Efficient Classification of Diffuse Lung Disease in Class Imbalance Data

Shyla Raj and B.S Mahanand

Abstract—**Diffuse lung disease is a group of complex lung disorders that causes lung scarring. Emphysema, ground glass opacity, fibrosis, and micronodules are the lung disorders that are associated with this disease. These disorders appear as texture alterations in computed tomography and exhibit lower intraclass and higher inter class variations. These variations can be extracted using texture features, using modified intuitionistic fuzzy local binary pattern, in combination with gray level co-occurrence matrix, and Gabor filter bank. TALISMAN is one of the standard data sets used for diffuse lung classification problems. Analysis of the TALISMAN dataset reveals an imbalance in the sample distribution across different classes of diffuse lung diseases. Addressing sample imbalance is crucial, as it significantly impacts the performance of the classifiers. In this study, the issue of sample imbalance in diffuse lung disease classification is addressed by analyzing the effectiveness of the synthetic minority oversampling technique and weighted extreme learning machine classifier. The performance of both approaches is assessed and compared, with the results indicating that the weighted extreme learning machine performs better for diffuse lung disease classification, particularly in the presence of sample imbalance.**

Index Terms—**diffuse lung diseases, sample imbalance, synthetic minority oversampling technique, TALISMAN, weighted extreme learning machine.**

I. Introduction

DIFFUSE lung diseases (DLDs) are a diverse collection of lung diseases that affect the tissues surrounding the alveoli of the lungs [1]. High-resolution computed tomography (HRCT) of the lungs can aid better diagnosis of diffuse parenchymal diseases [2]. However, interpreting HRCT scans is a demanding task due to the large volume of data, significant similarities between diffuse lung patterns, and the subjective variability among radiologists. DLDs are progressive and irreversible, making early detection crucial in preventing disease progression. Detecting DLDs from HRCT scans in their early stages requires identifying subtle changes in lung patterns, which is challenging. Therefore, computer-aided methods have been developed to assist radiologists in detecting DLDs. DLDs manifest as texture alterations in the lung parenchyma. As a result, texture features are valuable for accurately analyzing diffuse lung patterns. The TALISMAN dataset [3] is one of the most extensively studied and publicly accessible datasets for DLD classification. This dataset has an imbalance in sample distribution across various disease types, which must be addressed, as it significantly impacts the performance of traditional classification techniques [4]. Machine learning algorithms generally aim for higher accuracy, which often leads to bias towards classes with more data samples. In such cases, these algorithms may classify all cases as belonging to the majority class, resulting in high overall accuracy but precision of identifying minority class samples will be low. Misclassifying a minority class sample (positive) as a majority class sample (negative) can have significant consequences in applications such as fraud detection and medical classification. A false positive prediction can cause unnecessary anxiety, while a false negative prediction may delay medical intervention. Therefore, addressing the imbalance in the dataset is critical.

Manuscript received November 6, 2024; revised May 29, 2025.

Shyla Raj is an associate professor in the Department of Information Science and Engineering, JSS Science and Technology University, Mysuru 570 006, Karnataka, India (email: shylaraj@jssstuniv.in).

B.S.Mahanand is a professor in the Department of Information Science and Engineering, JSS Science and Technology University, Mysuru 570 006, Karnataka, India (Corresponding author to provide email:bsmahanand@sjce.ac.in).

Most studies on DLD classification use traditional classifiers such as Bayesian classifiers [5], random forest classifiers [6], k-nearest neighbors (k-NN) [7], support vector machines (SVM) [8]-[9], and neural networks [10], which do not address the issue of sample imbalance. In contrast, convolutional neural network (CNN) models employ augmentation techniques to increase training samples and mitigate overfitting caused by limited training data. Augmentation techniques applied in CNN models include rotation and flipping, and combination of both to add samples to the minority class [11]. Label-preserving geometric transformations are applied to avoid overfitting in [12], where it was reported that the augmentation technique improved classification performance by around 5%. Wang et al.[13] addressed imbalance by increasing overlapping size between adjacent patches in the minority class and by decreasing overlapping size in the majority class. Random sampling at various rates to the TALISMAN dataset to balance the data samples in [14]. Rotation-based augmentation is employed in [15], where images are augmented through several rotations. Rotation, horizontal and vertical flipping, shifting, and shearing techniques are used to increase minority samples in the training set in [16]. Lung tissue patches are rotated by multiples of 90^0 to generate more training samples in [17]. From these studies, it is evident that while some works on DLD classification employed traditional classifiers and did not address the imbalance in the dataset, CNN models handle the imbalance using augmentation techniques.

Many practical problems, including medical classification, often face challenges related to small sample sizes and imbalanced sample distributions. Approaches for handling sample imbalance include both sampling and algorithmic techniques. Sampling is a common method that is used to balance a dataset by altering its distribution. Sampling could be oversampling or undersampling. Undersampling technique reduces the samples in the majority class to match the minority class, whereas oversampling increases the samples in the minority class. Hasanin et al.[18] investigated the effects of random undersampling for addressing imbalance in bioinformatics datasets. Their study reported that, compared to random oversampling, random undersampling had a smaller computational overhead and resulted in faster training times, which is advantageous in data analytics. However, undersampling may risk excluding important data from the majority class. Whereas, oversampling often duplicates minority class samples, which can lead to overfitting [19]. Chawla et al.[20] proposed the synthetic minority oversampling technique (SMOTE) to overcome the disadvantages of random sampling and to address class distribution imbalance. SMOTE creates synthetic samples in the feature space rather than simply duplicating data in the feature space, and is widely used in many practical problems [21]. Another approach for handling sample imbalance is through algorithmic techniques, which aim to improve the performance of minority class classification without altering the sample distribution. Cost-sensitive learning is a popular approach in this category, where higher misclassification costs are assigned to the minority class to prevent it from being overlooked. Weighted extreme learning machine (WELM) extends the extreme learning machine (ELM) by incorporating a weight matrix to handle data imbalance [22]. WELM has been employed in many practical applications, including indoor positioning systems, user behavior prediction in social networks, fault diagnosis, network intrusion detection, software defect prediction, and tracking systems [23]-[24].

The primary goal of this work is to classify diffuse lung disease while addressing the class imbalance problem. Both sampling and classifier approaches are explored to handle the imbalance in the DLD dataset. In the sampling approach, the minority class samples are oversampled using SMOTE, which creates synthetic samples in the feature space. In the classifier approach, WELM is employed as a cost-sensitive classifier to handle the sample imbalance by assigning misclassification costs based on class distribution. The rest of the paper is structured as follows: Section II presents the dataset used in this study and the methodology employed. Section III discusses the evaluation metrics, parameter configurations, and the results achieved. Section IV presents the result analysis. Finally, Section V concludes the work.

II. Materials and Method

In this work, DLD is classified by extracting texture features and addressing sample imbalance through both sampling and cost-sensitive classifier approaches.

A. Dataset

For this work, the publicly available standard DLD dataset, TALISMAN [3], is used. The dataset was created at the University Hospitals in Geneva. It consists of 103 digital imaging and communications in medicine (DICOM) image series from 128 patients, each diagnosed with one of the 13 histologically identified DLDs. The database contains 3D annotated regions of interest (AROI), delineated by two experienced radiologists. The AROI is used to extract 14,356 overlapping square patches of size $32X32$, providing a precise representation of the DLD patterns. Pathological confirmation (through biopsy or bronchoalveolar lavage) or laboratory/specific tests were used to confirm the diagnosis for each case. Table I shows the class-wise distribution of images, AROI, and patches of DLD patterns. As seen in Table I, there is an imbalance in the sample distribution among DLD patterns. This imbalance is primarily due to the way radiologists mark the regions during annotation sessions. The size of the ROIs varies significantly depending on the type of lung tissue. Some patterns are dispersed throughout the lungs, allowing for the delineation of large areas, while others are more localized, resulting in smaller ROIs. Emphysema is a minority class with 1,177 samples, whereas healthy lung tissue is the dominant class with 5,530 samples. Additionally, the sample distribution among the other three classes differs from that of the healthy class.

TABLE I
DISTRIBUTION OF DIFFUSE LUNG DISEASE CLASS-WISE

DLD Type	No.Images	No.AROI	No.Samples
Emphysema (E)	5	66	1177
Ground Glass Opacity (GG)	37	427	2226
Micro-nodules (MN)	16	297	2384
Fibrosis (F)	38	473	3039
Healthy (H)	7	100	5530
Total	103	1363	14,356

B. Texture Feature Extraction

DLD is characterized by inter and intraclass diversity within the classes and texture analysis will help identify the patterns efficiently. As a result, this study extracts lung texture features using the modified intuitionistic fuzzy local binary pattern (MILBP), combined with gray level co-occurrence matrix (GLCM) and Gabor filter bank features [25]. To analyze the lung texture in local neighborhoods, MILBP features are extracted. MILBP extends the local binary pattern (LBP) by incorporating intuitionistic principles and associating membership, non-membership, and hesitation degrees with each pixel in the image. The Gabor filter bank performs multichannel representation and provides optimal localization of images in both the spatial and frequency domains. To obtain a multi-resolution representation of texture content in the DLD images, a rotation and scale-invariant Gabor filter bank is employed [26]. GLCM is employed to analyze the spatial distribution of pixels. GLCM is a global texture feature extraction technique that quantifies the frequency of occurrence of gray levels and spatial interdependencies between pixels [27]. After feature extraction, the sample imbalance in DLD is addressed using both sampling and classifier approaches.

C. Synthetic Minority Oversampling Technique

Sampling is one of the most widely used technique for addressing data imbalance. Random oversampling and random undersampling are two types of sampling. Without adding new information, random oversampling generates duplicate data in the minority class. To balance the sample distribution, random undersampling removes samples from the majority class at random. Oversampling may cause the decision boundary for the minority class to become overly specific, potentially leading to overfitting [28]. On the other hand, undersampling carries the risk of discarding vital data. To overcome the shortcomings of the aforementioned methodologies, this work employs SMOTE to address the imbalance in sample distribution [29]. SMOTE is an oversampling technique that generates synthetic samples for minority classes rather than merely duplicating existing ones. By interpolating in feature space instead of data space, the method creates synthetic samples from existing minority samples.

D. Weighted Extreme Learning Machine

The second approach for addressing the imbalance in sample distribution is by modifying the classifier. Traditional classifiers assign equal misclassification costs to all classes, and in the case of sample imbalance, they tend to favor the majority class in order to achieve higher accuracy. Although the overall accuracy may be high, the recognition of minority class samples is poor. The algorithmic or classifier-based approach addresses class imbalance by enhancing the learning of classification algorithms concerning the minority class by assigning higher misclassification costs, thereby preventing its neglect. In this work imbalance in sample distribution is addressed by employing the cost-sensitive classifier WELM [22].

WELM is a single-layer feed-forward NN that generates hidden nodes randomly and is unaffected by the training data or the output of the hidden layer. WELM automatically constructs a misclassification cost matrix to address the class imbalance. The misclassification cost is inverse to the number of samples in each class. As a result, minority class samples incur a higher misclassification cost, while majority class samples incur a lower cost. This way, the minority class gains strength, and the relative influence of the majority class diminishes.

WELM tries to reduce 'ζ_i' training error while improving the marginal distance ||B||/2 between classes. The following is a mathematical representation of this optimization:

$$Minimize : L_{WELM} = \frac{\|B\|^2}{2} + \frac{1}{2}CW\sum_{i=1}^{M}\|\zeta_i\|^2 \quad (1)$$

$$Subject\ to : h(y_i)B = {s_i}^T - {\zeta_i}^T, i = i, \ldots, M \quad (2)$$

The regularization parameter 'C' represents the trade-off between minimizing training errors and maximizing the margin distance. 'M' denotes the sample size, and 'W' is the cost matrix of size MXM. The matrix 'W' is diagonal, with the diagonal elements indicating the misclassification cost for each sample. The 'W' determines the extent of re-balancing and how far the decision boundary can be pushed towards the majority class. The weight matrix 'W' can be computed using two different weighting systems, $W1$ and $W2$, depending on the degree of re-balancing required.

$$Weighting\ Scheme \quad W1\ : W_i = \frac{1}{No.(s_i)} \quad i = 1...M \quad (3)$$

$$\text{Weighting Scheme} \quad \text{W2=} \begin{cases} W_i = 1/No.(s_i) & if \quad s_i \leq AVG(s_i) \\ W_i = 0.618/No.(s_i) & if \quad s_i > AVG(s_i) \end{cases} \quad (4)$$

where $No.(s_i)$ is the number of data samples. W1 re-balances the minority class samples with majority class samples in the ratio 1:1. Whereas the W2 re-balances in the ratio 0.618:1. The training model of WELM is described in Algorithm.1. During training, the output weight 'B' is calculated, and in the testing, the class labels of the unknown samples are determined using 'B'.

Algorithm 1: Weighted extreme learning machine

Data: Set of training samples D={(x_i, s_i) i=1...M } f(x)-activation function and 'N'-number of hidden nodes.

Result: Trained WELM model

Step 1: Determine the weight/cost matrix 'W' associated with each training sample x_i

Step 2: Generate randomly the input weight vector $'w_i'$ and hidden node bias $'a_i'$, i=1...M.

Step 3: Hidden layer output matrix 'L' is calculated as follows

L=($w_1 \ldots w_L$,$a_1 \ldots a_L$,$x_1 \ldots x_L$)

$$\begin{bmatrix} f(w_1, a_1, x_1) \ldots f(w_N, a_N, x_1) \\ \vdots \qquad\qquad \vdots \\ f(w_1, a_1, x_M) \ldots f(w_N, a_N, x_M) \end{bmatrix}_{MXN}$$

Step 4: Output weight B is calculated as

$$B = L^T(\frac{1}{c} + WLL^T)^{-1}WT$$

where T= $[t_1 \ldots t_M]$

III. Results

This section presents the evaluation metrics, parameter configuration, and experimental results. The classification results for DLD, using both the sampling and cost-sensitive classifier approaches, are compared and contrasted.

A. Evaluation Metrics and Parameter Configuration

In the vast majority of classification problems, the accuracy metric is used to evaluate performance. However, accuracy may not be the optimal statistical measure in the case of sample imbalance, as it is highly influenced by the class with more samples [29]. To evaluate an imbalanced problem, alternative metrics such as precision, recall, F-measure, and geometric mean (G-measure) are often used. Another prominent metric for assessing classifier performance is the receiver operating characteristic (ROC) curve and the area under the curve (AUC).

The performance of WELM is determined by the parameters of the Gaussian activation function, specifically the kernel width 'σ' and the trade-off constant 'C'. A grid search strategy is employed to determine the optimal values for 'σ' and 'C' when training the WELM classifier. From grid search it is found that 'σ' value of 2^{14} and a 'C' value of 2^{28} yield the best classification performance. Similarly, the performance of the two weighting schemes, W1 and W2, as well as the Gaussian and sigmoid activation functions, is examined. The experiments reveal that the W1 weighting scheme with the Gaussian or radial basis function (RBF) activation function provides superior performance, and thus, it is chosen.

TABLE II
PERFORMANCE COMPARISON OF DT, K-NN AND SVM CLASSIFIERS IN TERMS OF G-MEASURE IN IMBALANCED SCENARIO

Method	k-NN	SVM	DT
E	0.60	0.79	0.61
GG	0.85	0.89	0.86
MN	0.86	0.90	0.85
F	0.90	0.93	0.88
H	0.89	0.93	0.88
Mean	0.82	0.88	0.81

B. DLD classification using support vector machine

This section discusses the discriminating ability of the extracted texture features, namely MILBP, the Gabor filter bank, and GLCM, in classifying DLD. The experiments are carried by dividing the dataset into 70% training set and 30% testing set. The experiments are conducted 10 times with stratified random combinations for better generalization.

Table II presents the results of SVM, k-NN, and DT in the class-imbalance scenario. It can be observed from Table II that for all classifiers, the G-measure of ground glass, micro nodules, fibrosis, and the healthy class are higher than that of the minority class sample, emphysema. When comparing the performance of individual classifiers, SVM achieved a G-measure of 0.88, k-NN 0.82, and DT 0.81. From Table II, it can be inferred that SVM recognizes all DLD types better than k-NN and DT in an imbalance scenario.

TABLE III
CLASSIFICATION PERFORMANCE OF SVM IN CLASS IMBALANCE SCENARIO

DLD Type	Recall	Precision	F-measure	G-measure
E	0.62	0.91	0.74	0.79
GG	0.80	0.89	0.84	0.89
MN	0.84	0.82	0.83	0.90
F	0.88	0.88	0.88	0.93
H	0.95	0.86	0.91	0.93
Mean	0.82	0.88	0.84	0.88

Table III presents the classification performance of SVM in the class-imbalance scenario in terms of precision, recall, F-measure, and G-measure. Overall, precision is 0.88, recall is 0.82, F-measure is 0.82, and G-measure is 0.84. It is evident that the recognition of majority class samples is considerably higher than that of minority class samples in the class-imbalance situation.

C. DLD classification on employing Synthetic Minority Oversampling Technique

SMOTE is employed to create synthetic samples in the minority class to re-balance the dataset. Table IV presents the sample distribution in the original set, and Table V presents the sample distribution after re-sampling using SMOTE. It can be observed from Table IV that there is sample imbalance between various diffuse lung patterns in both the training set and the testing set. To re-balance the dataset, SMOTE is employed to increase the samples in the minority classes (emphysema, fibrosis, ground glass, and micro nodules) to match that of the majority class (healthy). Sampling is performed solely on the training set, while the testing set remains intact.

TABLE IV
THE DLD CLASS WISE DISTRIBUTION BEFORE SAMPLING

DLD Type	E	GG	MN	F	H
Training Set	824	1558	1669	2127	3871
Testing Set	353	668	715	912	1659

TABLE V
THE DLD CLASS WISE DISTRIBUTION AFTER SAMPLING THROUGH SMOTE

DLD Type	E	GG	MN	F	H
Training Set	3861	3873	3811	3857	3871
Testing Set	353	668	715	912	1659

Table VI presents the results obtained from SVM, k-NN, and DT in a balanced class scenario. By comparing the results of Table II and Table VI, it can be observed that the overall performance of all classifiers has improved as a result of the sampling technique. More importantly, the G-measure of the minority class, emphysema has significantly improved in all SVM, k-NN, and DT classifiers after increasing the samples. Along the same lines, it can also be observed that the G-measure of the majority class, healthy has been slightly reduced. The increase in minority class samples is responsible for this improvement, as increasing the samples leads to better recognition. In comparison, with a G-measure of 0.85 for k-NN and 0.83 for DT, the SVM classifier has achieved a higher overall G-measure of 0.90.

Table VII presents the classification performance of SVM for the balanced training set in terms of precision, recall, F-measure, and G-measure. Overall, precision is 0.85, recall is 0.84, F-measure is 0.84, and G-measure is 0.90 for SVM.

D. DLD classification using weighted extreme learning machine

Although the sampling approach by SMOTE increased the recognition of minority class samples, creating synthetic samples may not be appropriate for medical classification problems. Instead, the classifier can be modified to address the issue of class imbalance. To address the sample imbalance in the TALISMAN dataset, WELM, a cost-sensitive classifier, is employed. The effects of sample imbalance are handled by WELM by automatically creating the misclassification cost matrix based on the training set. The results obtained using WELM are shown in Table VIII.

TABLE VI
PERFORMANCE COMPARISON OF DT, K-NN AND SVM CLASSIFIERS IN TERMS OF G-MEASURE MEASURE FOR BALANCED DATASET

DLD Type	KNN	SVM	DT
E	0.72	0.85	0.74
GG	0.87	0.90	0.88
MN	0.87	0.91	0.82
F	0.88	0.92	0.88
H	0.88	0.92	0.82
Mean	0.85	0.90	0.83

TABLE VII
CLASSIFICATION PERFORMANCE OF SVM CLASSIFIER IN CLASS BALANCE SCENARIO

DLD	Recall	Precision	F-measure	G-measure
E	0.74	0.76	0.75	0.85
GG	0.83	0.86	0.84	0.90
MN	0.86	0.81	0.83	0.91
F	0.88	0.90	0.89	0.92
H	0.90	0.90	0.90	0.92
Mean	0.84	0.85	0.84	0.90

TABLE VIII
PERFORMANCE OF COST SENSITIVE WELM CLASSIFIER

DLD	Recall	Precision	F-measure	G-measure
E	0.75	0.81	0.78	0.86
GG	0.89	0.84	0.86	0.93
MN	0.85	0.88	0.87	0.91
F	0.84	0.95	0.89	0.91
H	0.94	0.88	0.91	0.93
Mean	0.85	0.87	0.86	0.91

From Table VIII, it can be seen that WELM achieves an overall recall/sensitivity of 0.85, precision of 0.87, 0.86 F-measure, and 0.91 G-measure. Furthermore, when the G-measure of individual classes is analyzed, it can be observed that WELM achieves a G-measure of 0.86 for emphysema, 0.93 for ground glass, 0.91 for micro nodules, and 0.91 for fibrosis. For the healthy class, the G-measure is 0.93. This suggests an improvement in the identification of the minority class, emphysema.

IV. DISCUSSION

Generally, to achieve higher overall accuracy, machine learning classifiers tend to favor the class with more samples. While this approach results in higher overall accuracy, the recognition of classes with fewer samples tends to be lower. This is evident in the case of DLD classification, as presented in Table II and Table III. Fig. 1 shows the ROC curve for SVM in the class imbalance scenario. From Fig. 1, it can be observed that the AUC is 0.81 for emphysema, 0.89 for ground glass, micro nodules is 0.90, 0.93 for fibrosis, and for the healthy class it is 0.93.

From Table II, Table III, and Fig. 1, it can be observed that, in the case of class imbalance, traditional classifiers tend to prioritize the class with more samples to achieve higher overall performance. As a result, the class with fewer samples is often overlooked. To overcome this bias, addressing the sample imbalance is crucial.

In this work, the imbalance is addressed using SMOTE, a sampling technique, and WELM, a cost-sensitive classifier approach. From Table VI, it can be inferred that employing SMOTE has improved the recognition of the minority class sample, emphysema. Since SVM yielded better results than k-NN and DT, the ROC curve for SVM with SMOTE is shown in Fig. 2. The AUC values for the classes are as follows: emphysema = 0.87, ground glass = 0.91, micro nodules = 0.89, fibrosis = 0.93, and healthy = 0.92, as shown in Fig. 2. Comparing Fig. 1 and Fig. 2, it is evident that

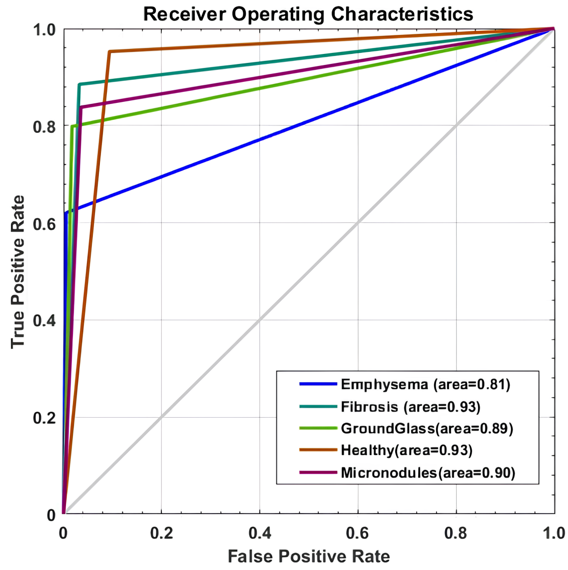

Fig. 1. Receiver operating characteristic curve and area under curve of SVM in class imbalance scenario

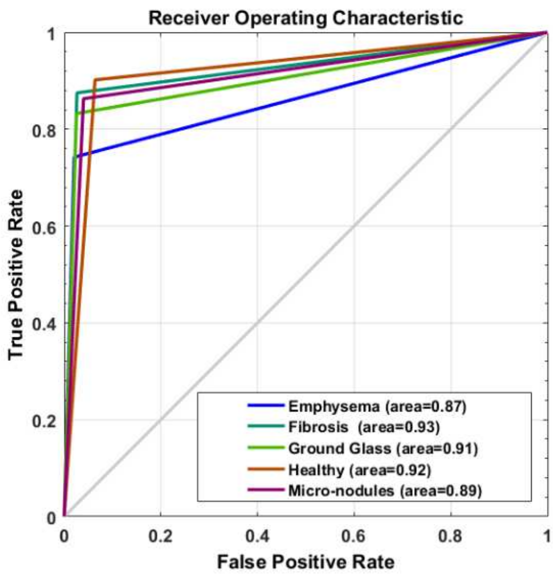

Fig. 2. Receiver operating characteristic curve and area under curve of SVM classifier using SMOTE

the recognition of the minority class sample, emphysema, has increased from 0.81 to 0.87, while the bias toward the majority class sample, healthy, has decreased from 0.93 to 0.91. This demonstrates that by balancing the dataset, the accuracy in classifying minority class samples has improved.

ROC and AUC for the WELM classifier is presented in Fig. 3. It can be seen from Fig. 3, that the AUC values are 0.87 for emphysema, 0.92 for ground glass, 0.93 for micro nodules, 0.92 for fibrosis, and 0.93 for the healthy class. Although the results from WELM and SVM with SMOTE are similar, creating synthetic samples may not be appropriate in medical image/data classification. In contrast, employing a cost-sensitive classifier like WELM is a more suitable approach for addressing class imbalance in such scenarios.

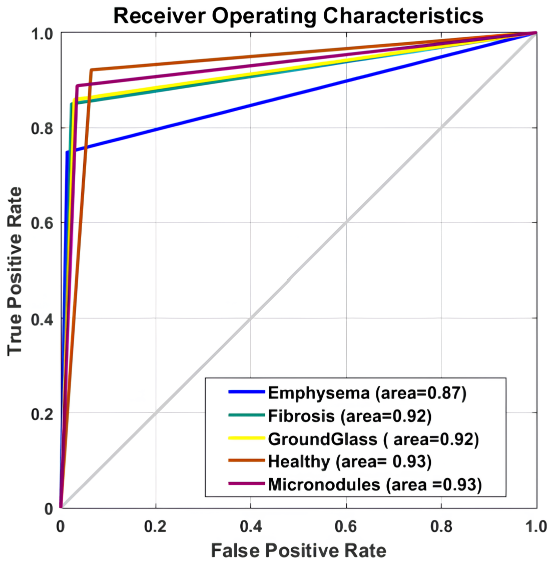

Fig. 3. Receiver operating characteristic curve and area under curve of WELM

Additionally an analysis is made to investigate the impact of sample imbalance in automatic feature extraction using pre-trained networks. Deep features in this work are extracted using the pre-trained VGG-19 (Visual Geometry Group) network through transfer learning approach, followed by classification using WELM classifier. WELM classifier with deep features achieves 0.76 recall, a precision of 0.74, F-measure of 0.75, and a G-measure of 0.85. The G-measure for the emphysema class is 0.73, for ground glass opacity it is 0.88, for micro nodules it is 0.85, for fibrosis it is 0.90, and for the healthy class, it is 0.89. This indicates that the recognition of minority class samples has decreased when only deep features are utilized. When deep features are concatenated with handcrafted features WELM classifier achieves 0.84 recall, precision of 0.82, a F-measure of 0.83, and 0.90 G-measure. The G-measure for the minority class samples is 0.84 for emphysema, 0.90 for ground glass opacity, 0.90 for micro nodules, 0.93 for fibrosis, and 0.92 for the healthy class. The comparsion results indicate that handcrafted features with WELM classifier outperform the classification results obtained using deep features and concatenation of handcrafted and deep features.

On analyzing the results it is clear that hand-crafted features outperform deep learning features in class imbalance scenario. This decrease in performance with deep features could be attributed to the fact that deep learning models are data-driven and typically require large datasets. However, obtaining a large annotated medical dataset is difficult due to confidentiality constraints, financial limitations, and the significant time required for experts to annotate the data [30]. On the other hand, hand-crafted features tend to perform well even with a limited number of samples. Another possible reason for the lower performance with deep features is that our dataset is small and unbalanced, and fine-tuning a pre-trained network on such a dataset could lead to overfitting. Additionally, no augmentation technique was applied in this work, which could have improved the deep learning model's generalization ability.

TABLE IX
PERFORMANCE COMPARISON OF SAMPLING AND CLASSIFIER APPROACH

DLD	WELM	SMOTE-SVM	SMOTE-ELM	SMOTE-RVFL
E	0.86	0.85	0.81	0.85
GG	0.93	0.90	0.87	0.93
MN	0.91	0.91	0.91	0.92
F	0.91	0.92	0.92	0.90
H	0.93	0.92	0.93	0.93
Mean	0.91	0.90	0.88	0.91

In terms of problem formulation and network architecture, there are connections between ELM and SVM [31]. ELM also shares similarities with random vector functional link (RVFL) networks in terms of implementation [32]. Both are feedforward NN with single-hidden-layer and randomly initialized weights. The key differences, however, lie in the connections between the input and output layers, as well as the training methods. RVFL features a direct connection between the input and output layers, unlike ELM, which lacks this connection. In terms of training, RVFL generates and fixes the hidden layer randomly before training, while WELM initializes the weights connecting the input layer to the hidden layer randomly and optimizes them during the training phase using a least squares regression algorithm [33]-[34].

Table IX compares the results of WELM and SMOTE with SVM, SMOTE with ELM, and SMOTE with RVFL in terms of the G-measure measure. As seen in Table IX, WELM achieves an overall G-measure of 0.91 while SVM with SMOTE achieves 0.90, ELM with SMOTE achieves 0.88, and RVFL with SMOTE achieves a G-measure of 0.91. Although the difference in the G-measure between the individual classifiers is minimal, when the recognition of individual classes is considered, WELM shows superior accuracy in recognizing the minority classes emphysema and ground glass. Despite both WELM and RVFL achieving the same G-measure, it is important to note that generating synthetic samples for medical classification problems may not be appropriate.

WELM classifier's performance is compared with cost sensitive classifiers namely robust energy-based least squares twin SVM (RELS-TSVM) and least squares SVM (LS-SVM) [35]-[36] in Table X. As shown in Table X, WELM achieves a G-measure of 0.86 for emphysema, while RELS-TSVM attains a G-measure of 0.84 and LS-SVM achieves 0.76. Overall, WELM reaches a G-measure of 0.91, compared to 0.89 for RELS-TSVM and 0.88 for LS-SVM, demonstrating an approximate 3% performance improvement.

TABLE X
WELM CLASSIFIER PERFORMANCE COMPARED TO COST SENSITIVE LS-SVM AND RELS-TSVM CLASSIFIERS

DLD	LS-SVM	RELS-TSVM	WELM
E	0.76	0.84	0.86
GG	0.89	0.92	0.93
MN	0.89	0.91	0.91
F	0.93	0.92	0.91
H	0.92	0.90	0.93
Mean	0.88	0.89	0.91

The results obtained in this work are compared with previous research on the TALISMAN dataset in Table XI, focusing on recall. The studies compared include Riesz and deep features with softmax [15], Riesz wavelet with the SVM classifier [9], rotation-invariant Gabor-LBP and multi-coordinate histogram of oriented gradient [37], isotropic wavelet features with the SVM classifier [8], texture, intensity and shape features [38], and deep CNN models [12][14]. From Table XI, it can be observed that, in class imbalance scenario, the recognition of the majority classes, healthy (H) and fibrosis (F), is comparable to existing works, but the recognition of emphysema (E) is lower. When the sampling technique is employed, SVM performs better in recognizing ground glass (GG) and emphysema, but the results still fall short compared to other studies in the literature. The combination of hand-crafted features with the WELM classifier achieves the highest recall for the healthy and ground glass classes when compared to other works. When only deep features are used with the WELM classifier, the recognition across all classes is lower than other methods. However, when deep and hand-crafted features are fused, the recognition of ground glass is the highest compared to other studies. By comparing the results of the proposed work with existing works in the literature, it is clear that better DLD classification performance can be achieved by using texture features such as MILBP, Gabor filter bank, and GLCM, along with addressing the sample imbalance using the WELM classifier.

TABLE XI
PERFORMANCE COMPARISON OF PROPOSED WORK WITH SIMILAR METHODS IN THE LITERATURE

Method	E	GG	MN	F	H
Imbalance	0.59	0.77	0.86	0.87	0.90
SVM+SMOTE	0.74	0.83	0.86	0.88	0.90
WELM Handcrafted	0.75	0.89	0.85	0.84	0.94
WELM Deep	0.57	0.81	0.82	0.83	0.79
WELM Handcrafted+Deep	0.64	0.91	0.82	0.87	0.83
Vishraj et al.[38]	0.71	0.79	0.87	0.80	0.81
Joyseeree et al.[15]	0.54	0.77	0.81	0.88	0.63
Joyseeree et al.[9]	0.57	0.73	0.88	0.82	0.73
Gao et al. [12]	0.83	0.82	0.88	0.89	0.91
Shin et al. [14]	0.91	0.70	0.79	0.83	0.68
Song et al.[37]	0.81	0.83	0.81	0.81	0.88
Depeursinge et al.[8]	0.78	0.81	0.81	0.81	0.59

V. CONCLUSION

This work presents a classification approach for diffuse lung patterns in a class-imbalanced scenario. Lung texture analysis is conducted using modified intuitionistic fuzzy local binary pattern, grey-level co-occurrence matrix, and Gabor filter bank. The issue of class imbalance is addressed by exploring both sampling and algorithmic approaches. Synthetic minority over-sampling technique is employed to create synthetic samples for the minority classes, instead of replicating the existing data. The support vector machine classifier with SMOTE achieved an average recall/sensitivity of 0.84, precision of 0.85, F-measure of 0.84, and G-measure of 0.90 for a balanced training set. Although the SMOTE addresses the sample imbalance, it may not be suitable for medical classification tasks due to the risk of generating synthetic data. On the other hand, the weighted extreme learning machine approach uses a cost matrix instead of synthetic sample generation. WELM assigns higher misclassification costs to classes with fewer samples, increasing their importance, while reducing the misclassification costs for classes with more samples, thus minimizing their dominance. WELM achieved an average recall/sensitivity of 0.85, 0.87 precision, F-measure of 0.86, and 0.91 G-measure for diffuse lung disease classification. When results of SVM with SMOTE and WELM are compared, it is evident that WELM performs better in classifying minority class samples. Further comparison of WELM with other cost-sensitive classifiers, such as RELS-TSVM and LS-SVM, shows an approximate 3% better performance of WELM. WELM's performance is also compared to previous studies, and the results indicate that DLD classification can be significantly improved by extracting robust texture features and effectively addressing sample imbalance.

REFERENCES

[1] K. M. Antoniou, G. A. Margaritopoulos, S. Tomassetti, F. Bonella, U. Costabel, and V. Poletti, "Interstitial lung disease," *European Respiratory Review*, vol. 23, no. 131, pp. 40–54, 2014.

[2] G. Raghu, M. Remy-Jardin, J. L. Myers, L. Richeldi, C. J. Ryerson, D. J. Lederer, J. Behr, V. Cottin, S. K. Danoff, F. Morell *et al.*, "Diagnosis of idiopathic pulmonary fibrosis. an official ats/ers/jrs/alat clinical practice guideline," *American Journal of Respiratory and Critical Care Medicine*, vol. 198, no. 5, pp. 44–68, 2018.

[3] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P.-A. Poletti, and H. Müller, "Building a reference multimedia database for interstitial lung diseases," *Computerized Medical Imaging and Graphics*, vol. 36, no. 3, pp. 227–238, 2012.

[4] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 50–59, 2004.

[5] J. K. Dash, S. Mukhopadhyay, M. K. Garg, N. Prabhakar, and N. Khandelwal, "Multi-classifier framework for lung tissue classification," in *IEEE Students Technology Symposium*. IEEE, 2014, pp. 264–269.

[6] M. Anthimopoulos, S. Christodoulidis, A. Christe, and S. Mougiakakou, "Classification of interstitial lung disease patterns using local dct features and random forest," in *Int. Conf. Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 6040–6043.

[7] P. D. Korfiatis, A. N. Karahaliou, A. D. Kazantzi, C. Kalogeropoulou, and L. I. Costaridou, "Texture-based identification and characterization of interstitial pneumonia patterns in lung multidetector ct," *IEEE Trans. Inf. Technol. Biomed*, vol. 14, no. 3, pp. 675–680, 2009.

[8] A. Depeursinge, D. Van de Ville, A. Platon, A. Geissbuhler, P.-A. Poletti, and H. Muller, "Near-affine-invariant texture learning for lung tissue analysis using isotropic wavelet frames," *IEEE Trans. Inf. Technol. Biomed*, vol. 16, no. 4, pp. 665–675, 2012.

[9] R. Joyseeree, H. Müller, and A. Depeursinge, "Rotation-covariant tissue analysis for interstitial lung diseases using learned steerable filters: Performance evaluation and relevance for diagnostic aid," *Computerized Medical Imaging and Graphics*, vol. 64, pp. 1–11, 2018.

[10] A. A. Dudhane and S. N. Talbar, "Multi-scale directional mask pattern for medical image classification and retrieval," in *Int. Conf. Computer Vision and Image Processing*. Springer, 2018, pp. 345–357.

[11] S. Huang, F. Lee, R. Miao, Q. Si, C. Lu, and Q. Chen, "A deep convolutional neural network architecture for interstitial lung disease pattern classification," *Medical Biological Engineering Computing*, pp. 1–13, 2020.

[12] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers *et al.*, "Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, vol. 6, no. 1, pp. 1–6, 2018.

[13] Q. Wang, Y. Zheng, G. Yang, W. Jin, X. Chen, and Y. Yin, "Multi-scale rotation-invariant convolutional neural networks for lung texture classification," *IEEE J. Biomed. Health Inform*, vol. 22, no. 1, pp. 184–195, 2018.

[14] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[15] R. Joyseeree, S. Otálora, H. Müller, and A. Depeursinge, "Fusing learned representations from riesz filters and deep cnn for lung tissue classification," *Medical Image Analysis*, vol. 56, pp. 172–183, 2019.

[16] D. Bermejo-Peláez, S. Y. Ash, G. R. Washko, R. S. J. Estépar, and M. J. Ledesma-Carbayo, "Classification of interstitial lung abnormality patterns with an ensemble of deep convolutional neural networks," *Scientific Reports*, vol. 10, no. 1, pp. 1–15, 2020.

[17] V. Sukanya Doddavarapu, G. B. Kande, and B. Prabhakara Rao, "Differential diagnosis of interstitial lung diseases using deep learning networks," *The Imaging Science Journal*, vol. 68, no. 3, pp. 170–178, 2020.

[18] T. Hasanin, T. M. Khoshgoftaar, J. Leevy, and N. Seliya, "Investigating random undersampling and feature selection on bioinformatics big data," in *Int. Conf. Big Data Computing Service and Applications*. IEEE, 2019, pp. 346–356.

[19] C. Drummond, R. C. Holte *et al.*, "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," *Workshop on Learning from Imbalanced Datasets*, vol. 11, no. 1, pp. 1–8, 2003.

[20] V. C. Nitesh, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[21] A. M. Idrees, N. S. Elhusseny, and S. Ouf, "Credit card fraud detection model-based machine learning algorithms." *IAENG International Journal of Computer Science*, vol. 51, no. 10, pp. 1649–1662, 2024.

[22] W. Zong, G.-B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, pp. 229–242, 2013.

[23] J. Zhang, H. Wang, and Y. Ren, "Robust tracking via weighted online extreme learning machine," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30 723–30 747, 2019.

[24] X. Luo, C. Jiang, W. Wang, Y. Xu, J.-H. Wang, and W. Zhao, "User behavior prediction in social networks using weighted extreme learning machine with distribution optimization," *Future Generation Computer Systems*, vol. 93, pp. 1023–1035, 2019.

[25] S. Raj, B. Mahanand, and D. Vinod, "Diffuse lung disease classification based on texture features and weighted extreme learning machine," *Multimedia Tools and Applications*, vol. 80, no. 28, pp. 35 467–35 479, 2021.

[26] J. K. Dash, S. Mukhopadhyay, and R. D. Gupta, "Multiple classifier system using classification confidence for texture classification," *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 2535–2556, 2017.

[27] R. M. Haralick, K. Shanmugam *et al.*, "Textural features for image classification," *IEEE Trans. Syst. Man. Cybern.*, vol. 3, no. 6, pp. 610–621, 1973.

[28] R. C. Holte, L. Acker, B. W. Porter *et al.*, "Concept learning and the problem of small disjuncts." in *Int. Joint Conf. Artificial Intelligence*, vol. 89, 1989, pp. 813–818.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[30] S. Khan and S.-P. Yong, "A comparison of deep learning and hand crafted features in medical image modality classification," in *Int. Conf. Computer and Information Sciences*. IEEE, 2016, pp. 633–638.

[31] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst. Man. Cybern.*, vol. 42, no. 2, pp. 513–529, 2011.

[32] Y.-H. Pao, "Functional-link net computing: theory, system architecture, and functionalities," *Computer*, vol. 25, no. 5, pp. 76–79, 1992.

[33] Y. Peng, Q. Li, W. Kong, F. Qin, J. Zhang, and A. Cichocki, "A joint optimization framework to semi-supervised rvfl and elm networks for efficient data classification," *Applied Soft Computing*, vol. 97, pp. 106–756, 2020.

[34] T. Goel, R. Sharma, M. Tanveer, P. Suganthan, K. Maji, and R. Pilli, "Multimodal neuroimaging based alzheimer's disease diagnosis using evolutionary rvfl classifier," *IEEE J. Biomed. Health Inform.*, pp. 1–9, 2023.

[35] B. Richhariya and M. Tanveer, "A robust fuzzy least squares twin support vector machine for class imbalance learning," *Applied Soft Computing*, vol. 71, pp. 418–432, 2018.

[36] R. Bharat and T. Muhammad, "A reduced universum twin support vector machine for class imbalance learning," *Pattern Recognition*, vol. 102, pp. 107–150, 2020.

[37] Y. Song, W. Cai, Y. Zhou, and D. D. Feng, "Feature-based image patch approximation for lung tissue classification," *IEEE Trans. Med. Imaging*, vol. 32, no. 4, pp. 797–808, 2013.

[38] R. Vishraj, S. Gupta, and S. Singh, "Ecm-iltp: an efficient classification model for categorization of interstitial lung tissue patterns," in *Int. Conf. Advances in Computing, Communication Control and Networking*. IEEE, 2021, pp. 481–485.

Shyla Raj is an Associate Professor in the Department of Information Science and Engineering, JSS Science and Technology University, Mysuru. She received her Bachelor degree in Computer Science and Engineering, M.Tech degree in Software Engineering and Ph.D degree from Visvesvaraya Technological University, Belgavi. She has published her research papers in reputed journals and conferences. Her research interests include medical image processing and machine learning.Email id : shylaraj@jssstuniv.in

B S Mahanand is a Professor in the Department of Information Science and Engineering; he carries 22 years of teaching and research experience at Sri Jayachamarajendra College of Engineering, JSS Science and Technology University, Mysuru, India. Dr. Mahanand has published more than 50 research articles in various peer-reviewed journals and reputed conferences. His research interests are in the areas of artificial intelligence, machine learning and medical image analysis. Email id : bsmahanand@sjce.ac.in