision Analysis

The precision improvement analysis, as illustrated in Fig 11, showcases significant strides made by different machine learning models in comparison to the referenced ANN within the context of lung cancer classification. The findings reveal substantial advancements: 6.89% enhancement over DT, 4.17% improvement over K-NN, and 2.35% boost over SVM. These improvements emphasize the selected methodology's efficacy in achieving more precise and reliable lung cancer classifications. The varied improvements across algorithms highlight their unique strengths and capabilities within this specific context. Ultimately, these insights offer valuable guidance for tailored model selection, potentially leading to more refined and accurate lung cancer diagnosis methodologies and improved patient outcomes.

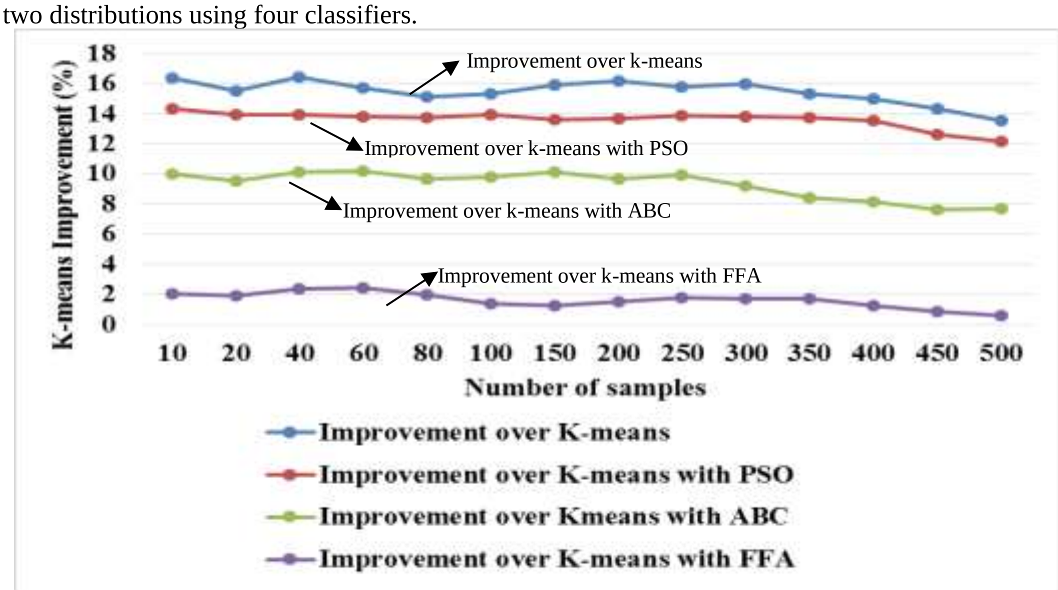

Fig. 10. Improvement Accuracy Analysis of Segmentation Techniques

Table II.Analysis of Machine Learning based Lung Cancer Classification

Sample Images	DT	KNN	SVM	ANN
10	0.8641	0.8882165	0.9006303	0.9230567
20	0.865106	0.8921989	0.902483	0.9269329
40	0.86868	0.8927591	0.9067354	0.9270733
60	0.871208	0.8954844	0.9072042	0.9286658
80	0.872908	0.896837	0.9085074	0.9298326
100	0.875229	0.8978071	0.9128708	0.9335901
150	0.877572	0.8990183	0.9150535	0.9377428
200	0.87849	0.8991941	0.9162113	0.9404841
250	0.879095	0.9034083	0.9181335	0.9448038
300	0.883354	0.9050878	0.9227224	0.9465817
350	0.885535	0.9082404	0.9254794	0.9476081
400	0.886333	0.9083824	0.929864	0.948793
450	0.889082	0.9106796	0.9318532	0.9489045
500	0.891041	0.9114123	0.9351163	0.9505099
		80:20 Distribution		
10	0.8895	0.8940114	0.9120519	0.9239248
20	0.890698	0.8971519	0.9149486	0.9278771
40	0.892251	0.8990347	0.9190011	0.9324453
60	0.892917	0.8999083	0.9194381	0.9336355
80	0.894016	0.9010292	0.9210351	0.9369384
100	0.896703	0.9012467	0.9246246	0.9382364
150	0.897577	0.9050973	0.9271314	0.9389645
200	0.898354	0.9063255	0.9300074	0.9423829
250	0.898544	0.908011	0.9317744	0.9457787
300	0.899793	0.9108408	0.9319673	0.9493111
350	0.903864	0.912637	0.932505	0.9506963
400	0.906497	0.9152802	0.9360221	0.9583658
450	0.908674	0.9196991	0.9371856	0.9655451
500	0.910693	0.924252	0.9380161	0.9688778

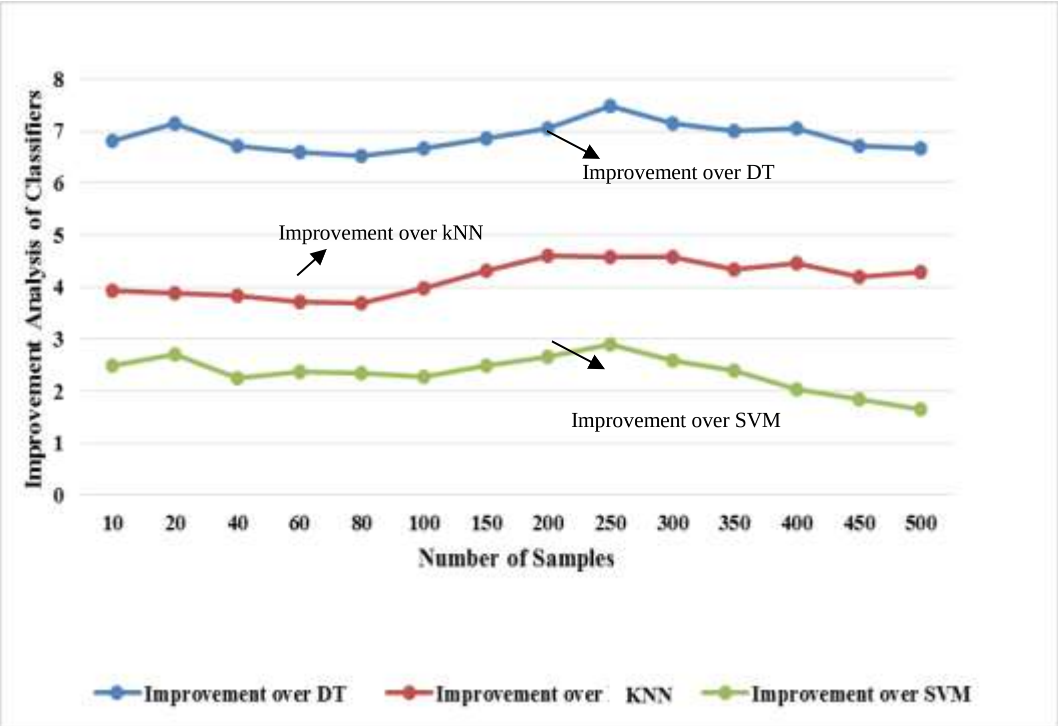

Fig. 11. Improvement Analysis of Precision for 70:30 Distribution.

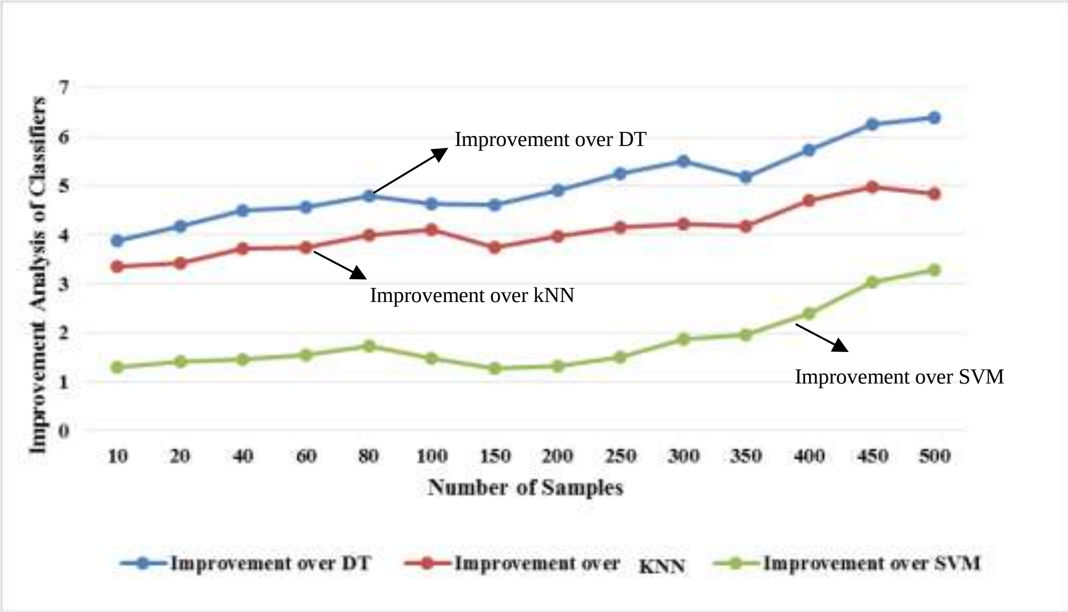

Fig. 12. Improvement Analysis of Precision for 80:20 Distribution.

The improvement analysis of precision as represented in Figure 12 for an 80:20 distribution presents compelling insights into the comparative advantages of proposed ANN over various machine learning models. The results showcase substantial strides in performance metrics, where the ANN method surpasses the other framework by 5.03%, demonstrating its robustness and efficacy in capturing intricate patterns within the dataset. Notably, the improvement over K-NN by 4.08% signifies a considerable enhancement in precision, emphasizing the superiority of the chosen methodology in handling classification tasks, especially in scenarios reliant on proximity-based algorithms. Moreover, the noteworthy 1.82% improvement over SVM underscores the nuanced intricacies the selected method encapsulates, outperforming a widely-used algorithm known for its versatility in high-dimensional spaces. These findings not only underscore the effectiveness of the selected approach but also emphasize its potential in delivering superior precision rates, crucial in domains demanding accurate and reliable predictions.

In Figure 13, the comparison of average recall values between the ANN and alternative machine learning techniques across the 70:30 and 80:20 distributions highlight distinct performance patterns in lung cancer classification., in the 70:30 distributions, the ANN maintains superior recall at 0.9240, surpassing DT (0.8862), K-NN (0.8943), and SVM (0.9126). Similarly, within the 80:20 split, the ANN consistently exhibits higher recall values compared to DT, K-NN, and SVM. Starting at 0.9486 for ANN, the recall decreases gradually for DT (0.9064), K-NN (0.9210), and SVM (0.9278). These findings underscore the ANN's consistent advantage in accurately capturing positive instances within lung cancer datasets across both distributions.

G. *Recall Analysis*

Recall Analysis for lung cancer classification across two distributions 70:30 and 80:20 distributions highlight distinct performance in lung cancer classification in 70:30 distribution, The ANN outperforms

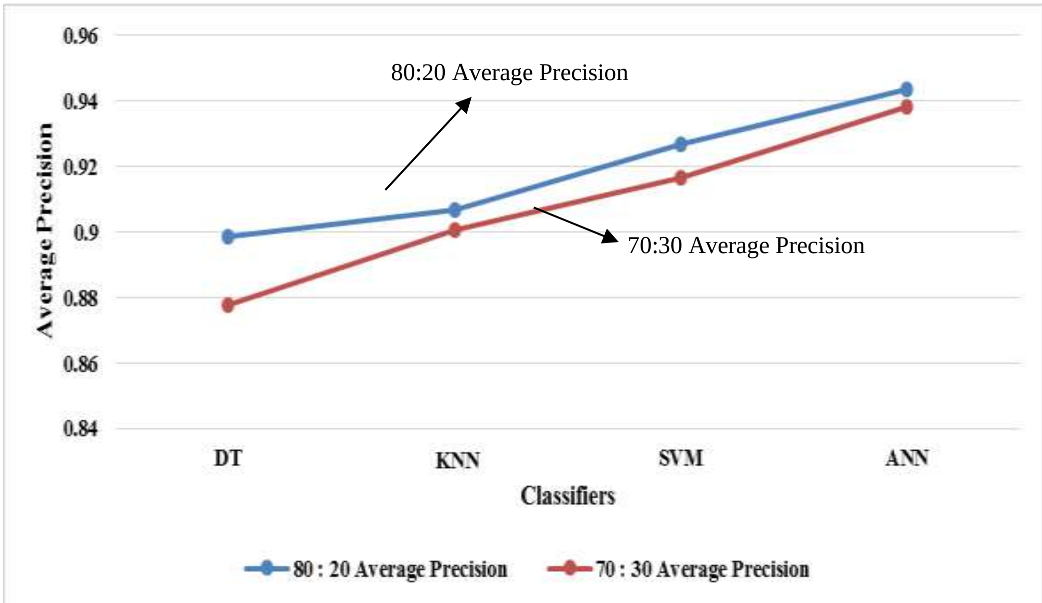

Fig. 13. Average Comparison of Precision.

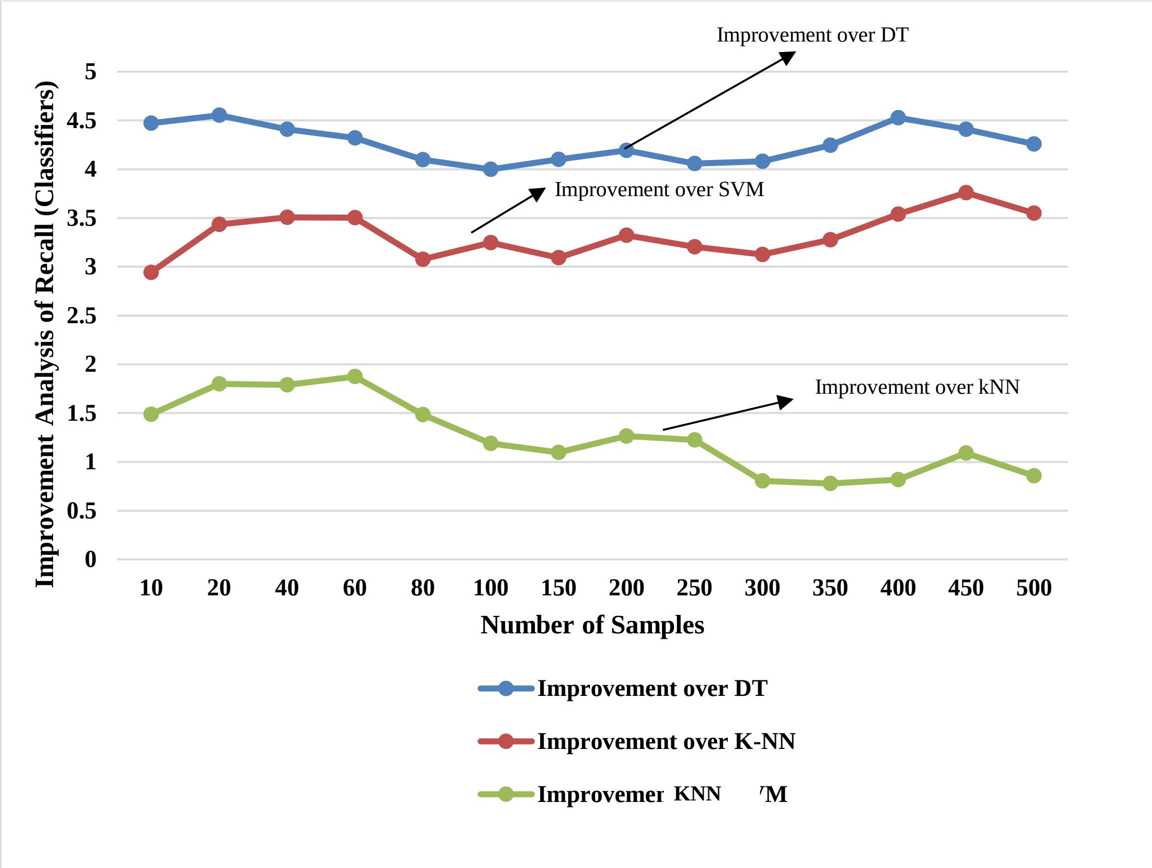

Fig. 14. Analysis of Recall for 70:30 Distribution

In Figure 14, the improvement analysis of recall for lung cancer classification showcases how the proposed ANN compared to various machine learning models performs. The recorded improvement percentages, 4.65% over DT, 2.99% over K-NN, and 2.24% over SVM, highlight the superior performance of these models in accurately identifying In Figure 15, the comparison of average precision values between the ANN and other techniques across 70:30 and 80:20 distributions is visually depicted. This graphical representation illustrates how different machine learning models perform concerning precision metrics within these specific dataset splits. For the 70:30 distribution, the ANN maintains a higher precision starting at 0.9240 compared to DT (0.8862), K-NN (0.8943), and SVM (0.9126). Similarly, in the 80:20 split, the ANN demonstrates notably higher precision values compared to DT, K-NN, and SVM. Starting from 0.9486 for ANN, the precision values gradually decrease for the other techniques, with DT at 0.9064, K-NN at 0.9210, and SVM at 0.9278 The depiction in Figure 16 underscores the significance of the ANN as a preferred choice for precision-driven classification tasks, providing a compelling argument for its superiority over DT, K-NN, and SVM in terms of precision in the context of this analysis.

H. F-measure Analysis

The Improvement Analysis for the 70:30 F-measure distribution as illustrated in Figure 16 reveals substantial advancements of the proposed ANN over the alternative machine learning models in lung cancer classification. The recorded improvements as 5.57% over DT, 3.74% over K-NN, and 1.80% over SVM underscore the efficacy of these positive instances within lung cancer datasets. These improvements demonstrate the efficacy of ANN over the alternative models in achieving better recall rates compared to the baseline. In the realm of lung cancer classification, where precise identification of malignancies is crucial, the ANN emphasize the potential to better detect cancerous cells models in achieving higher F-measure values. These percentages indicate significant enhancements of ANN in precision and recall balance for identifying lung cancer instances compared to these techniques. This data emphasizes the potential superiority of the ANN over diverse models in accurately identifying malignancies within the 70:30 dataset split. The Improvement Analysis of F-measure for lung cancer classification in the 80:20 distribution shown in Figure 17 demonstrates a significant advancement of ANN over the alternative machine learning models. The reported improvement percentages, 4.84% over DT, 3.54% over K-NN, and 2.03% over SVM highlight these models' superior performance in achieving higher F-measure values. These findings emphasize the potential of these models to enhance the balance between precision and recall in identifying lung cancer instances, crucial for accurate diagnoses and improved patient outcomes. It provides valuable insights for model selection, aiming to achieve better and more comprehensive diagnoses in lung cancer classification, thereby potentially contributing to improved patient outcomes. In Figure 18, the comparison of average F-measure values across the 70:30 and 80:20 and distributions represent the performance differences among the ANN and alternative machine learning techniques in lung cancer classification.

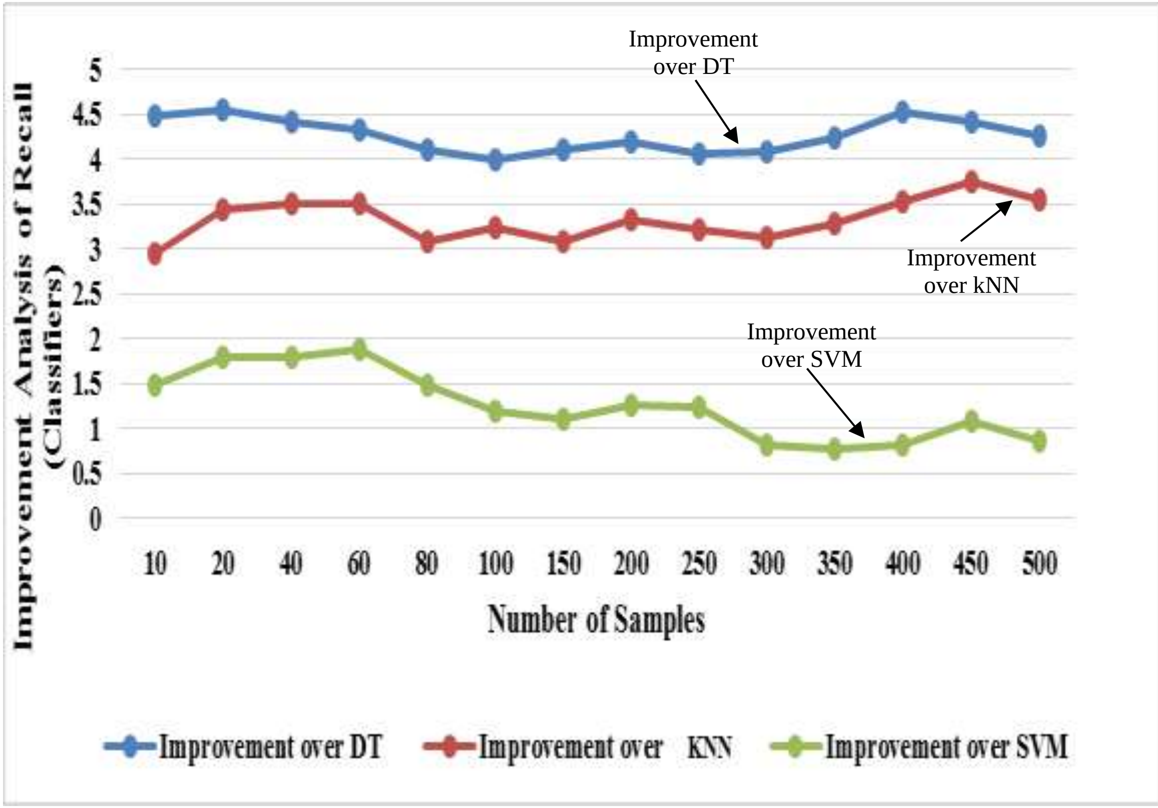

Fig. 15. Improvement Analysis of Recall for 80:20 Distribution

These metrics shed light on the distinct trends observed across both distributions. Within, the 70:30 distribution, the ANN maintains a superior F-measure of 0.9311, outperforming DT (0.8819), K-NN (0.8974), and SVM (0.9146). Similarly, in the 80:20 split, the ANN consistently demonstrates higher F-measure values compared to DT, K-NN, and SVM. Starting at 0.9462 for ANN, the F-measure decreases gradually for DT (0.9025), K-NN (0.9138), and SVM (0.9273).

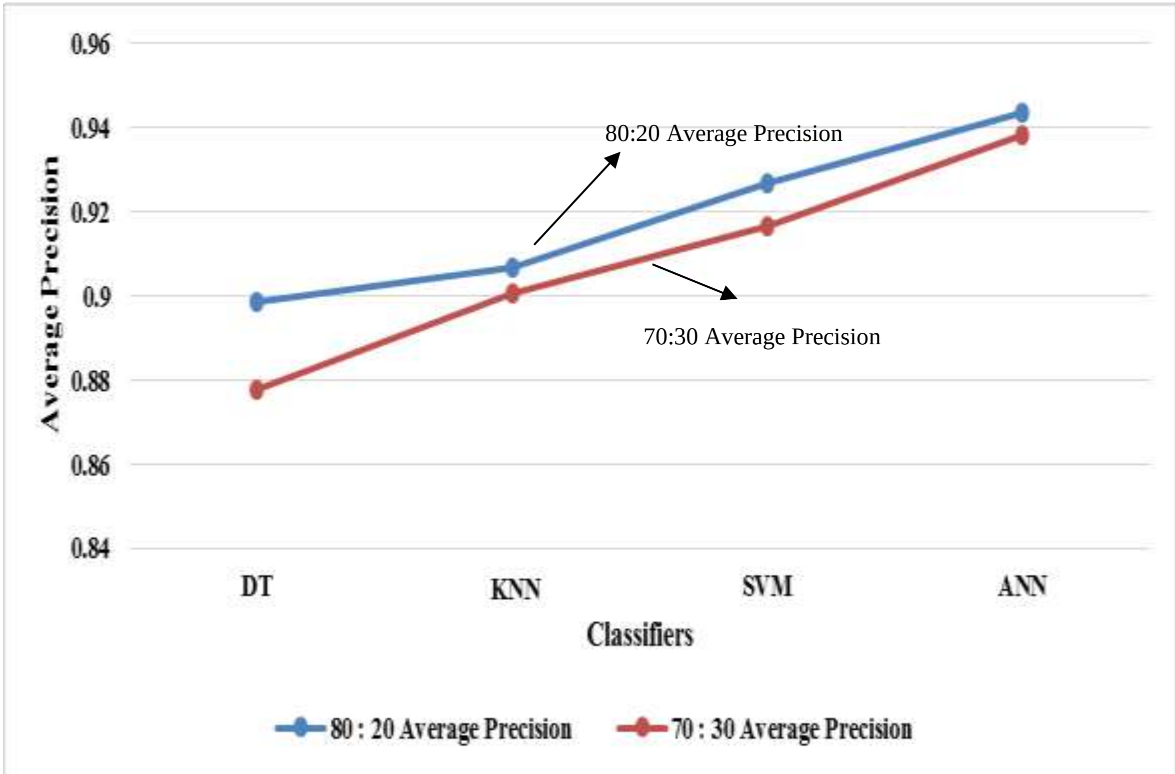

Fig. 16. Average Comparison of Recall.

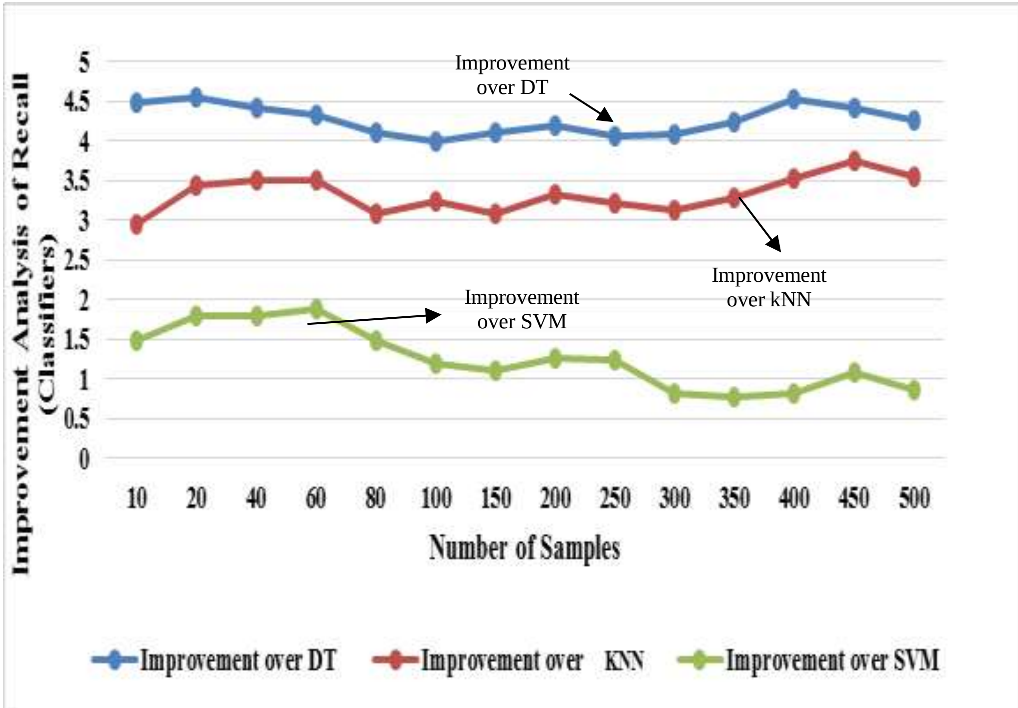

Fig. 17. Improvement Analysis of F-measure for 70:30 Distribution.

I. Accuracy Analysis

The Accuracy Analysis of various machine learning models concerning lung cancer classification across two distinct distributions 70:30 and 80:20. It provides comprehensive view of how different machine learning models perform in classifying lung cancer cases across varied data splits.

In Figure 19, the improvement percentages as, 3.14% over DT, 1.73% over K-NN, and 1.52% over SVM highlighted in the improvement analysis of accuracy within the 70:30 distribution, demonstrate the superior performance of ANN compared to the alternative models in lung cancer classification. While the improvements are comparatively lower than in some other distributions, they still indicate the potential of these models to provide increased accuracy rates in identifying lung cancer instances. This underscores the importance of exploring ANN over various machine learning approaches to optimize accuracy in lung cancer classification tasks within different dataset distributions.

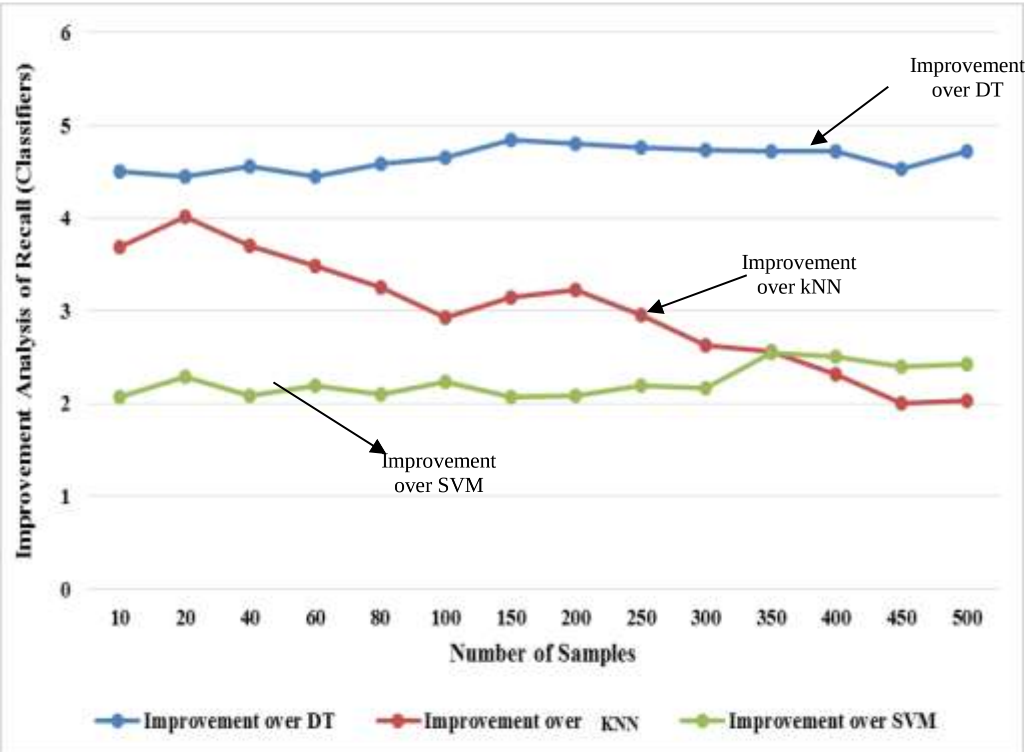

Fig. 18. Improvement Analysis of F-measure for 80:20 Distribution.

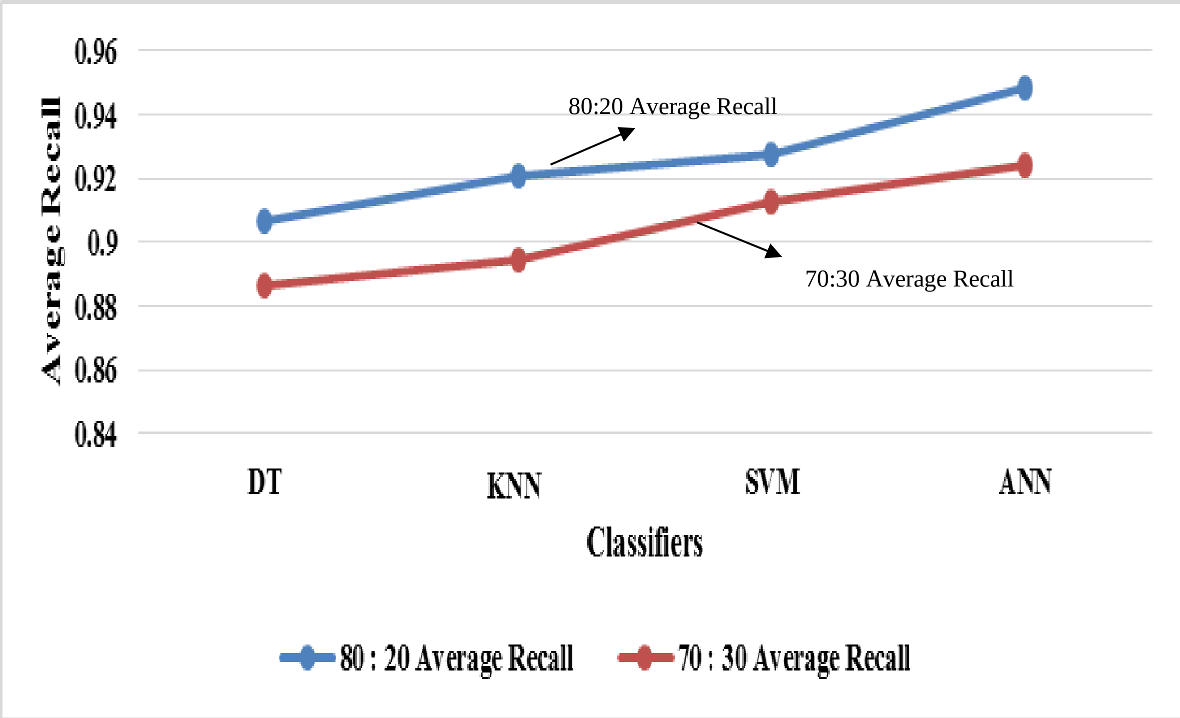

Fig. 19. Average Comparison of F-measure.

In Figure 20, the improvement values are depicted, showcasing 3.99% enhancement over DT, 2.04% over K-NN, and 2.34% over SVM compared to the proposed ANN within the 80:20 accuracy distribution for lung cancer classification. These figures underscore the alternative models' increased performance, suggesting the potential of ANN to outperform the other models in accuracy rates for identifying lung cancer instances. This highlights the significance of exploring diverse machine learning models, as depicted in Figure 20. To potentially achieve superior accuracy in lung cancer classification tasks.

In Figure 22, the comparison between average accuracy values across the 70:30 and 80:20 distributions demonstrate the performance variations between the ANN and alternative machine learning techniques in lung cancer classification. Within, the 70:30 distributions, the ANN starts at 93.15%, surpassing DT (90.31%), K-NN (91.57%), and SVM (91.75%). Similarly, the 80:20 split, the ANN exhibits higher accuracy compared to DT, K-NN, and SVM. Starting at 95.50% for ANN, accuracy decreases gradually for DT (91.83%), K-NN (93.59%), and SVM (93.31%). These results highlight the ANN's advantage in accuracy over alternative models across both distributions. However, within the 70:30 split, all models showcase lower accuracy rates. This emphasizes the importance of dataset configurations in model performance. While the ANN consistently outperforms other models, the comparison indicates the potential competitiveness of alternative models in specific dataset distributions. Thus, the findings in Figure 21 underscore the significance of selecting appropriate models tailored to dataset characteristics for optimizing lung cancer classification accuracy.

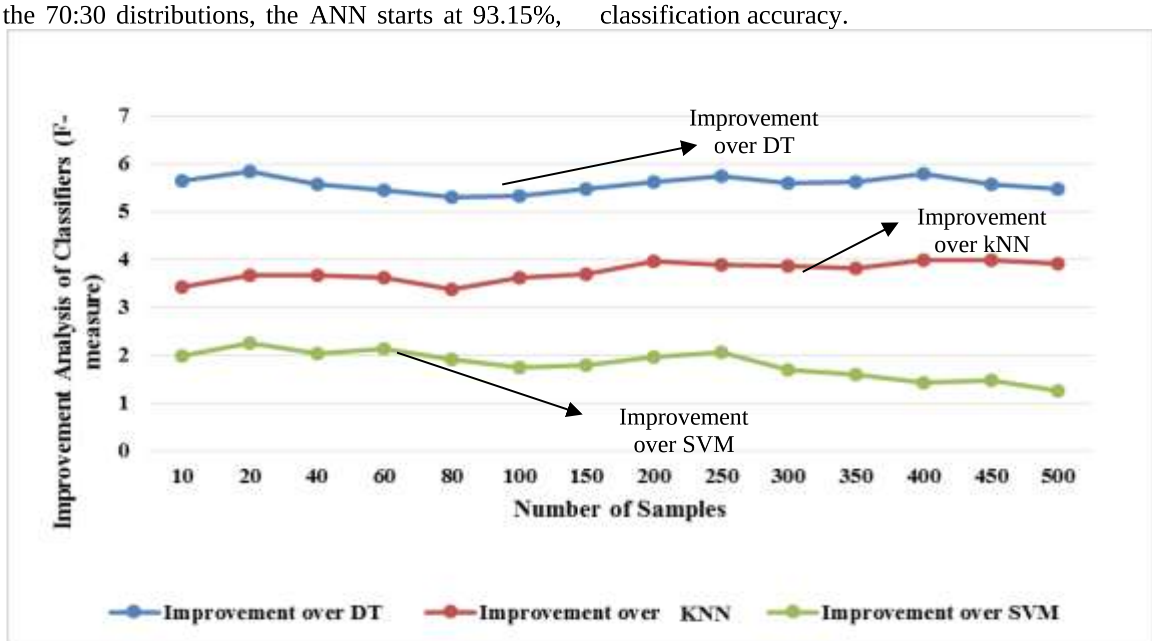

Fig. 20. Improvement Analysis Accuracy for 70:30 Distribution.

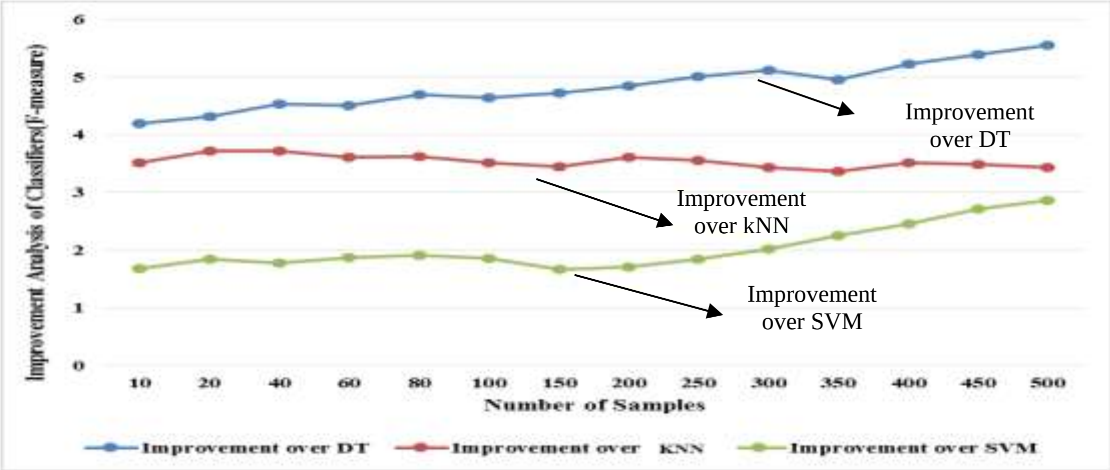

Fig. 21. Improvement Analysis of Accuracy for 80:20 Distribution.

J. Execution Time Analysis

The execution time for the proposed model is evaluated while considering both the training to testing dataset distributions in the following section.

K. Training and Testing Analysis for 70:30 Distribution

The below graph presents a comprehensive analysis of time consumed in Training and Testing phases utilizing a 70:30 distribution strategy. This distribution method dedicates 70% of the dataset for training the machine learning models, while the remaining 30% is utilized for direct testing. The analysis across this dataset distribution provides invaluable insights into the performance and efficacy of the models in lung cancer classification.

L. Overall Training Time Analysis

Table III showcases the training time analysis across two distinct dataset distributions, utilizing 70% and 80% of the available data for model training. This comparison aims to explore the performance variations and model adaptability concerning varying data proportions. Figure 24 and Table III present a comparative analysis of the training times for various classifiers using 70% and 80% of the dataset for model training. The training times are depicted in hours for DT, K-NN, SVM, and ANN. In the 70% training scenario, the classifiers took 2.79 hours to 2.87 hours for training, with ANN requiring slightly more time at 2.88 hours. However, when the training data increased to 80%, the training times across all classifiers increased, with ANN taking the longest at 3.04 hours. This comparison indicates that as the dataset size increases, especially with 80% of the data, the training time for all classifiers, Particularly ANN, escalates, suggesting higher computational demand for training larger datasets. Despite its longer duration for training, ANN potentially yield superior predictive accuracy and demonstrate proficiency in recognizing intricate patterns within lung cancer data.

M. Overall Testing Time Analysis

This comparison aims to highlight the impact of dataset size on model performance during the testing phase. The testing time analysis in both Figure 23 and Table III highlights the computational durations for various classifiers on lung cancer data. Across DT, K-NN, SVM, and ANN, the testing time is consistently longer for the 20% testing dataset compared to the 30% dataset.

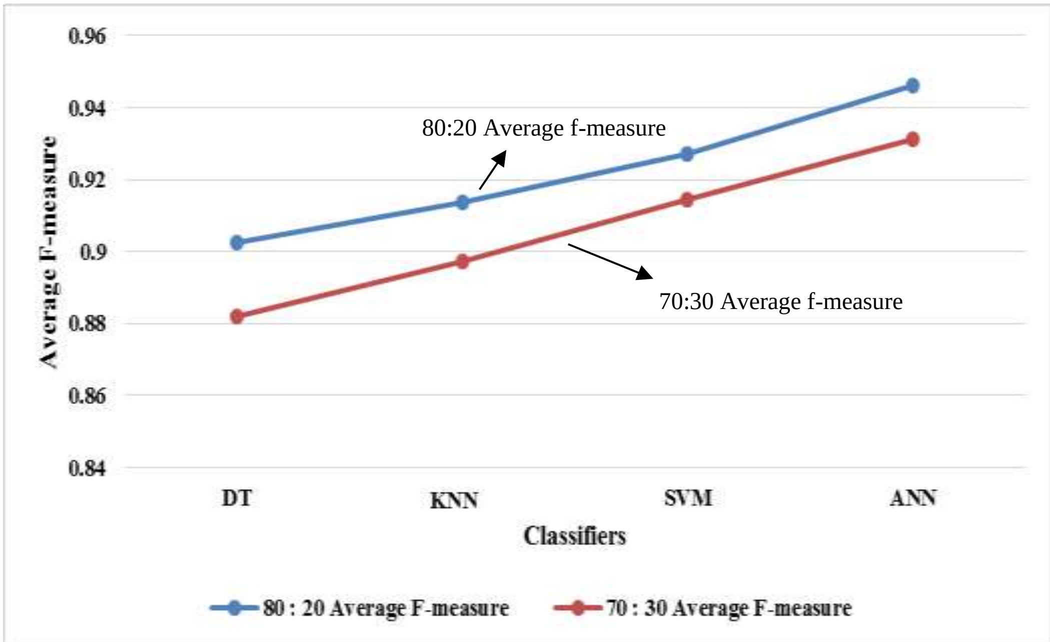

Fig. 22. Average Comparison of Accuracy.

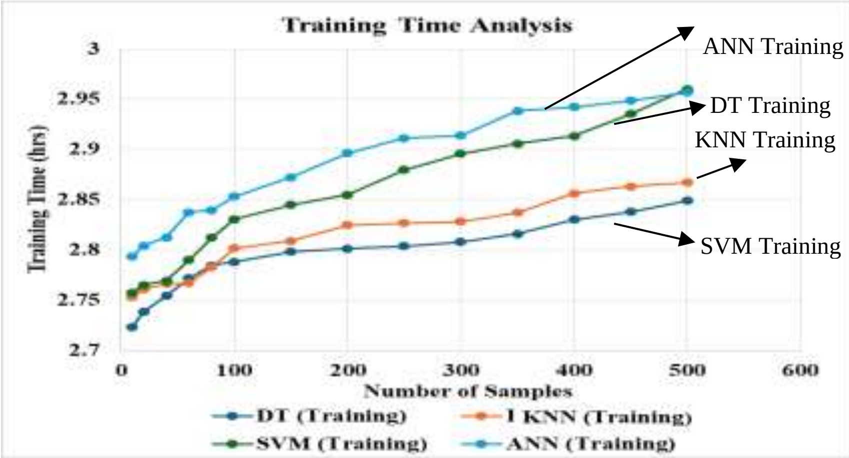

Fig. 23. Training Time Analysis using 70:30 Distribution

For the 30% testing data, the values for DT, K-NN, SVM, and ANN are approximately 38.57819, 39.165571, 39.236453, and 39.908907 milliseconds, respectively. Conversely, for the 20% dataset, the corresponding values are approximately 38.006227, 39.24751, 39.826375, and 40.445804 milliseconds. This illustrates that the 20% dataset consistently requires more computational time for evaluation across all classifiers. Specifically, the ANN classifier exhibits notably longer testing durations compared to other techniques in both scenarios. These results underscore the impact of dataset size on computational load during testing, with larger datasets demanding more time for evaluation, and ANN showcasing extended testing durations in both 30% and 20% dataset scenarios. The testing time analysis demonstrates that although the ANN classifier requires more computational time compared to other techniques for lung cancer classification, it showcases a commendable ability to predict lung cancer accurately

N. Comparative Analysis

Table V provides a comparative analysis of various classifiers across different years, focusing on their Precision, Recall, F-measure, and Accuracy metrics. This table offers an insightful overview of how different classifiers from distinct studies perform across these crucial evaluation criteria. The Comparative Analysis of Different Classifiers, as depicted in Table V, provides an insightful overview of various models' performance metrics sourced from different studies across multiple years. In 2018, Arulmurugan and Anandakumar introduced the FFBPNN, achieving commendable Precision, Recall, and F-measure scores of 0.9124, 0.8934, and 0.9028, respectively. The subsequent year, Sweetlin et al. presented a SVM model with improved Precision, Recall, and F-measure scores of 0.9142, 0.9012, and 0.9076535. In 2021, Jena et al. introduced the RBCNN with a competitive performance, securing scores of 0.9045, 0.9128, and 0.908631 for Precision, Recall, and F-measure.

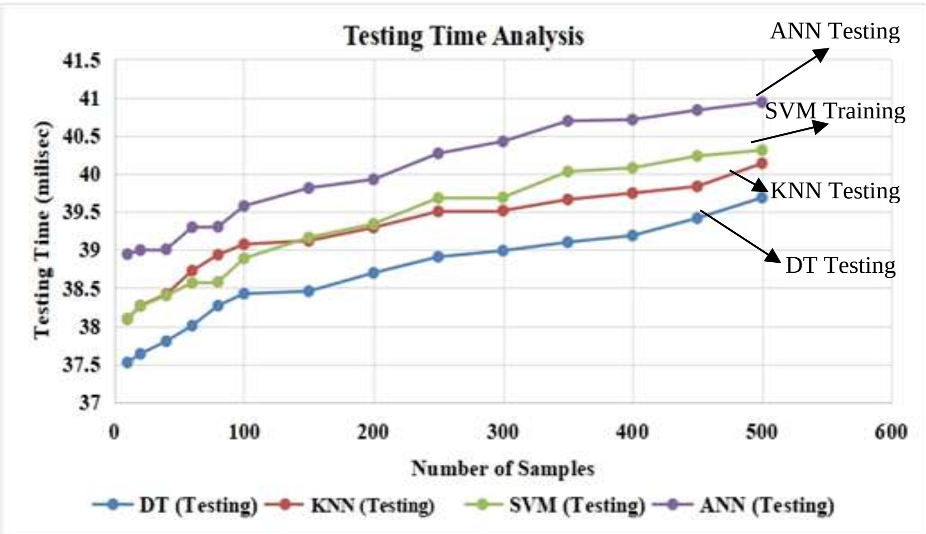

Fig. 24. Testing Time Analysis using 70:30 Distribution

TABLE III. COMPARATIVE ANALYSIS OF THE TRAINING TIMES FOR VARIOUS CLASSIFIERS USING 70% AND 80% OF THE DATASET FOR MODEL TRAINING.

	DT (hrs)	KNN (hrs)	SVM (hrs)	ANN (hrs)
70% Training	2.793 2157	2.8102555	2.8509576	2.8798604
80% Training	2.904 2244	2.9710403	2.9610269	3.0357692

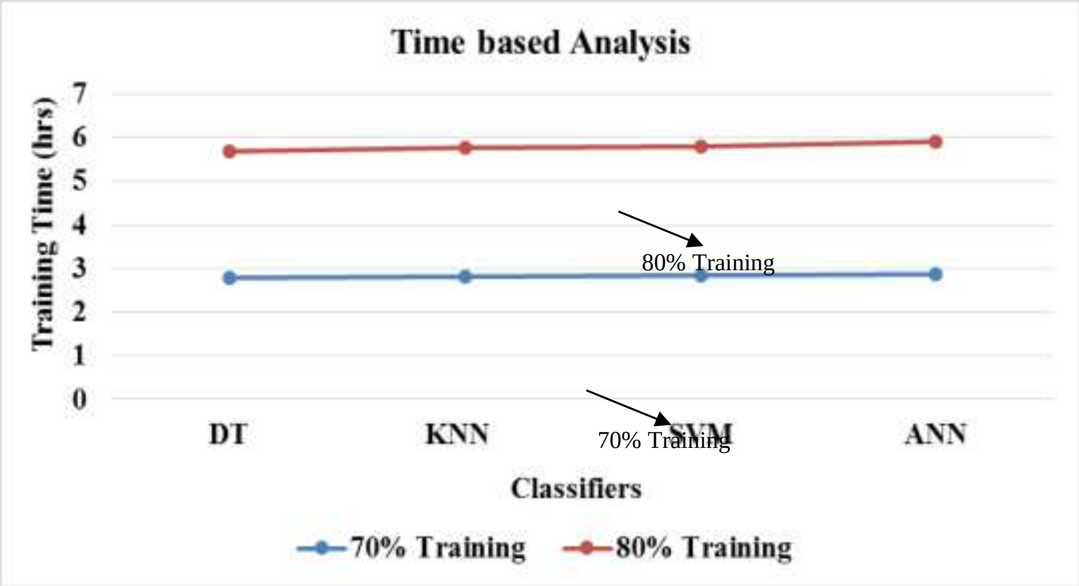

Fig. 25. Overall Training Time Analysis.

The graphical representation provides visual comparison of above classifiers overtime, highlighting the consistent the trend of advancement in performance matrix particular see in the ANN models. However, the most recent studies in 2022 showcased even more promising results. He et al. introduced an ANN, achieving substantial improvements with Precision, Recall, and F-measure scores of 0.9298, 0.9321, and 0.9309486, respectively. Remarkably, the proposed ANN model demonstrated further enhancement, reaching impressive scores of 0.9437843, 0.9485698, and 0.9461641 for Precision, Recall, and F-measure. The graphical representation in Figure 24 provides a visual comparison of these classifiers over time, highlighting the consistent trend of advancement in performance metrics, particularly seen in the ANN-based models. This comparison underscores the evolution and progression of classifier models, emphasizing the ongoing efforts to improve predictive accuracy and model robustness within this domain.

TABLE IV. TESTING COMPARATIVE ANALYSIS FOR VARIOUS CLASSIFIERS.

Testing	DT (milisec)	KNN (milisec)	SVM (milisec)	ANN (milisec)
30%	38.57819	39.165571	39.236453	39.908907
20%	38.006227	39.24751	39.8263	40.445804

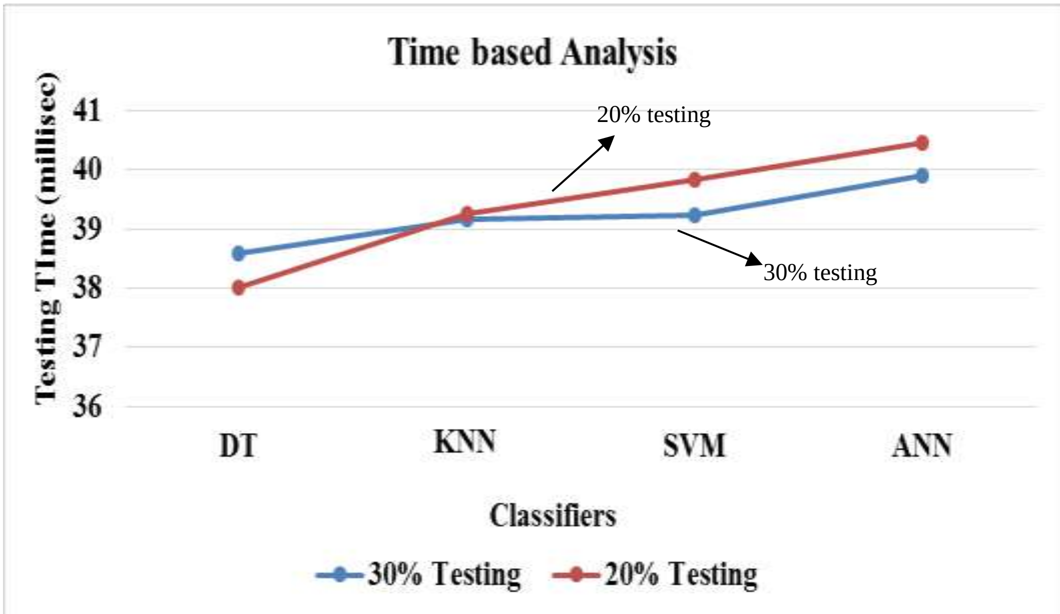

Fig. 26. Overall Testing Time Analysis.

TABLE V. COMPARATIVE ANALYSIS OF VARIOUS CLASSIFIERS ACROSS DIFFERENT YEARS.

Referred Work	Classifiers	Precision	Recall	F-measure
Arulmurugan and Anandakumar (2018) [33]	FFBPNN	0.9124	0.8934	0.9028
Sweetlin et al. (2019) [98]	SVM	0.9142	0.9012	0.9076535
Jena et al. (2021) [106]	Region-based Convolutional Neural Network (RBCNN)	0.9045	0.9128	0.908631
He et al. (2022) [37]	ANN	0.9298	0.9321	0.9309486
Proposed	ANN	0.9437843	0.94856 98	0.9461641

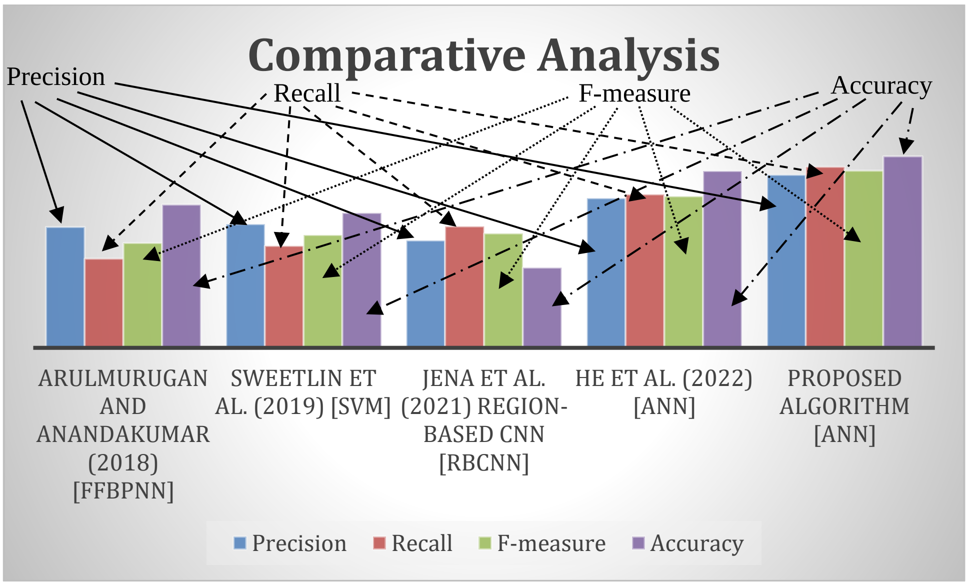

Fig. 27. Comparative Analysis of Parameters.

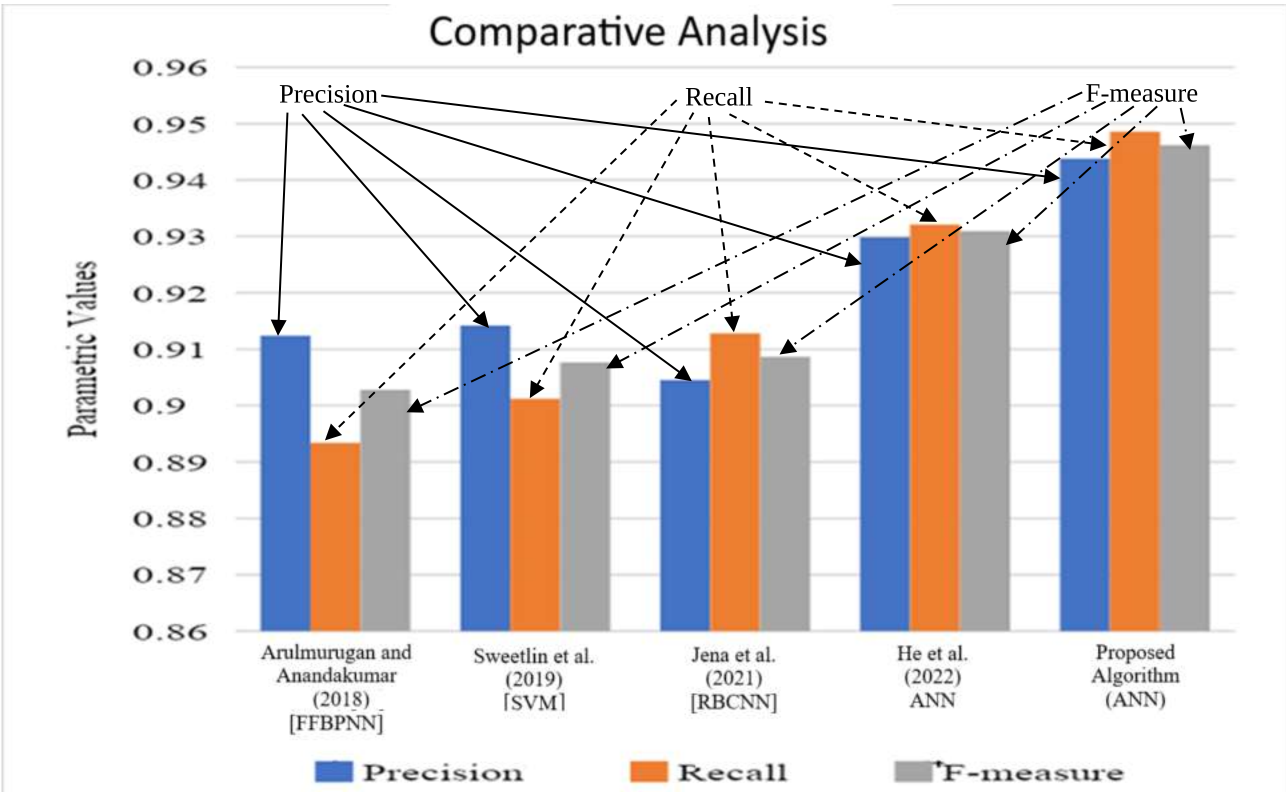

Fig. 29. Comparative Analysis of Accuracy.

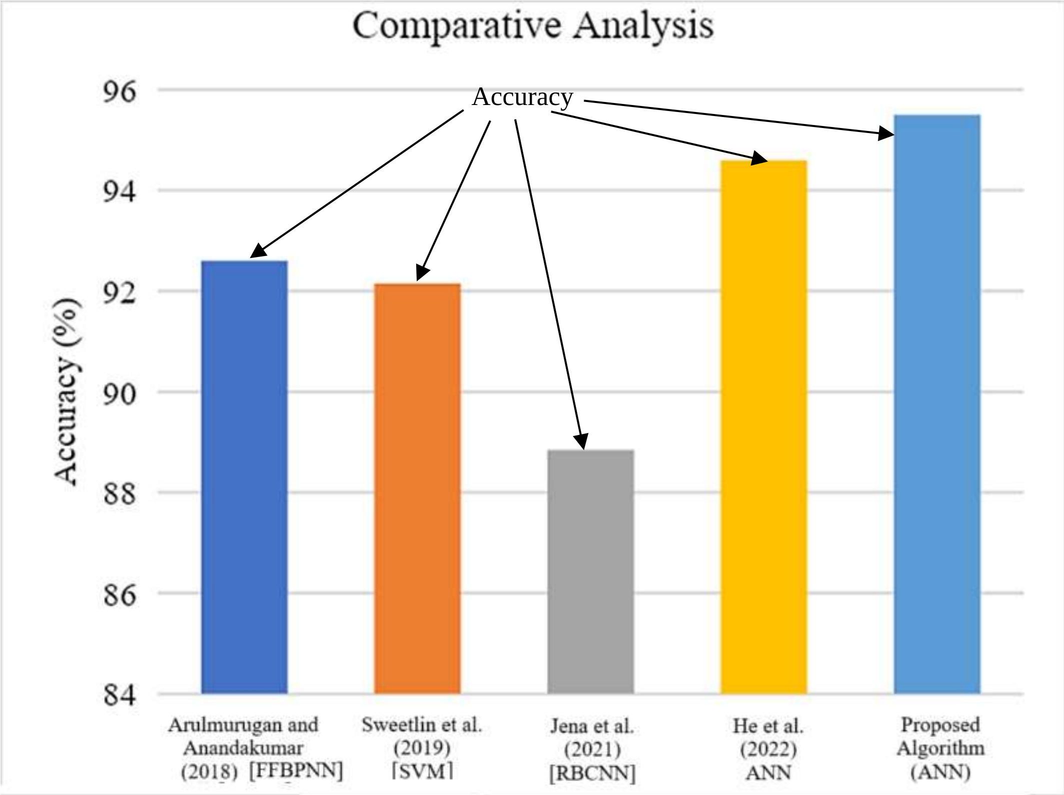

Fig. 30. Comparative Analysis of Parametric Values.

Figure 29. Comparative analysis of accuracy, presents a comprehensive overview of accuracy across different classifiers reported in various studies spanning from 2018 to 2022. The depicted accuracy values showcase the performance of each classifier in different research works. In 2018, Arulmurugan and Anandakumar introduced FFBPNN with an accuracy of 92.6%. Sweetlin et al. in 2019 reported an accuracy of 92.15% using SVM. Jena et al. in 2021 introduced Region-based Convolutional Neural Network (RBCNN) with an accuracy of 88.84%. Subsequently, He et al. in 2022 proposed an ANN classifier achieving an accuracy of 94.6%. Notably, the Proposed ANN in outperformed previous classifiers, reaching an accuracy of 95.496861, indicating significant advancements in classification accuracy compared to earlier methods. This comparison underscores the progression and enhancements achieved in accuracy by different classifiers over the referenced years.

V. Conclusion

Utilizing the Lung Image Database's meticulously annotated CT scan images, this study employed advanced image analysis techniques. The Scale Invariant Feature Transform (SIFT) expedited feature extraction, significantly reducing processing time in subsequent classification phases. This comprehensive assessment spanned diverse dataset distributions (70:30 and 80:20), shedding light on the efficacy of various machine learning models for lung cancer classification. Among the models examined, the ANN consistently stood out, demonstrating unparalleled performance in accurately identifying and delineating lung cancer instances. Across both dataset distributions, the ANN showcased exceptional metrics: recall (0.9240 / 0.9486), precision (0.9240 / 0.9486), F-measure, and accuracy (93.15% / 95.50%), surpassing Decision Trees, K-Nearest Neighbors, and SVM. Moreover, this study emphasizes the critical alignment of model selection with dataset configurations to optimize accuracy, providing a strategic pathway for refining lung cancer diagnostic systems and potentially enhancing patient outcomes. Additionally, it reaffirms the significance of integrating k-means with CSA for highly precise lung cancer segmentation, further bolstering the study's credibility and offering avenues for future research and implementation. In essence, this investigation substantiates the ANN's efficacy in lung cancer classification, underlining the pivotal role of advanced segmentation techniques. The study's findings offer crucial insights for advancing diagnostic approaches and treatment protocols in combating this formidable health challenge. The execution time analysis and comparative analysis also justified the effectiveness of the proposed lung cancer detection and classification. Despite of the fact that utilization of ANN in the methodology resulted in slight increase in the training and testing time. However, it can be overlooked due to overall improved performance owing to the involvement of ANN.

References

[1] Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. Radiology, 278(2), 563-577. doi: 10.1148/radiol.2015151169.

[2] Sethi, M., & Goo, J. M. (2019). Artificial intelligence in chest radiography: A primer. Clinical Radiology, 74(5), 357-366. doi: 10.1016/j.crad.2018.12.002.

[3] Zhao, B., Tan, Y., Tsai, W. Y., Qi, J., Xie, C., Lu, L., & Schwartz, L. H. (2016). Reproducibility of radiomics for deciphering tumor phenotype with imaging. Scientific Reports, 6, 23428. doi: 10.1038/srep23428.

[4] National Cancer Institute. (2021, March 16). Lung cancer screening. Retrieved from https://www.cancer.gov/types/lung/screening-fact-sheet

[5] Lee, H., Park, H., Lee, S. H., & Kim, H. (2019). Computer-Aided Diagnosis of Lung Cancer: An Overview and Recent Advances. Journal of Clinical Medicine, 8(8), 1160. https://doi.org/10.3390/jcm8081160.

[6] Li, Y., Hu, Y., Liu, X., Gao, Y., Wu, F., & Shi, G. (2021). Detection of lung cancer from CT images using deep learning and radiomics: A systematic review and meta-analysis. Medical Physics, 48(4), e240–e260. https://doi.org/10.1002/mp.14846

[7] Mansoor, A., Bagci, U., Foster, B., Xu, Z., Papadakis, G. Z., Folio, L. R., … Udupa, J. K. (2018). Segmentation and Image Analysis of Abnormal Lungs at CT: Current Approaches, Challenges, and Future Trends. RadioGraphics, 38(7), 1958–1975. https://doi.org/10.1148/rg.2018180031.

[8] Sharma, A., & Madabhushi, A. (2019). Artificial Intelligence in Lung Cancer Pathology Image Analysis. Current Pathobiology Reports, 7(3), 65–70. https://doi.org/10.1007/s40139-019-00202-1

[9] Arulmurugan R, Anandakumar H. Early detection of lung cancer using wavelet feature descriptor and feed forward back propagation neural networks classifier. Lecture Notes in Computational Vision and Biomechanics. 2018; 28:103–10.

[10] Senthil Kumar K, Venkatalakshmi K, Karthikeyan K, others. Lung cancer detection using image segmentation by means of various evolutionary algorithms. Computational and mathematical methods in medicine. 2019; 2019.

[11] Sweetlin JD, Nehemiah HK, Kannan A. Computer aided diagnosis of drug sensitive pulmonary tuberculosis with cavities, consolidations and nodular manifestations on lung CT images. International Journal of Bio-Inspired Computation. 2019; 13(2):71–85.

[12] Lenin Fred A, Kumar SN, Padmanaban P, Gulyas B, Ajay Kumar H. Fuzzy-crow search optimization for medical image segmentation. Applications of hybrid metaheuristic algorithms for image processing. 2020; 413–39.

[13] Jena SR, George ST, Ponraj DN. Lung cancer detection and classification with DGMM-RBCNN technique. Neural Computing and Applications. 2021; 33:15601–17.

[14] He B, Hu W, Zhang K, Yuan S, Han X, Su C, et al. Image segmentation algorithm of lung cancer based on neural network model. Expert Systems. 2022; 39(3):e12822.

[15] Guleria K, Sharma S, Kumar S, Tiwari S. Early prediction of hypothyroidism and multiclass classification using predictive machine learning and deep learning. Measurement: Sensors. 2022; 24:100482.

[16] Karthick P, Mohiuddine SA, Tamilvanan K, Narayana moorthy S, Maheswari S. Investigations of color image segmentation based on connectivity measure, shape priority and normalized fuzzy graph cut. Applied Soft Computing. 2023; 139:110239.

[17] Jian M, Wu R, Chen H, Fu L, Yang C. Dual-Branch-UNet: A Dual-Branch Convolutional Neural Network for Medical Image Segmentation. CMES-Computer Modeling in Engineering \& Sciences. 2023; 137(1).

[18] Shanmuganathan S. Artificial neural network modelling: An introduction. Studies in Computational Intelligence 2016; 628: 1–14.

[19] Tkáč M, Verner R. Artificial neural networks in business: Two decades of research. Applied Soft Computing 2016; 38: 788–804.

[20] Zhou D, Thompson WK, Siegle G. MATLAB toolbox for functional connectivity. Neuroimage 2009; 47: 1590–1607