IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 May 2024, accepted 17 August 2024, date of publication 5 September 2024, date of current version 24 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3454966

RESEARCH ARTICLE

FibroRegNet: A Regression Framework for the Pulmonary Fibrosis Prognosis Prediction Using a Convolutional Spatial Transformer Network

PARDHASARADHI MITTAPALLI AND V. THANIKAISELVAN, (Member, IEEE)
School of Electronics Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India
Corresponding author: V. Thanikaiselvan (thanikaiselvan@vit.ac.in)

ABSTRACT Predicting the growth of idiopathic pulmonary fibrosis (IPF) is crucial for effectively treating patients affected by the disease. While the Forced Vital Capacity (FVC) serves as one of the indicators of lung functionality, accurately determining its decline solely based on previous FVC values presents a significant obstacle. We propose the utilization of a multimodal system called FibroRegNet, which capitalizes on the recent achievements in cross-model learning across general domains. FibroRegNet is designed to acquire knowledge through the regression function which maps the multimodal inputs, including CT scan and demographic information, to the coefficients of the quadratic polynomial ridge regression of FVC as outputs. FibroRegNet estimates the lung volume from a fraction of CT slices, encodes the demographic information, and combines these features with the convolutional features, from selected CT slices, that are learned through convolutional spatial transformer modules in three identical parallel streams. Trained on a publicly available database, FibroRegNet has shown significant improvement in the results compared to the related past works with a modified Laplace log-likelihood score of −6.64. Furthermore, we believe that this network has the potential to provide advantages in research domains related to the development of networks aimed at enhancing the predictive precision of IPF.

INDEX TERMS CNN, fibrosis prognosis, IPF, lung CT scans, regression, spatial transformer.

I. INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is a multifaceted and irrepressible lung ailment, with its insidious progression marked by the gradual accumulation of scar tissue within the delicate tissues of the lungs. This relentless and unforgiving process ultimately leads to a dismal prognosis for those affected by it, with a bleak future ahead [1]. The incidence and the prevalence rates recorded per 10,000 individuals are 1.30 & 4.51 in Asia-Pacific countries, 0.49 & 2.51 in Europe, and 0.93 & 2.98 in North America respectively [2], [3]. Albeit declaring it as one of the rare diseases, the average longevity rate for individuals afflicted by IPF ranges from two to four years subsequent to the identification of the condition [4]. An evaluation founded on primary care data from the United Kingdom over a period of decade, 2000-2012, estimated a upsurge of 78% in the occurrence, 100% in prevalence, at a proportion of approximately 38.8 out of 100,000 [5].

A precise and consistent assessment of the conduct of this harmful disease is crucial in order to anticipate the prognosis for individuals afflicted with IPF caused by various factors. This is especially decisive given the existence of therapies aimed at decelerating the advancement of the disease, despite the absence of a known remedy. Additionally, it is of paramount importance to identify those patients in the early stages of the condition who are most susceptible to further deterioration.

Visual inspection using conventional imaging, such as High Resolution Computed Tomography (HRCT) interpretation, has become a common practice in detecting early-stage interstitial lung disease (ILD), the other methods being clinical diagnostic tests that include lung function tests and tissue

The associate editor coordinating the review of this manuscript and approving it for publication was Zhen Ren.

biopsy. Even though the visual evaluation of CT scans for the assessment of PF progression is relatively easy compared to the other two, nevertheless, it is crucial to acknowledge that it necessitates multiple CT scans, a requirement that is prohibitively expensive, time-consuming, and potentially hazardous to patients due to their exposure to ionizing radiation. At the same time their visual evaluation is encumbered by inter-observer variability. This has motivated researchers into the development of computer-based, objective CT assessment techniques [6], [7], [8], [9].

When viewed through high-resolution computed tomography scans, idiopathic pulmonary fibrosis has a unique appearance in the lungs such as areas of scarring, honeycombing, and scattered ground-glass spots that may or may not have fluid presence in the pleural space [10]. However the development of CAD systems for the visual discovery of these features can be challenging due to (1) the characteristics of IPF and pneumonia can exhibit similarity and overlay in terms of CT abnormalities, making it challenging to visually differentiate between the two, consequently, discerning between various types of pulmonary fibrosis can pose difficulty, (2) high resolution CT is also sensitive to comorbidities, such as lung cancer, however, the presence of these comorbidities can complicate the distinction between fibrosis and the other lung diseases, and (3) HRCT may not possess the required sensitivity to detect subtle changes in lung tissue and depends on the precision of the corresponding equipment [9], [11], [12], [13], [14], [15]. However, as a result of the highly volatile and uncertain nature of this illness, it presents a formidable undertaking even for proficient radiologists, thus compounding the difficulty in establishing the prognosis for individuals afflicted with IPF [10].

Recently, deep learning has penetrated into and empowered many fields. Models that are successful on general benchmark datasets started showing better results on the new datasets of other downstream tasks. Medical domain has many daunting tasks which often needs the analysis of different variety of information. However, unavailability of standard medical image and correlating clinical datasets is always barring the adaptation of rapidly growing research in deep learning. The basic objective of our research effort is the development of a method for the prediction of pulmonary fibrosis progression using publicly available OSIC multimodal dataset [16]. This paper presents our contributions to help solve the task of predicting the FVC providing solutions to the challenges thereof as below:

1. FVC decline, unlike as a measure of its slope, is empirically formulated as quadratic polynomial ridge regression mechanism resulting in three coefficients of the regression as target variables to be predicted by the model from which FVC can be calculated.
2. Options for learning embedding representations from demographic information is explored and concluded that its use require more clinical data that is descriptively correlated to CT scans, hence choose encoding techniques as a better option to obtain features from demographic modality.
3. Employed a pretrained lung segmentation model and estimated the lung volume as a feature from CT modality.
4. A customized convolutional spatial transformer modules are utilized to learn the spatially invariant CT features in the first phase of training. Then all the three multimodal features were fused and the second phase of training was accomplished for the prediction of three regression coefficients.

In the subsequent sections of the paper, we contemplate and summarize the related literature in section II, preprocessing of multimodal information, and the details of the development of our proposed model FibroRegNet in section III, the evaluation metrics used, experiments conducted, and ablation study carried out are presented in section IV, followed by discussion and conclusion in section V and VI respectively.

II. RELATED WORK

This section provides an exposition of recent investigations that have employed deep learning methodologies to examine thoracic CT images, with particular relevance to the evaluation of FVC decline in relation to various modalities of available data. The advantages and limitations of each of them are also summarized in TABLE 1. Refaee et al. [4] worked on a private dataset and evaluated their method on LTRC dataset [17]. They discussed that hand crafted radiomics and the automatic features from deep learning network can differentiate between IPF and other ILDs using HRCT scans and accordingly developed an HCR model, a DL model, and an ensemble of HCR and DL model. Finally concluded that the ensemble model performed well with an accuracy of 85.2%+/-2.7%. They also interpreted their models using GRAD-CAM plots [18]. Moua et al. [19] built a custom layer called the attention gate which produces two outputs, first the domain knowledge-based attention map and the second being a feature map that is a weighted combination of the input feature map and the attention map. This feature map was fed into the subsequent feature learning residual modules of the CNN. Two attention-based loss functions each with a different scale and the IPF prediction probabilities from the final classification layer of the CNN were fed to a random forest where the final classification task is carried out. Recent methods of quantifiable radiological scoring was studied in [20]. U-net with DenseNet as its backbone architecture developed and evaluated in [21]. Cystic fibrosis is classified for the purpose of studying the therapy response by [22]. Yadav et al. [10] proposed the combined use of metadata, degree of honey combing measured from CT scans using traditional image computation methods, and the features learned from the selected slices of CT scans in the penultimate dense layer to predict the decline trend in the FVC. Further, [23] employed a 3 layered ResNet model to determine the decline. Nazi et al. [24] studied multiple quantile, ridge, and elastic ridge regression technics for the purpose assessing

TABLE 1. Related work.

Ref.	Datasets	Task	Method	Advantages	Shortcomings
Rafaee et al., 2022 [4]	Training on private dataset, testing on public HRCT	IPF & non-IPF ILD	A random forest model and Densnet-121 based deep learning models combined to distinguish IPF from non-IPF ILD.	Utilized the hand crafted and neural networks based automatically extracted features from HRCT	The sensitivity and specificity of the ensemble model was not studied when HCR and DL models were not in agreement. IPF Prognosis not predicted.
Wenxi et al., 2023 [19]	Private-HRCT	IPF & non-IPF ILD	A two stage approach consisting of multi scale attention model guided by domain specific knowledge followed by a random forest classifier was proposed.	Population level domain specific knowledge is utilized. Customized attention gates and the use of RF classifier in the final decision stage improved the model performance.	Population level domain specific knowledge is necessary.
Kasara et al., 2019 [22]	Private-HRCT	Cystic Fibrosis classification	A base line CNN model trained on 1100 lung HRCT slices.	Relatively a simple network	High false positive rate
Taneja et al., 2023 [23]	Public- OSIC	IPF progression	ResNet based Sky-Net was proposed	Predicts the FVC decline	Demographic information not utilized
Mandal et al., 2020 [24]	Public- OSIC	IPF progression	Several regression techniques studied	Elastic Net regression found to be the better one.	CT features extraction and their utilization in the progression prediction not studied extensively.
Nazi et al., 2021 [25]	Public- OSIC	IPF progression	Convolutional Self Attention model was proposed.	The Self Attention (SA) module played crucial role in extracting class specific features	Linear regression in the estimation of FVC decline from its baseline value
Kim et al., 2021 [26]	Private-HRCT	IPF progression	Random forest classifier optimized by quantum particle swarm algorithm.	Voxel level classification of the whole lung region enables the classification of a specified annotated ROI into "expected to progress" and "expected to be stable" predictions.	HRCTs are collected only from IPF subjects.
Aoki et al., 2022 [27]	Private-HRCT	ILD classification & IPF progression	An existing CNN based Ziosoft Informatics Platform was used to find the percentage of the fibrotic lesion volume.	Measured the extent of IPF lesions using deep learning methods and found that FVC had a negative correlation with lesion extent.	Demographic information not explored, not tested on any public datasets.
Wu et al.., 2022 [28]	Private-HRCT	IPF progression	Using CT analysis and pulmonary function grades a comprehensive final risk prediction model was built	The proposed method provides the leverage of fast incremental learning system.	This method relied upon the segmentation of honeycombing only.
Wong et al., 2021 [29]	Public- OSIC	IPF progression	Proposed a selective long range connectivity to a baseline CNN architecture.	The last dense layer of the network is given with demographical clinical data preventing the possible overfitting.	Even though state of the art results were obtained, the model is not production ready
Kotowski et al., 2023 [30]	Public-CHAOS Private- CLIN	Detection of liver cirrhosis that leads to liver fibrosis	Random forest classifier was trained on clinically inspired radiomic features extracted from CT scans	The pipeline not only classifies scans but also identifies discriminative features, aiding in understanding the disease progression.	Understanding how specific features contribute to the classification of cirrhotic and non-cirrhotic states could be challenging, potentially hindering clinical adoption.
Du et al., 2022 [31]	Private-HRCT	IPF progression	Fibrotic lung volume was quantified by Synapse software.	The formulas used for the assessment of fibrotic lung volume were confirmed to be with high sensitivity and specificity.	Performance is limited by the data quality, validation of the model was done on a smaller cohort.
Kawahara et al, 2022 [32]	Private-HRCT	IPF progression	Radiomics features were extracted from CT slices using a hybrid auto-segment method.	Study showed the visual scores significance of reticulation and honeycomb in prognosis assessment.	Evaluated on a very small dataset.

of IPF progression. Self-attention mechanism was utilized for CT feature extraction, followed by the fusion of lung volume feature and demographic features [25]. Syed et al. [9] applied transfer learning on six popular image classification models pretrained on ImageNet for the classification of IPF using [16]. After applying several conventional preprocessing, and data augmentation techniques, and performing the optimization of hyper parameters they have pronounced ResNet50V2 as the best-performing model. Aoki et al. [26] employed statistical learning paradigm. Wu et al. [27] proposed deep learning based analysis for the classification of typical ILD. The lesion extent they predicted correlated well with the FVC indicating that CT scan analysis can provide the measurement of pulmonary fibrosis progression. Wong et al. [28] performed honeycomb segmentation for the estimation of risk. Kotowski et al. [29] proposed a dense connected like selective long range connectivity to a baseline CNN architecture.

Vision transformers (ViTs) have played a significant role in advancing medical image analysis in recent years [33], [34]. But their extensive use in the analysis of medical images is hindered by the following reasons: (1) ViTs require large datasets for training, (2) ViTs lack inductive biases related to image structure and translational invariance, and (3) self-attention mechanism in ViTs only operates between image patches which can lead to missed critical information when

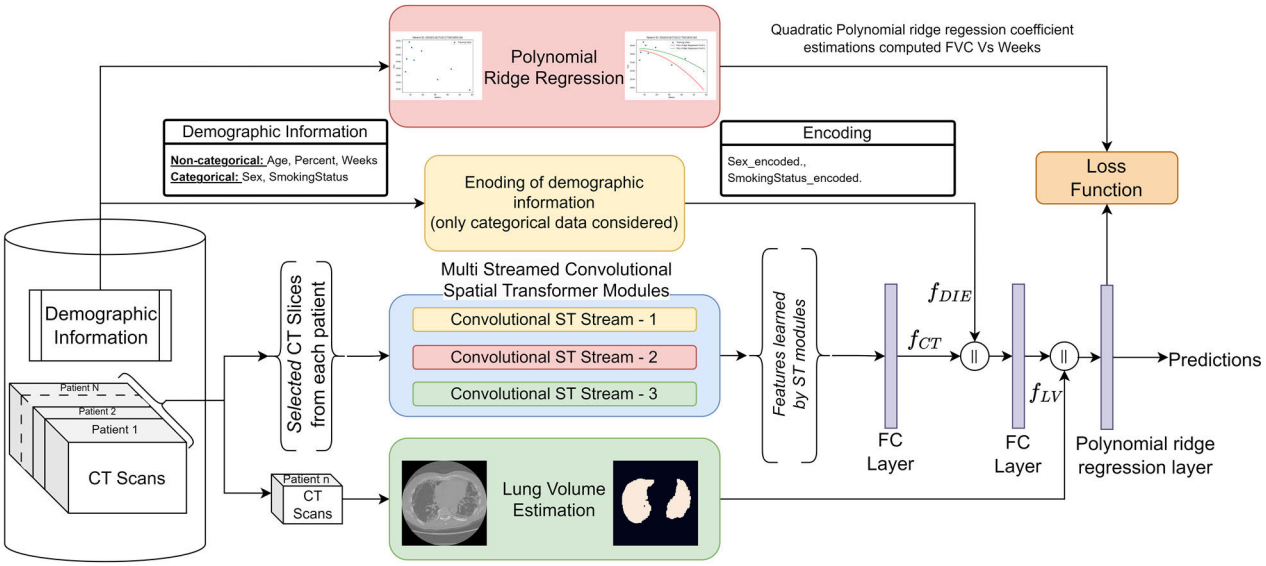

FIGURE 1. The proposed FibroRegNet for the prediction of pulmonary fibrosis prognosis.

segmenting small objects or regions with blurred boundaries. Ongoing research aims to address the limitations by developing more efficient architectures, incorporating inductive biases, and leveraging large-scale pretraining. But the small dataset size remains a fundamental challenge. Finally, this literature review reveals that the studies on the prediction of disease progression is meagre except a few works [10], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32].

III. PROPOSED WORK

In the proposed work, we explored novel approaches in regression techniques for FVC data preprocessing for the purpose of deciding whether linear or nonlinear regression coefficients are to be used as target variables. We empirically decided to use quadratic polynomial ridge regression. Examined encoding and embedding techniques for demographic data preprocessing. We took decision in favor of encoding as it precedes over embedding in the cases of low cardinal and ordinal data. The proposed system made a significant advancement by integrating deep learning-based image segmentation models for precise lung volume estimation. In particular, we pioneered the application of the Spatial Transformer Network (STN) in medical image analysis, marking an innovative use of STN in this domain. We introduced the concept of employing parallel streams of cascaded STN layers to enhance feature extraction for improved generalization. This innovative method aimed to extract features with enhanced generalization capabilities, pushing the boundaries of feature extraction in medical image analysis.

Furthermore, our research culminated in a unique approach where preprocessed demographic features, estimated lung volume feature, and CT features extracted by cascaded STN streams were synergistically combined at the dense layers, developing a novel polynomial ridge regression layer. Integrating diverse data sources through innovative feature extraction techniques represents a significant advancement in medical image analysis and predictive modelling.

The complete structure of the proposed FibroRegNet is depicted in Figure 1. The two phase training framework that we propose contains the following essential steps: (i) the extraction of f_{CT} features from the preprocessed CT slices using a CNN model equipped with a Spatial Transformer (ST) modules in the first phase of the training, and (ii) in the second phase of training, the integration of those f_{CT} features with demographic features f_{DIE} and shallow lung volume estimation features f_{LV} one after the other in the subsequent fully connected layers, followed by a linear layer that predicts the quadratic polynomial ridge regression coefficients of FVC.

A. FVC DATA PREPROCESSING

In this section, we discuss our concerns about preprocessing the target variable, FVC, for the network to do regression on. We do have FVC measurements nonetheless there are multiple measurements for each patient across several weeks irregularly distributed ranging from -5 to 133 weeks of time, where sign of the number indicates FVC measurements that were taken weeks before and after the acquisition of the CT scan. We preferred to consider the coefficients of polynomial ridge regression as the target variables rather than linear regression, the decision was made empirically (detailed discussion is in Section IV) looking at the data and the root mean square errors of both the regressions. To avoid overfitting to the data, we restricted ourselves to the quadratic polynomial ridge regression only. The process has resulted in three coefficients which eventually became the target variables for our model. Overall, the prediction of fibrosis prognosis is formulated as a multivariate regression problem that is to

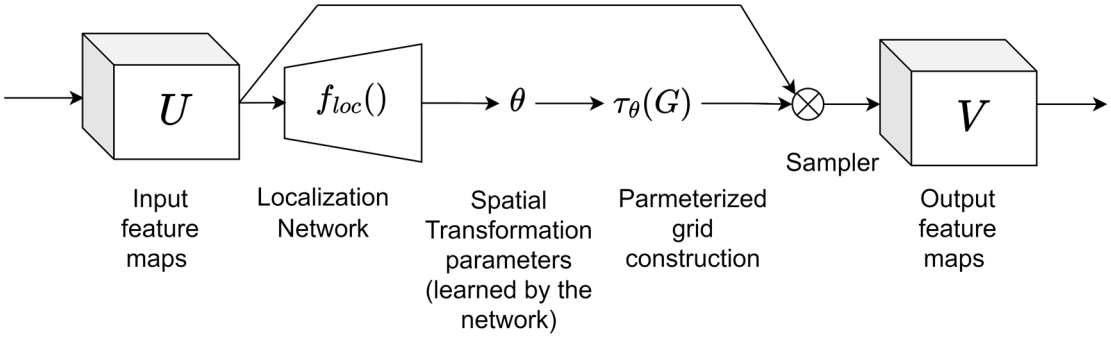

FIGURE 2. Spatial transformation module. Localization network learns the spatial transformation parameter vector of size 2 × 3 which will be different for different input feature maps. Constructs the parameterized grid on the input. Finally, produces the output using bilinear interpolation.

be solved using FibroRegNet model that is to be trained on multimodal data that include CT scans and the demographic information.

B. DEMOGRAPHIC DATA PREPROCESSING

The clinical data published by the OSIC dataset contains a wealth of valuable demographic information in the form of both categorical and non-categorical data which can greatly enhance the capabilities of our model. By extracting relevant feature representations from this data, we can improve the accuracy and effectiveness of our predictions. Typically, the categorical text data can be converted into a robust numerical representations using two techniques embeddings and encodings [35]. Embeddings is a learning technique that can be incorporated into the machine learning models which generate dense numerical representation feature vectors during the training phase. We discussed why we have not used embeddings in section IV even though they are recent and gained wide popularity in dealing with text data in the domain of natural language processing. On the other hand encoding comes under preprocessing techniques. The popular encoding techniques those can be used in preprocessing stage are one-hot encoding, label encoding, or ordinal encoding. We have identified that in many cases, learning models struggle to properly interpret these encoded representations due to the challenges such as high dimensionality, sparse representations, correlated features, and/or ordinal mismatch. As a result, even small changes in the encoded data can lead to significantly different model outcomes.

However, after examining our data we have excluded the 'Patient' field as it simply serves as an identifier and does not contain any meaningful information. Among the remaining fields, 'Weeks', 'Percent', and 'Age' are all non-categorical continuous numerical data. The other columns 'Sex' and the 'SmokingStatus' are the only two columns that are existing as categorical non numerical features with a cardinality of 2 and 3 respectively. As both feature's cardinality is less and also doesn't have any issues like inherent ordering, we decided to do one-hot-encoding of the both to get their numerical representations.

C. LUNG VOLUME ESTIMATION

In addition to considering demographic factors, we pre-computed volume of the lung based on CT scans for each individual. The primary rationale behind estimating the lung volume was to include an approximation of this measure as a part of the feature set. The process require the lung segmentation, but as OSIC dataset hasn't provided any lung masks, we resorted to the use of pre trained lung segmentation model [36]. They examined the importance of training data diversity for the automation of lung delineation in routine Computed Tomography (CT) imaging. Their model showed superior performance while inferencing on diverse routine publicly available databases and a large dataset privately collected that are showing a variety of pathologies as well. These facts motivated us to use their U-net (R231) system as our pre-trained model for OSIC lung CT scan segmentations.

In order to avoid the computational complexity associated with including all CT slices for extracting volumetric features, we selectively employed the U-net (R231) on a set of 85 percent of CT slices to generate a binary lung segmentation map. Through the elementwise product of this segmentation result and the original CT image, we obtained the segmented lung image. The process of generating the lung volume, can be formalized in terms of voxel volume and voxel count. The basic formula is straightforward which is given (in milliliters) as:

$$\text{Lung Volume}(f_{LVE}) = N \times V_w \times V_h \times V_d \quad (1)$$

where N is the count of lung voxels, and $\mathbf{V_w}$, $\mathbf{V_h}$, and $\mathbf{V_d}$ are the width, height, and depth of the single voxel retrieved from the metadata of the DICOM files.

D. NETWORK ARCHITECTURE OF FibroRegNet

The fundamental building block of our proposed model is the convolutional spatial transformer module. We present a detailed explanation of the formulation of a spatial transformer here. Subsequently we discuss the adaptation of this module to learn the CT modality features f_{CT}. Finally, we perform the fusion of the multimodal features learned from shallow CT lung volume f_{LV} and demographic information

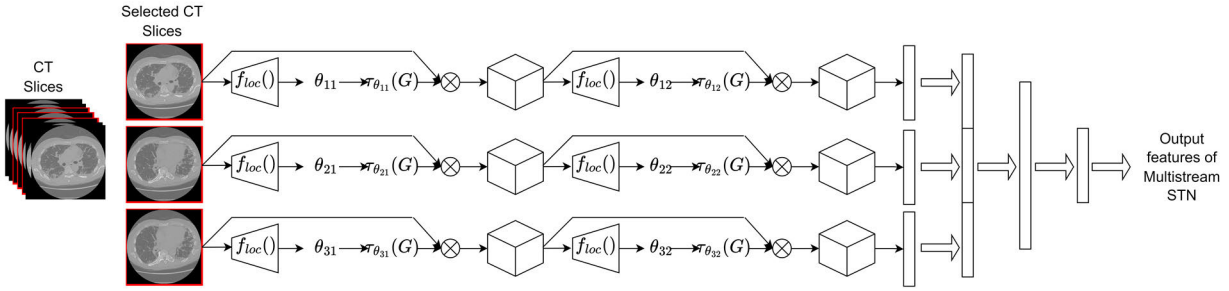

FIGURE 3. The three streams of convolutional spatial transformer modules. Three slices are randomly selected from the central ten percent of the slices. These slices are given as input to the first stage spatial transformer module one in each of the streams.

encodings f_{DIE} with that of CT features f_{CT} learned by the spatial transformer CNN model.

1) SPATIAL TRANSFORMER

A spatial transformer module as proposed in [37], can be defined as module that is differentiable and processes input features by taking a spatial transformation in a single forward pass. The spatial transformations are input dependent. This means the spatial transformation matrix is learned through training and it will be different for different inputs based on the input feature map and the each channel in the input undergoes same transformations. The mechanism has three procedures to be applied in sequence, as illustrated in Figure 2, a localization network, parameterized sampling grid construction, and differentiable image sampling and interpolation.

Localization Network: The intricate network of localization receives $\mathbf{U}\in\mathbf{R}^{(\mathbf{H}\times\mathbf{W}\times\mathbf{C})}$ as its input. It then outputs the parameters of the spatial transformation $\mathbf{A}_{\boldsymbol{\theta}}$, denoted as $f_{loc}(\mathbf{U})$. The function $f_{loc}(\mathbf{U})$ utilized for this transformation, in our case, has three 2D convolution blocks. These blocks are composed of convolution, 2D max pool and ReLU layers where the kernel sizes are 7×7, 5×5, and 3×3 respectively. We also did one 1×1 convolution and ReLU to capture the channel attention. These convolution features are then fed to two fully connected layers first one with ReLU non-linearity and the second one performing linear regression to predict six parameter values necessary for an affine transformation. The parameter vector θ is then reshaped in to 2×3 transformation matrix $\mathbf{A}_{\boldsymbol{\theta}}$ given as

$$\mathbf{A}_{\boldsymbol{\theta}} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \tag{2}$$

Parameterized Grid Construction: This matrix $\mathbf{A}_{\boldsymbol{\theta}}$ will be responsible for executing rotation, translation, cropping, and scaling spatial transformations that are conditioned on specifics of the input feature vector. Next, the transformation parameters are utilized to construct a sampling grid G, composed of precise set of locations in input feature map that must be sampled to harvest the transformed output. In general, this output feature map $\mathbf{V}\in\mathbf{R}^{(\mathbf{H}'\times\mathbf{W}'\times\mathbf{C})}$ may differ in terms of its spatial size $H' \times W'$ from the input feature size $H \times W$ but they both will have same number of channels.

Sampling and Interpolation: Computation of this feature map involves to steps sampling and interpolation. Sampling is basically finding the pixel locations (x_i^s, y_i^s) in the input feature vector that are corresponding to the pixels (x_i^t, y_i^t) in the output feature vector. This is done through the sampling grid G that is constructed based on the parameterized spatial transformation $\tau_\theta(G)$ as given below.

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \tau_\theta\,(G) = \mathbf{A}_{\boldsymbol{\theta}} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{3}$$

Sampling is followed by interpolation which computes the pixel values of the output feature vector V based on the corresponding pixels in the input feature vector U using the bilinear interpolation as given by equation (4)

$$\mathbf{V}_\mathbf{i}^\mathbf{C} = \sum_{\mathbf{n}}^{\mathbf{H}} \sum_{\mathbf{m}}^{\mathbf{W}} \mathbf{U}_{\mathbf{mn}}^{\mathbf{C}} \mathbf{max}\left(\mathbf{1} - \left|\mathbf{x}_\mathbf{i}^\mathbf{s} - \mathbf{m}\right|\right) \mathbf{max}\left(\mathbf{1} - \left|\mathbf{y}_\mathbf{i}^\mathbf{s} - \mathbf{n}\right|\right) \tag{4}$$

CT Modality Feature Learning: The spatial transformer module, with its ability to learn features that are invariant to spatial variations, is a crucial addition to our FibroRegNet. We have incorporated this module into our design by implementing it in three streams, each containing two ST modules, as depicted in Figure 3. Each ST module is followed by two convolution blocks consisting of 3×3 convolution, 2×2 max pooling, and ReLU activation, along with a dropout layer for regularization. The resulting feature vectors are then flattened and passed through a fully connected ReLU layer in each stream before being concatenated and fed into two additional fully connected ReLU layers for further processing. This integration of the spatial transformer module adds an extra layer of adaptability and robustness to our network architecture.

2) FUSION OF FEATURE REPRESENTATIONS OBTAINED FROM MULTIMODAL INFORMATION

We have a total three categories of features two of them, f_{DIE} and f_{LV}, obtained in the preprocessing stage and the third, f_{CT}, is learned from convolutional ST streams of FibroRegNet. We can fuse them only through concatenation as they are

TABLE 2. The dataset has a total of 176 CT scans, 33026 slices, and 1549 FVC values.

Patient ID	Weeks	FVC	Percent	Age	Sex	Smoking Status
ID00007637202177411956430	-4	2315	58.25365	79	Male	Ex-smoker
ID00009637202177434476278	11	3895	90.75869	69	Male	Ex-smoker
ID00019637202178323708467	66	1778	78.62038	83	Female	Ex-smoker
ID00020637202178344345685	18	2297	117.7707	66	Female	Never smoked

learned from three different modalities of information as shown in Figure 1. First, we combined CNN features, f_{CT}, obtained from the CT modality through the spatial transformer driven CNN, with the encoded features, f_{DIE}, obtained from demographic information and forwarded to ReLU nonlinear layer for fine tuning the further information processing. Later, the lung volume, f_{LV}, feature is fused and forwarded to the last linear layer consisting of three neurons which will learn the three coefficients of the polynomial ridge regression. FVC is then computed using the following equation:

$$\mathrm{FVC}\,(w_i) = a \times w_i^2 + b \times w_i + c \tag{5}$$

where w_i is the i^{th} week for which the FVC is to be calculated, constants a, b, and c are the predicted coefficients.

IV. EXPERIMENTS

This section begins with the description of dataset, evaluation metrics, implementation and training strategies adapted in our work. Then we assessed the efficacy of the proposed FibroRegNet in predicting the pulmonary fibrosis progression. Initially, we emphasized the performance achieved by our suggested multi modal learning system through two-phase training process and subsequently compared it with recent approaches. Next, we presented the ablation study that we conducted with regard to the choice of considering quadratic polynomial ridge regression of FVC values and the encoding of demographic information. Finally we discuss the selection of hyper parameters and the network configurations that we tested and their associated results.

Dataset: The pulmonary fibrosis progression dataset by OSIC was used in our model training and evaluation was sourced from Kaggle [16]. This dataset consists of CSV metadata and CT scans for each patient. The metadata of the dataset comprises 1549 entries in seven columns, including the Patient's ID, Percent, Age, Smoking Status, Weeks, Sex, and Forced Vital Capacity (FVC). A few rows of the CSV file are presented in TABLE 2. The CT scans of the dataset are in DICOM format with the number of slices varying highly from patient to patient with a minimum of 12 scans and a maximum of 1018 scans per patient. The range of values for Weeks is from -5 to 133, negative number indicates the FVC measurement taken week number before the CT was acquired and the positive number after. We were provided with 6 to 10 FVC measurements per patient across unevenly distributed weeks and a CT scan. Our goal is to predict the value of FVC for any specified future week from the CT scan without any knowledge about the patient.

Evaluation Metric: We employed a modified version of the Laplace Log-Likelihood ($\mathrm{LLL_m}$) and the root mean square error (RMSE) as metrics for evaluating the performance of our proposed model. The selection of $\mathrm{LLL_m}$ was based on its ability of combining both the certainty of its predictions and accuracy to assess the confidence of the model. To determine the FVC and its associated confidence measure with respect to the actual FVC measurements, we utilized the following calculation method which was given in [16]:

$$\sigma_{\mathrm{clipped}} = \max(\sigma, 70) \tag{6}$$

$$\Delta = \min\left(\left|\mathrm{FVC_{True}} - \mathrm{FVC_{Predicted}}\right|, 1000\right) \tag{7}$$

$$\mathrm{LLL_m} = -\frac{\sqrt{2}\Delta}{\sigma_{\mathrm{clipped}}} - \ln\left(\sqrt{2}\sigma_{\mathrm{clipped}}\right) \tag{8}$$

Here, σ represents the standard deviation, and the threshold of 1000 ml was set to avoid imposition of any unfavorable penalty resulting from large errors. To account for the approximate measurement uncertainty in FVC, we capped the confidence values at 70 ml. The final score was derived by averaging the metric across all weeks. It is worth noting that the metric value was consistently negative, with lower values indicating better performance.

Implementation Details: We employed the publicly accessible PyTorch framework [38] to develop our proposed FibroRegNet. The network training is aimed at the minimization of the L1 loss. The experimentation was conducted on hardware comprising an Intel(R) Xeon(R) Silver 4208 CPU and an Nvidia Tesla V100 PCIE GPU with 187GB and 32GB of physical RAM respectively. We adapted a two phase training strategy considering the inherent nature of the fusion of the data modalities involved in the model training. The training process involved optimizing models using the Adam optimizer. The weight decay regularization is applied independently of the learning rate updates. This means that the weight decay term, set to the value of 0.01, is not directly multiplied with the learning rate as in the original Adam algorithm.

The dataset is split into training (80%) and testing (20%) sets. The split is stratified by patient to ensure that all scans from a single patient are contained within the same set, preventing data leakage and ensuring the model's ability to generalize to unseen data. K-fold cross-validation (with K=5) is employed to assess the model's performance and

TABLE 3. Cross-validation statistics of the training and testing datasets. The dataset has a total of 176 CT scans, 33026 slices, and 1549 FVC values. The train-test split is stratified by patient IDs.

Folds	Training			Testing		
	#Patient IDs	#Slices	#FVC values	#Patient IDs	#Slices (test)	#FVC values
0	140	25934	1229	36	7092	320
1	141	26765	1240	35	6261	309
2	141	26660	1237	35	6366	312
3	141	26522	1241	35	6504	308
4	141	26223	1249	35	2803	300

TABLE 4. Performance comparison between different modalities. LLL_m, RMSE.

Modality	LLL_m	RMSE
Demographic + Lung volume	-6.69±0.35	174.92±19.8
CT modality	-6.67±0.24	174.51±16.4
Multimodality	**-6.64±0.42**	**172.35±16.2**

TABLE 5. Performance comparison of FibroRegNet with recent works in terms of LLL_m and RMSE. Our proposed FibRegNet outperforms the existing works on predicting the progression of pulmonary fibrosis. (where c.f means 'copied from').

Reference	Type of Regression	LLLm	RMSE
Wong et al [29] Fibrosis-Net	Elastic Net	-6.82	-
Mandal et al [24]	Quamtile	-6.92	-
	Ridge	-6.81	-
	Elastic Net	-6.72	-
Nazi et al [25] Fibro-CoSANet	EffiecientNet-b2	-6.68 ± 0.31	181.5 ± 25.88
	ResNet50	-6.68 ± 0.31	181.6 ± 22.89
	EffiecientNet-b1	-6.68 ± 0.28	182.58 ± 24.04
	EffiecientNet-b3	-6.68 ± 0.28	183.96 ± 22.89
OSIC 2020 (c.f [29])	Kaggle 1st Place	-6.8305	-
OSIC 2020 (c.f [29])	Kaggle 2nd Place	-6.8331	-
OSIC 2020 (c.f [29])	Kaggle 3rd Place	-6.8336	-
Yadav et al [10]	FVC-Net	-6.641	-
Poulou et al [42]	Regression based approach	-6.8950	-
Shehab et al [43]	CNN + Quantile Regression	-6.6409	-
Our model FibroRegNet	Three streams with two STN layers in each	**-6.64 ± 0.42**	**172.35 ± 16.2**

TABLE 6. Comparison of in terms of Macs (G), number of parameters (M), inference (s), LLL_m, and RMSE.

Reference	Macs (G)	Params (M)	Inference (s)	LLL_m	RMSE
Nazi et al [25]	0.1	6.65	0.7	−6.68 ± 0.28	183.96 ± 22.89
Wong et al [29]	-	1.38	0.053	−6.68188	-
Our model FibroRegNet	0.63	22.4	0.81	**-6.64±0.42**	**172±16.2**

stability across different subsets of the data. This technique involves dividing the training set into k smaller sets, using $k-1$ of them for training and the remaining set for validation. This process is repeated k times, with each of the k sets used exactly once as the validation data. The quantitative characteristics of the cross-validation is presented in TABLE 3.

A. RESULTS OF THE PROPOSED FibroRegNet

To determine the effectiveness of our proposed technique that combines multiple modalities in feature fusion, we conducted a careful study with different modalities individually. This involved comparing our technique with other available modes to demonstrate its superiority. We carried out experiments using three distinct modes: (i) Shallow Modality, which trained the model using only lung volume and demographic features without any CNN backbone, (ii) CT modality, which solely utilized convolutional spatial transformer features from CT scans, and (iii) Multimodality, which incorporated, demographics, lung volume features, and convolutional spatial transformer features from CT images. Our results show that the multimodality modes outperformed both standalone CT modality and shallow modality in terms of LLL_m and RMSE as shown in TABLE 4. This suggests that while the demographics along with either lung volume or CT scan features individually result in reasonable performance, combining these two modalities enhances overall performance.

B. PERFORMANCE COMPARISON WITH RECENT METHODS

In our study, we performed a thorough comparative analysis with the recent approaches which have reported notable results in the prediction of pulmonary fibrosis progression and presented in TABLE 5. Nazi et al. [24] implemented a combination of three different types of regression that include elastic net, ridge, and multiple quantile regression in their predictive model, with the elastic net method yielding the

most successful results $-\text{LLL}_\text{m}$ at a value -6.72. Kotowski et al. [29] achieved a value of $-\text{LLL}_\text{m}$ of -6.82 using their own methods. Kim et al. [25] proposed the inclusion of lung volume estimation and CT features and performed linear regression on FVC values and achieved an $-\text{LLL}_\text{m}$ of -6.68. The winners of the OSIC Pulmonary Fibrosis Progression contest hosted by Kaggle in the first three places yielded $-\text{LLL}_\text{m}$ values of -6.8305, -6.8311, and -6.8336 respectively. Poulou et al. [42] using their regression approach achieved -6.8950 score. Yadav et al. [10] with their FVC-Net and Shehab et al. [43] with their CNN and quantile regression methods were able pull the $-\text{LLL}_\text{m}$ score to -6.41 and -6.409 respectively. Through the implementation of FibroRegNet, our proposed algorithm outperformed them, achieving an impressive $-\text{LLL}_\text{m}$ score of –6.64 and an RMSE of 172.35 ± 16.2 as shown in TABLE 5. The close competition in the performances of [10] and [25], tells us that the combination of EfficientNet and Quantile regression can perform better in these kind of tasks. Besides that, Yadav et al. [10] and Kim et al. [25] have presented an interesting approach of involving estimated honeycomb and lung volumes respectively as a feature in the prediction process.

The proposed model is different from these existing models in terms of many aspects yet outperforming all of them. We got motivated by [10], [25], [29], [42], and [43] about the use of regression techniques. But unlike the quantile regression, as used by all of them, we used quadratic polynomial ridge regression in the preparation of the target variables from the FVC values. The quadratic polynomial ridge regression can model the relationship between the predictor(s) and the mean of the outcome variable using a quadratic function which is relatively an easy task compared to quantile regression. We definitely got inspired by [10] and [25] for using lung volume as one of the feature. But unlike honeycomb volume used in [10], we used a pretrained lung volume estimation model. For the extraction of CT features, all most all models relied upon already existing state of the art CNN models like the variants of EfficientNets. Instead, we incorporated convolutional STN based feature extraction layers. This is the first time STNs being used in the medical image analysis as per our knowledge. Finally, we adapted a unique approach where preprocessed demographic features, the estimated lung volume feature, and CT features extracted by cascaded STN streams were synergistically combined at the dense layers, developing a novel polynomial ridge regression layer. All these adaptations and innovative approaches lead to the superior performance of the proposed model.

C. COMPUTATIONAL COMPLEXITY

The computational complexity in terms of Macs (G), number of trainable parameters (M), inference time (s), LLLm, and RMSE in comparison with other existing models based on as much as the availability of those metrics are presented in TABLE 6. The main goal of this work is to enhance the prediction performance and we achieved it at the cost of requiring more computational resources.

D. ABLATION STUDY

1) WHY POLYNOMIAL RIDGE REGRESSION RATHER THAN LINEAR

We began by testing a quadratic polynomial regression against a linear regression, using the null hypothesis that "*the quadratic model is not necessary*." The obtained F-statistic and p-values of 1.53 and 0.28 respectively rejected this null hypothesis, indicating that the data can be accurately modelled using quadratic polynomial regression. However, due to the small amount of data per patient, we chose to use quadratic polynomial ridge regression as a safeguard against overfitting. We selected an alpha value of 1000 based on the range of FVC values in our database. The fitting curves for some of the patients were shown in Figure 4. To ensure the model was not overfitting, we also performed residual analysis. The results were satisfactory with mean R-squared and mean MSE scores of 0.85 and 174.98 respectively. After examining the residual plots, which showed asymmetry around the ideal zero residual axis as expected, we gained confidence in using quadratic polynomial ridge regression. This has resulted in three coefficients, the target variables, which the model has to learn.

2) WHY NEED TO CONSIDER DEMOGRAPHIC INFORMATION

According to the OSIC database, there is a negative correlation of -0.09 between FVC and Age. It is important to keep in mind that correlation does not necessarily imply causation and the strength of a relationship cannot be determined solely by the correlation coefficient. The FVC in case of healthy adults who have never smoked would decline at a rate of 0.2 liters per decade [39], but its decline may be more closely correlated with age once the disease sets in. Additionally, it can be observed that males tend to have higher FVC values than females regardless of age or smoking status, as shown in Figure 5. An interesting fact is that most female patients are non-smokers yet still have lower FVC values, unlike most of the male patients who are either ex-smokers or current smokers. The same pattern can also be seen concerning the 'Percent' feature values as well. These findings suggest that not only do initial FVC values play a role in predicting outcomes, but the rate of decline may also be influenced by variables such as age, gender and smoking status. Our model has taken these factors into account, resulting in improved prediction performance as shown in TABLE 4 and TABLE 5.

Why embeddings of the categorical data is not considered?

In section III, we mentioned that the one-hot-encoding is being used to get the numerical representations of categorical data. In this section we present our study on using word embeddings in place of one-hot- encoding of such categorical data. Embeddings can be learned by the model conditioned on the input data, unlike encoding which can be considered as data preprocessing task rather than learning task. We used the process called entity embedding as described in [40] for this purpose. Entity embeddings capture relationships between

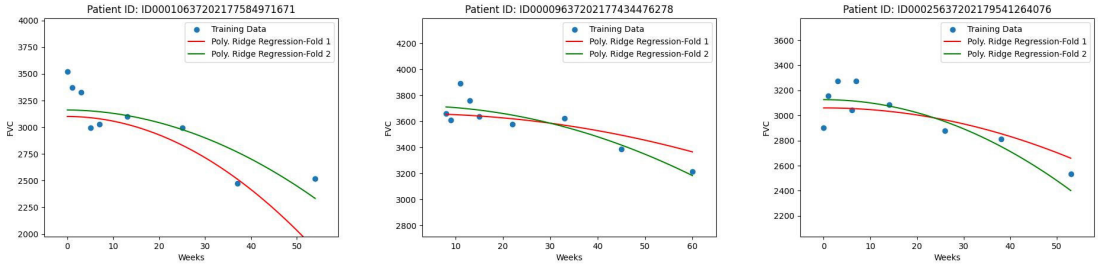

FIGURE 4. The results of quadratic polynomial ridge regression of a few patients. We performed this regression over two folds for every patient. The result of regression in the first fold is displayed in red color and in the second fold with green color (best viewed in color).

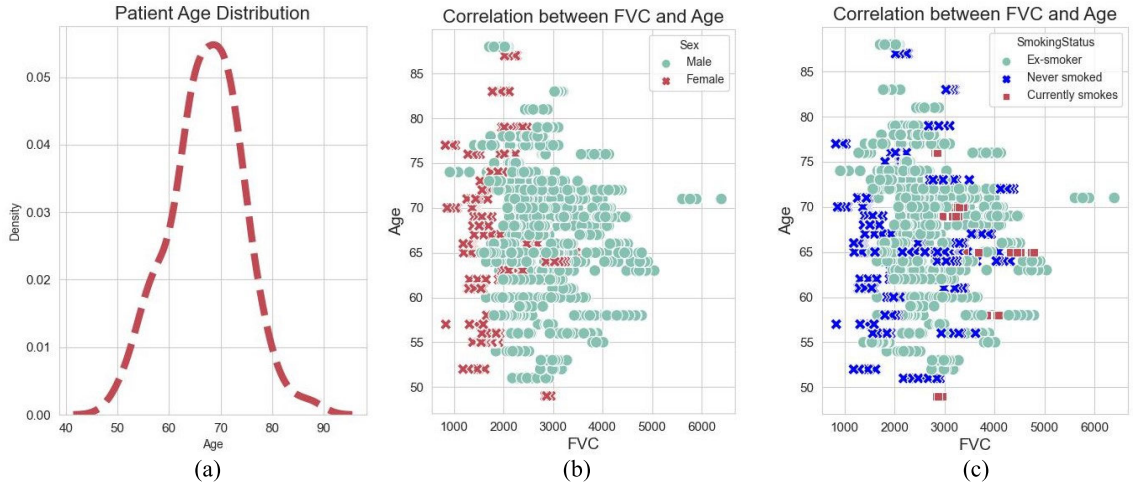

FIGURE 5. (a) Age distribution of the patients involved in the study of the dataset, the Gaussian bell shape indicates that the Age feature is well distributed suitable for any nature of system. (b), and (c) presents the correlation between FVC vs Age against Sex and Smoking Status respectively (best viewed in color).

categorical values by representing them as dense vectors of real numbers. These vectors are learned through neural network training, where the network adjusts the embeddings to minimize the loss function during the learning process. As a result, similar categorical values end up having similar embedding vectors, and the distances and directions between these vectors can reflect the relationships and similarities between the categorical values. This allows the embeddings to capture the underlying structure and patterns within the categorical variables, enabling more effective learning and prediction.

However, in our case, the results obtained are inferior with an $-\text{LLL}_\text{m}$ score of –6.71. This inferior performance may be because of the lack of high cardinality and ordinality in the data as it is being medical data. But the idea can be utilized in future when the sufficient amount of clinical demographic data can be provided by the medical community.

3) NETWORK CONFIGURATIONS

The spatial transformer module is regarded as a crucial component of our proposed model. We conducted experiments by integrating this module with various configurations into the model as illustrated in TABLE 7, and the corresponding outcomes were presented in the TABLE 8. The model variations with multiple streams and either one or two layers of the Spatial Transformer modules tend to exhibit lower values of Log-Likelihood Loss (LLL_m), indicating a superior performance in terms of likelihood estimation. Moreover, they also demonstrate lower values of Root Mean Square Error (RMSE), implying enhanced accuracy in predicting the target values. Among the different configurations, the ‘‘Multiple streams with two spatial transformer module in each’’ configuration has exhibited the most favorable performance, as evidenced by both LLL_m and RMSE metrics as given in the TABLE 8. We will now discuss the factors that may have contributed to this improvement in performance.

Spatial Transformer modules offer a superior alternative to traditional Convolutional layers with attention mechanisms. Typically, images contain areas of interest that could belong to unidentified categories and may exist in multiple occurrences [37]. Networks enabled with spatial transform module not only utilize spatial attention in localizing them, but also possess the unique ability to align them for optimal performance. This is achieved through rotational, scale, and translation invariance capabilities of ST module, allowing the model to automatically learn the locations of these regions within an image. As a result, the model can perform precise cropping and scale-normalization on these crucial areas,

TABLE 7. (a) First Spatial transformer module that comes in the first stage along each of the streams. The convolution layers learns the features. The fully connected (FC) layers learns the parameters required for taking the affine spatial transformation that includes scaling, rotation, and cropping. The 6 parameters obtained in the last FC layer are reshaped into (2, 3) vector. (b) First usual forward pass that comes in between two spatial transformer modules. This is basically meant for spatial feature learning, so we kept a fixed 3 × 3 kernel size in each convolution layer. (c) Second spatial transformer module. This module has only three convolution layers that are sufficient for lower input shape. But the FC layers for learning the parameters required for spatial transformation are exactly identical to that of first spatial transformer module. (d) Second usual forward pass which comes after the second spatial transformer module. In this the features are feed forwarded through two convolution followed by four FC layers.

(a)

Module	Layer	Kernel Size	Stride	Input Shape	Output Shape
Localization	Conv2D	7x7	2	(1,256,256)	(8,125,125)
	MaxPool2D	2x2	1	(8,125,125)	(8,62,62)
	Conv2D	5x5	2	(8,62,62)	(16,29,29)
	MaxPool2D	2x2	1	(16,29,29)	(16,14,14)
	Conv2D	3x3	2	(16,14,14)	(32,6,6)
	MaxPool2D	2x2	1	(32,6,6)	(32,3,3)
	Conv2D	1x1	1	(32,3,3)	(8,3,3)
Affine transformation	FC	-	-	72	36
	FC	-	-	36	6

(b)

Module	Layer	Kernel Size	Stride	Input Shape	Output Shape
Usual Forward Pass	Conv2D	3x3	1	(1,256,256)	(4,254,254)
	MaxPool2D	2x2	1	(4, 254,254)	(4,127,127)
	Conv2D	3x3	2	(4,127,127)	(8,125,125)
	MaxPool2D	2x2	1	(8,125,125)	(8,62,62)

(c)

Module	Layer	Kernel Size	Stride	Input Shape	Output Shape
Localization	Conv2D	7x7	2	(8,62,62)	(16,23,23)
	MaxPool2D	2x2	1	(16,23,23)	(16,11,11)
	Conv2D	5x5	1	(32,11,11)	(32,6,6)
	MaxPool2D	2x2	1	(32,6,6)	(32,3,3)
	Conv2D	1x1	1	(32,3,3)	(8,3,3)
Affine transformation	FC	-	-	72	36
	FC	-	-	36	6

(d)

Module	Layer	Kernel Size	Stride	Input Shape	Output Shape
Usual Forward Pass	Conv2D	3x3	1	(8,62,62)	(16,60,60)
	MaxPool2D	2x2	1	(16, 60,60)	(16,30,30)
	Conv2D	3x3	2	(16,30,30)	(32,14,14)
	MaxPool2D	2x2	1	(32,14,14)	(32,7,7)
	FC	-	-	1568	800
	FC	-	-	800	400
	FC	-	-	400	200
	FC	-	-	200	50

TABLE 8. Performance comparison of FibroRegNet employing different configurations of the spatial transformer module.

Model	Type of Regression	LLL_m	RMSE
Our proposed FibroRegNet model	Single stream with two STN layers	-6.67 ± 0.38	179.02±22.5
	Multiple streams with single STN layer in each	-6.65 ± 0.34	176.22±21.3
	Multiple streams with two STN layers in each	**-6.64 ± 0.42**	**172.35±16.2**

which proves highly beneficial in downstream tasks such as classification and segmentation. Given the potential occurrence of honeycomb cysts in the lungs due to IPF [12], the CT image must prioritize these specific areas over others. The rationale behind the enhanced performance of our proposed model, FibroRegNet, lies in its capacity to effectively employ visually significant indicators of clinical importance with the presence of multiple streams, each equipped with two spatial transform modules, facilitating the detection and differentiation of multiple occurrences of these particular regions. The accuracy and clinical relevance of FibroRegNet's decision-making capabilities under different modalities are evidenced by the outcomes presented in TABLE 4.

4) SELECTION OF HYPERPARAMETERS

As we dive into model training, there are a few key hyperparameters that require fine-tuning. Our model consists of two different spatial transformer modules, each seamlessly integrated into each of the three identical streams. In each stream, the two spatial transformer modules are cascaded in sequence having a standard forward pass blocks after each of them. The type & number of layers, kernel sizes, and strides of each these modules are as specified in the accompanying TABLE 7. The trainable parameters for the network were initialized according to the method [41]. All hidden layers within the network utilize ReLU as the nonlinear activation function, including the fully connected layers. And finally,

the ultimate regression layer remains linear. Utilizing the L1 loss function, our network is trained to minimize errors. In light of the multimodality data fused at various depths of our network, we considered a two-phase training approach. In the first phase, to effectively capture CT features, our network was exclusively trained for 40 epochs on CT slices with a batch size of 16. Later on, we progressively integrated demographic information and estimated lung volume into subsequent fully connected layers. For an additional 20 epochs, our network was trained using the best weights from the first phase as initialization - this time allowing changes to only the last three fully connected layers while freezing all others. We set learning rate at 0.01 initially and subsequently decreased it by a factor of 0.1 every 10 epochs and we used Adam as our optimizer throughout the entire training process.

V. DISCUSSION

We discuss here the factors that have contributed in the performance enhancement of the proposed method. We estimated the quadratic polynomial ridge regression coefficients of the Forced Vital Capacity (FVC) based on an empirical prior decision and used them as target variables to determine the trend of a patient. We utilized the encoded categorical non numerical data of the demographic information along with non-categorical numerical data as one set of features during the network training. We employed 85 percent of the slices in the preprocessing stage to calculate the volume and used it as another feature during neural network training. We also verified by incorporating other higher order statistics (mean, skew, and kurtosis) as obtained from the slices as features; however, their inclusion in our model yielded unsatisfactory results. Hence, we solely utilized the volume, which captures majority of information about the lung vital capacity.

Although we employed majority of the CT slices of patient in the estimation of lung volume, we employed only three slices for training our multi stream convolutional spatial transformer network to extract our third set of features called CT features. We introduced a multi stream spatial transformer modules in between convolutional layers to enhance the representation power of convolutional layers by enabling the network to pay attention on specific regions. Our extensive experiments substantiated that our proposed approach showed better performance than the recent models tested on the same dataset [24], [25], [29]. The reasons for adopting this setup are as follows.

- The dataset exhibited significant inconsistency in the number of slices amongst the patients, while the neighboring slices displayed minimal variation. Consequently, it proved difficult to determine a fixed number of slices that are optimal in achieving better generalizability for the model.
- Employing all the slices for learning CT features was also infeasible due to the fact that varying number of slices per patient results in substantial disparities. To address this challenge, we utilized a 2DCNN approach, as employing a 3DCNN with this limited data would be exceedingly difficult. This necessitated the utilization of the slices in a channel-wise manner. Additionally, we conducted experiments involving single and multiple streams in a channel-wise fashion, and the outcomes revealed that the model's performance enhanced while having number of streams up to three and beyond that there is no significant improvement is evident sometimes showing decline in the performance primarily due to overfitting.
- Learned representations of demographic information which can map the categorical and non-categorical data in to an embedding space was also not feasible as there are at max three categories. Thus requiring to use only the encoding of the demographic information in the preprocessing step itself.
- Quadratic polynomial ridge regression is more relevant than the linear regression as the trend of the FVC values is quadratic. This was empirically determined after carefully validating that this formulation is not overfitting the coefficients of the regression to the data using R-square and MSE values along with the visual inspection of the residual plots.

We have provided outcomes using two evaluation measures, namely the Laplace Log-Likelihood Score and the RMSE. We performed cross validation with K equal to 5 with 20 percent of the patients reserved for testing the model evaluation in each fold, and the average metrics were reported across the five test folds. Nevertheless, we firmly believe that our pipeline is significantly simpler. We assert the effectiveness of our proposed pipeline from two perspectives. Firstly, we managed to avoid the burdensome analysis of high-volume biomedical data by utilizing only three slices from each patient, which typically curtail the increase in latency. Secondly, the prior polynomial assumption facilitated better prediction results. Our model has 22.4 million parameters, 0.63G macs, and takes 0.81 seconds of inference time. It was observed that all the three are marginally higher compared to what the recent works have portrayed, which was a trade-off we made to attain a model that is both robust and generalizable.

Despite the commendable performance, a significant constraint of our proposed methodology was the variability of the prognosis for pulmonary fibrosis from patient to patient. This particular aspect restricted our capacity to forecast the FVC at each temporal instant. To address this, we employed the quadratic polynomial assumption as a means to provide regularization and prevent model from overfitting. As previously discussed, during the testing phase, only CT scan is provided for each patient along with a baseline FVC. That is nothing but employing the input of single modality. Furthermore, despite attempts to employ deeper architectures, FibroRegNet failed to yield improved performance, potentially due to the availability relatively limited sample size

dataset. Lastly, we maintained a fixed set of hyperparameters in all our experiments. However, by meticulously selecting the most optimal hyperparameters, the overall performance of each backbone can be further enhanced.

VI. CONCLUSION

We have put forth a pioneering convolutional spatial transformer based learning pipeline which takes multi modal data as input for prognosticating the outcomes of Idiopathic Pulmonary Fibrosis (IPF). We comprehensively incorporated both demographic information and CT scan in an end-to-end fashion. It is to be observed that each patient in the dataset has only one CT scan. The CT scans were recorded in irregular intervals of time within the study period. Our objective was to present a framework to the research community that can be employed on datasets of significantly larger volumes of CT scans and the correlated clinical parameters in the future. Given the significance of precise prediction of IPF progression in patients and the scarcity of IPF-based datasets, our proposed algorithm could illuminate novel approaches in constructing reliable algorithms for IPF prognosis. In addition to that, with the advent of medical image-text datasets pertaining to IPF, which are greatly inferior at moment, would pave a way in adapting the recent approaches such as contrastive language image pretraining [44], [45], [46] in the medical domain as well.

REFERENCES

[1] K. R. Flaherty, A. U. Wells, V. Cottin, A. Devaraj, S. L. F. Walsh, Y. Inoue, L. Richeldi, M. Kolb, K. Tetzlaff, S. Stowasser, C. Coeck, E. Clerisme-Beaty, B. Rosenstock, M. Quaresma, T. Haeufel, R.-G. Goeldner, R. Schlenker-Herceg, and K. K. Brown, "Nintedanib in progressive fibrosing interstitial lung diseases," *New England J. Med.*, vol. 381, no. 18, pp. 1718–1727, Oct. 2019, doi: 10.1056/nejmoa1908681.

[2] T. M. Maher, E. Bendstrup, L. Dron, J. Langley, G. Smith, J. M. Khalid, H. Patel, and M. Kreuter, "Global incidence and prevalence of idiopathic pulmonary fibrosis," *Respiratory Res.*, vol. 22, no. 1, Dec. 2021, Art. no. 197, doi: 10.1186/s12931-021-01791-z.

[3] R. S. Gupta, A. Koteci, A. Morgan, P. M. George, and J. K. Quint, "Incidence and prevalence of interstitial lung diseases worldwide: A systematic literature review," *BMJ Open Respiratory Res.*, vol. 10, no. 1, Jun. 2023, Art. no. e001291, doi: 10.1136/bmjresp-2022-001291.

[4] T. Refaee, Z. Salahuddin, A.-N. Frix, C. Yan, G. Wu, H. C. Woodruff, H. Gietema, P. Meunier, R. Louis, J. Guiot, and P. Lambin, "Diagnosis of idiopathic pulmonary fibrosis in high-resolution computed tomography scans using a combination of handcrafted radiomics and deep learning," *Frontiers Med.*, vol. 9, pp. 1–10, Jun. 2022, doi: 10.3389/fmed.2022.915243.

[5] H. Strongman, I. Kausar, and T. M. Maher, "Incidence, prevalence, and survival of patients with idiopathic pulmonary fibrosis in the U.K.," *Adv. Therapy*, vol. 35, no. 5, pp. 724–736, May 2018, doi: 10.1007/s12325-018-0693-1.

[6] X. Wu et al., "Computed tomographic biomarkers in idiopathic pulmonary Fibrosis. The future of quantitative analysis," *Amer. J. Respiratory Crit. Care Med.*, vol. 199, no. 1, pp. 12–21, Jan. 2019, doi: 10.1164/rccm.201803-0444pp.

[7] S. L. F. Walsh, S. M. Humphries, A. U. Wells, and K. K. Brown, "Imaging research in fibrotic lung disease; applying deep learning to unsolved problems," *Lancet Respiratory Med.*, vol. 8, no. 11, pp. 1144–1153, Nov. 2020, doi: 10.1016/s2213-2600(20)30003-5.

[8] S. L. F. Walsh, L. Calandriello, M. Silva, and N. Sverzellati, "Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: A case-cohort study," *Lancet Respiratory Med.*, vol. 6, no. 11, pp. 837–845, Nov. 2018.

[9] A. H. Syed, T. Khan, and S. A. Khan, "Deep transfer learning techniques-based automated classification and detection of pulmonary fibrosis from chest CT images," *Processes*, vol. 11, no. 2, p. 443, Feb. 2023, doi: 10.3390/pr11020443.

[10] A. Yadav, R. Saxena, A. Kumar, T. S. Walia, A. Zaguia, and S. M. M. Kamal, "FVC-NET: An automated diagnosis of pulmonary fibrosis progression prediction using honeycombing and deep learning," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, Jan. 2022, doi: 10.1155/2022/2832400.

[11] M. B. Gotway, M. M. Freemer, and T. E. King, "Challenges in pulmonary fibrosis middle dot 1: Use of high resolution CT scanning of the lung for the evaluation of patients with idiopathic interstitial pneumonias," *Thorax*, vol. 62, no. 6, pp. 546–553, Jun. 2007, doi: 10.1136/thx.2004.040022.

[12] B. Hochhegger, E. Marchiori, M. Zanon, A. S. Rubin, R. Fragomeni, S. Altmayer, C. R. R. Carvalho, and B. G. Baldi, "Imaging in idiopathic pulmonary fibrosis: Diagnosis and mimics," *Clinics*, vol. 74, p. 225, Jan. 2019, doi: 10.6061/clinics/2019/e225.

[13] ATS, ERS Joint Statement, "Idiopathic pulmonary fibrosis: Diagnosis and treatment," *Amer. J. Respiratory Crit. Care Med.*, vol. 161, no. 2, pp. 646–664, 2000, doi: 10.1164/ajrccm.161.2.ats3-00.

[14] A. Chen, R. A. Karwoski, D. S. Gierada, B. J. Bartholmai, and C. W. Koo, "Quantitative CT analysis of diffuse lung disease," *RadioGraphics*, vol. 40, no. 1, pp. 28–43, Jan. 2020, doi: 10.1148/rg.2020190099.

[15] A. Shahin, *OSIC Pulmonary Fibrosis Progression*. Mountain View, CA, UAA: Kaggle, 2020. [Online]. Available: https://kaggle.com/competitions/osic-pulmonary-fibrosis-progression

[16] B. Bartholmai, R. Karwoski, V. Zavaletta, R. Robb, and D. Holmes, "The lung tissue research consortium: An extensive open database containing histological, clinical, and radiological data to study chronic lung disease," *Insight J.*, 2006. [Online]. Available: http://hdl.handle.net/1926/221

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[18] W. Yu, H. Zhou, Y. Choi, J. G. Goldin, P. Teng, W. K. Wong, M. F. McNitt-Gray, M. S. Brown, and G. H. J. Kim, "Multi-scale, domain knowledge-guided attention + random forest: A two-stage deep learning-based multi-scale guided attention models to diagnose idiopathic pulmonary fibrosis from computed tomography images," *Med. Phys.*, vol. 50, no. 2, pp. 894–905, Feb. 2023, doi: 10.1002/mp.16053.

[19] T. Moua, A. S. Lee, and J. H. Ryu, "Comparing effectiveness of prognostic tests in idiopathic pulmonary fibrosis," *Expert Rev. Respiratory Med.*, vol. 13, no. 10, pp. 993–1004, Oct. 2019, doi: 10.1080/17476348.2019.1656069.

[20] Z. Wang, "Deep learning approach for auto-detecting idiopathic pulmonary fibrosis prediction," in *Proc. IEEE Int. Conf. Artif. Intell. Ind. Design (AIID)*, May 2021, pp. 283–290, doi: 10.1109/AIID51893.2021.9456590.

[21] K. Nezamabadi, Z. Naseri, H. A. Moghaddam, M. Modarresi, N. Pak, and M. Mahdizade, "Lung HRCT pattern classification for cystic fibrosis using convolutional neural network," *Signal, Image Video Process.*, vol. 13, no. 6, pp. 1225–1232, Sep. 2019, doi: 10.1007/s11760-019-01447-y.

[22] A. Taneja and A. Yadav, "Sky-net: A deep learning approach to predicting lung function decline in sufferers of idiopathic pulmonary fibrosis," in *Proc. 4th Int. Conf. Inf. Manage. Mach. Intell.* New York, NY, USA: Association for Computing Machinery, Dec. 2022, pp. 1–4, doi: 10.1145/3590837.3590883.

[23] S. Mandal, V. E. Balas, R. N. Shaw, and A. Ghosh, "Prediction analysis of idiopathic pulmonary fibrosis progression from OSIC dataset," in *Proc. IEEE Int. Conf. Comput., Power Commun. Technol. (GUCON)*, Oct. 2020, pp. 861–865, doi: 10.1109/GUCON48875.2020.9231239.

[24] Z. Al Nazi, F. R. Mashrur, M. A. Islam, and S. Saha, "Fibro-CoSANet: Pulmonary fibrosis prognosis prediction using a convolutional self attention network," *Phys. Med. Biol.*, vol. 66, no. 22, Nov. 2021, Art. no. 225013, doi: 10.1088/1361-6560/ac36a2.

[25] G. H. J. Kim, Y. Shi, W. Yu, and W. K. Wong, "A study design for statistical learning technique to predict radiological progression with an application of idiopathic pulmonary fibrosis using chest CT images," *Contemp. Clin. Trials*, vol. 104, May 2021, Art. no. 106333, doi: 10.1016/j.cct.2021.106333.

[26] R. Aoki, T. Iwasawa, T. Saka, T. Yamashiro, D. Utsunomiya, T. Misumi, T. Baba, and T. Ogura, "Effects of automatic deep-learning-based lung analysis on quantification of interstitial lung disease: Correlation with pulmonary function test results and prognosis," *Diagnostics*, vol. 12, no. 12, p. 3038, Dec. 2022. [Online]. Available: https://www.mdpi.com/2075-4418/12/12/3038

[27] X. Wu, C. Yin, X. Chen, Y. Zhang, Y. Su, J. Shi, D. Weng, X. Jiang, A. Zhang, W. Zhang, and H. Li, "Idiopathic pulmonary fibrosis mortality risk prediction based on artificial intelligence: The CTPF model," *Frontiers Pharmacol.*, vol. 13, pp. 1–12, Apr. 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphar.2022.878764

[28] A. Wong, J. Lu, A. Dorfman, P. McInnis, M. Famouri, D. Manary, J. R. H. Lee, and M. Lynch, "Fibrosis-net: A tailored deep convolutional neural network design for prediction of pulmonary fibrosis progression from chest CT images," 2021, *arXiv:2103.04008*.

[29] K. Kotowski, D. Kucharski, B. Machura, S. Adamski, B. G. Becker, A. Krason, L. Zarudzki, J. Tessier, and J. Nalepa, "Detecting liver cirrhosis in computed tomography scans using clinically-inspired and radiomic features," *Comput. Biol. Med.*, vol. 152, Jan. 2023, Art. no. 106378, doi: 10.1016/j.compbiomed.2022.106378.

[30] K. Du, Y. Zhu, R. Mao, Y. Qu, B. Cui, Y. Ma, X. Zhang, and Z. Chen, "Medium-long term prognosis prediction for idiopathic pulmonary fibrosis patients based on quantitative analysis of fibrotic lung volume," *Respiratory Res.*, vol. 23, no. 1, pp. 1–12, Dec. 2022, doi: 10.1186/s12931-022-02276-3.

[31] D. Kawahara, T. Masuda, R. Nishioka, M. Namba, N. Imano, K. Yamaguchi, S. Sakamoto, Y. Horimasu, S. Miyamoto, T. Nakashima, H. Iwamoto, S. Ohshimo, K. Fujitaka, H. Hamada, N. Hattori, and Y. Nagata, "Prediction model for patient prognosis in idiopathic pulmonary fibrosis using hybrid radiomics analysis," *Res. Diagnostic Interventional Imag.*, vol. 4, Dec. 2022, Art. no. 100017, doi: 10.1016/j.redii.2022.100017.

[32] A. Khan, Z. Rauf, A. R. Khan, S. Rathore, S. H. Khan, N. S. Shah, U. Farooq, H. Asif, A. Asif, U. Zahoora, R. U. Khalil, S. Qamar, U. H. Asif, F. B. Khan, A. Majid, and J. Gwak, "A recent survey of vision transformers for medical image segmentation," 2023, *arXiv:2312.00634*.

[33] G. Zhang, C. Zheng, J. He, and S. Yi, "PCT: Pyramid convolutional transformer for parotid gland tumor segmentation in ultrasound images," *Biomed. Signal Process. Control*, vol. 81, Mar. 2023, Art. no. 104498, doi: 10.1016/j.bspc.2022.104498.

[34] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *J. Big Data*, vol. 7, no. 1, Dec. 2020, Art. no. 28, doi: 10.1186/s40537-020-00305-w.

[35] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *Eur. Radiol. Experim.*, vol. 4, no. 1, Dec. 2020, Art. no. 50, doi: 10.1186/s41747-020-00173-2.

[36] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," 2015, *arXiv:1506.02025*.

[37] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.

[38] E. T. Thomas, M. Guppy, S. E. Straus, K. J. L. Bell, and P. Glasziou, "Rate of normal lung function decline in ageing adults: A systematic review of prospective cohort studies," *BMJ Open*, vol. 9, no. 6, Jun. 2019, Art. no. e028150, doi: 10.1136/bmjopen-2018-028150.

[39] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," 2016, *arXiv:1604.06737*.

[40] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, vol. 9. PMLR, 2010, pp. 249–256. [Online]. Available: http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021, *arXiv:2103.00020*.

[42] A. Poulou, M. Poulos, and M. Panas, "A regression-based machine learning approach for the prediction of lung function decline," in *Proc. 12th Int. Conf. Dependable Syst., Services Technol. (DESSERT)*, Dec. 2022, pp. 1–5, doi: 10.1109/DESSERT58054.2022.10018624.

[43] R. A. Shehab, K. A. Apurba, M. Ahsanuzzaman, and T. Rahman, "Accurate prediction of pulmonary fibrosis progression using EfficientNet and quantile regression: A high performing approach," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Sep. 2023, pp. 1–6, doi: 10.1109/tensymp55890.2023.10223673.

[44] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," 2022, *arXiv:2210.10163*.

[45] S. Subramanian, L. L. Wang, B. Bogin, S. Mehta, M. van Zuylen, S. Parasa, S. Singh, M. Gardner, and H. Hajishirzi, "MedICaT: A dataset of medical images, captions, and textual references," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2020, pp. 2112–2120, doi: 10.18653/v1/2020.findings-emnlp.191.

[46] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang, and M. Zhou, "CodeBERT: A pre-trained model for programming and natural languages," 2020, *arXiv:2002.08155*.

PARDHASARADHI MITTAPALLI received the B.Tech. degree from Jawaharlal Nehru Technological University Hyderabad, Andhra Pradesh, India, and the M.Tech. degree in visual information processing and embedded systems from the Indian Institute of Technology Kharagpur. He is currently pursuing the Ph.D. degree with Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India. So far, he has published three research articles in peer-reviewed Scopus indexed journals and one Scopus indexed conference paper.

V. THANIKAISELVAN (Member, IEEE) received the B.E. degree in electronics and communication engineering from Bharathidasan University, Trichy, in 2002, the M.Tech. degree in advanced communication systems from SASTRA University, Thanjavur, in 2006, and the Ph.D. degree in image steganography for information security from Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India, in 2014. He is currently a Professor with the Department of Communication Engineering, School of Electronics Engineering, VIT. He has produced seven Ph.D. candidates. He is guiding four Ph.D. candidates in the areas of information security and digital image processing. So far, he has published more than 100 research articles in peer-reviewed Scopus indexed journals and Scopus indexed conference papers. He is a reviewer of Elsevier, Wiley, and Springer journals.