ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Integrated ML-Based Strategy Identifies Drug Repurposing for Idiopathic Pulmonary Fibrosis

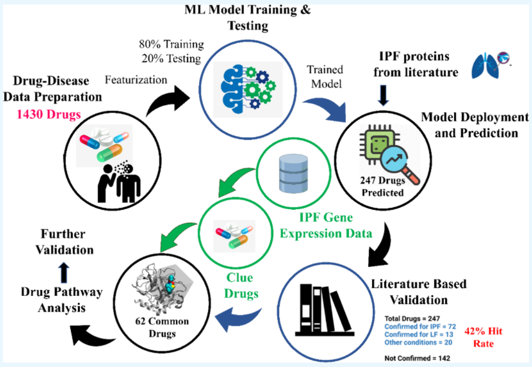
Faheem Ahmed, Anupama Samantasinghar, Myung Ae Bae, and Kyung Hyun Choi*

Cite This: *ACS Omega* 2024, 9, 29870−29883

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

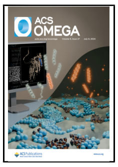
ABSTRACT: Idiopathic pulmonary fibrosis (IPF) affects an estimated global population of around 3 million individuals. IPF is a medical condition with an unknown cause characterized by the formation of scar tissue in the lungs, leading to progressive respiratory disease. Currently, there are only two FDA-approved small molecule drugs specifically for the treatment of IPF and this has created a demand for the rapid development of drugs for IPF treatment. Moreover, denovo drug development is time and cost-intensive with less than a 10% success rate. Drug repurposing currently is the most feasible option for rapidly making the drugs to market for a rare and sporadic disease. Normally, the repurposing of drugs begins with a screening of FDA-approved drugs using computational tools, which results in a low hit rate. Here, an integrated machine learning-based drug repurposing strategy is developed to significantly reduce the false positive outcomes by introducing the predock machine-learning-based predictions followed by literature and GSEA-assisted validation and drug pathway prediction. The developed strategy is deployed to 1480 FDA-approved drugs and to drugs currently in a clinical trial for IPF to screen them against "TGFB1", "TGFB2", "PDGFR-a", "SMAD-2/3", "FGF-2", and more proteins resulting in 247 total and 27 potentially repurposable drugs. The literature and GSEA validation suggested that 72 of 247 (29.14%) drugs have been tried for IPF, 13 of 247 (5.2%) drugs have already been used for lung fibrosis, and 20 of 247 (8%) drugs have been tested for other fibrotic conditions such as cystic fibrosis and renal fibrosis. Pathway prediction of the remaining 142 drugs was carried out resulting in 118 distinct pathways. Furthermore, the analysis revealed that 29 of 118 pathways were directly or indirectly involved in IPF and 11 of 29 pathways were directly involved. Moreover, 15 potential drug combinations are suggested for showing a strong synergistic effect in IPF. The drug repurposing strategy reported here will be useful for rapidly developing drugs for treating IPF and other related conditions.



1. INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is categorized as a rare, sporadic ailment. IPF impacts an estimated global population of around 3 million individuals.[1] The condition predominantly affects individuals aged 50 and above, with a higher prevalence observed among males compared to females. As per the information provided by the National Institutes of Health (NIH),[2] roughly 100,000 individuals in the United States are affected by IPF. Annually, there are approximately 30,000−40,000 newly diagnosed cases.[3] On a global scale, IPF impacts approximately 13−20 individuals out of every 100,000 people.[1] The progression of IPF can vary and is difficult to predict, leading to a gradual and irreversible decline in lung function in individuals with this condition. The prognosis for those diagnosed with IPF varies, but on average, the median survival time after diagnosis is approximately 2−3 years.[4] The disease can be influenced by a sudden deterioration in lung function, termed an acute exacerbation, which frequently results in mortality within a few months. Recognizing the severity of the disease and the urgency of treatment, early detection and treatment of IPF can also improve outcomes and reduce the risk of complications. However, to date, there are only two drugs Nintedanib and Pirfenidone approved for IPF.[5] A greater number of effective treatments are required to be developed to provide timely treatment to the patients. Drug discovery is a lengthy and expensive process that can take up to 15 years and cost billions of dollars.[6−10] Despite this, many drugs fail to gain regulatory approval due to safety or efficacy concerns. Drug repurposing (DR), also known as drug repositioning, offers an alternative approach to drug development by identifying new therapeutic uses for existing drugs. This strategy involves testing drugs that have already been approved or are in late-stage clinical trials for one indication to see if they could be effective for other diseases. Compared to traditional drug discovery, DR is a more efficient and cost-effective approach that has the potential to accelerate the delivery of new treatments to patients.[11,12] By

Received: April 20, 2024
Revised: May 30, 2024
Accepted: June 12, 2024
Published: June 27, 2024



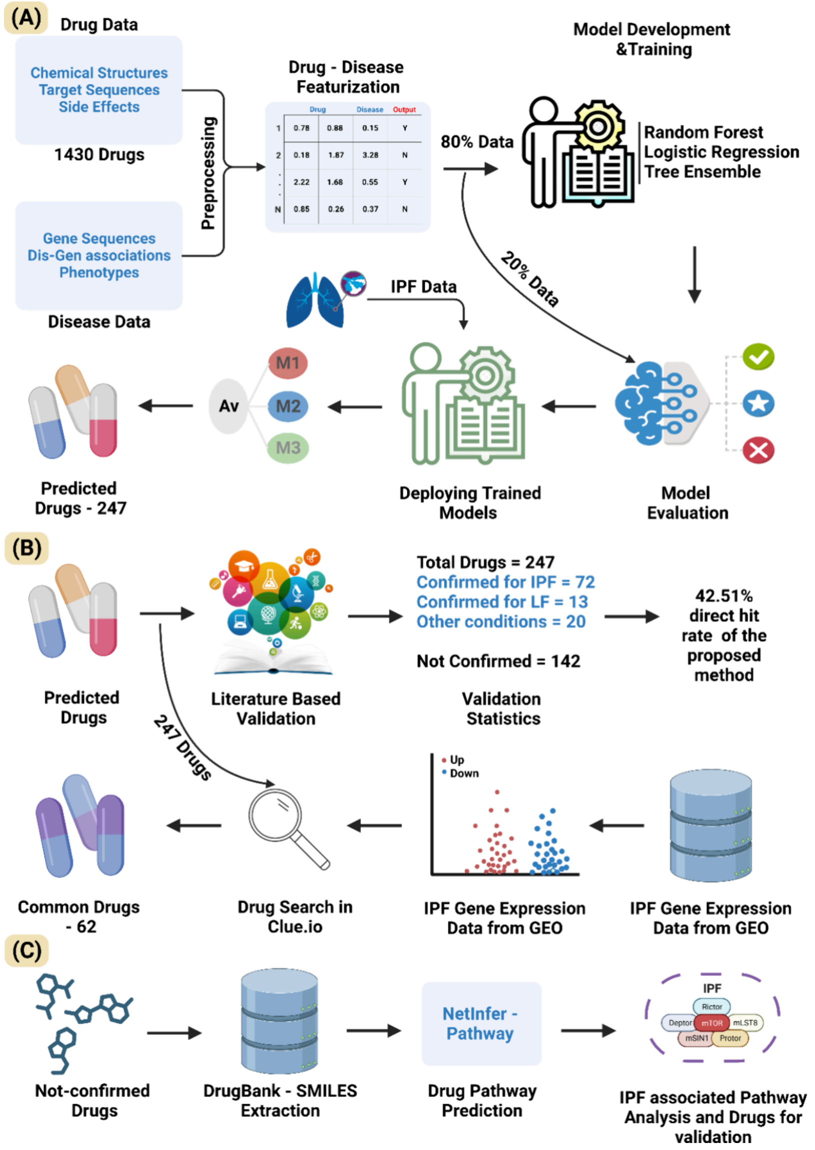


Figure 1. Drug Repurposing framework overview. (A) Process of data collection, preprocessing, and model development followed by deployment on IPF data to get the drug predictions. (B) Validation of the predicted drugs using literature and GSEA, and finally (C) depicts the drug-pathway association prediction for candidate drug prioritization and drug combination.

leveraging existing knowledge and clinical data, DR can save time and resources, and ultimately lead to the development of new treatments for a variety of diseases.[13]

Many approaches such as machine learning (ML),[14,15] virtual screening,[16] and network-science[17] have been used in DR for rare and sporadic diseases such as IPF and also in viral diseases to identify potential new uses for existing drugs. It is done by analyzing large amounts of data, such as gene expression, drug interactions, and clinical outcomes.[18,19] For example, a study published in[20] used ML to identify existing drugs repurposable for the treatment of Zika virus infection by analyzing gene expression profiles from infected human cells. Similarly, another study done in[21] performed DR through an in-silico analysis, as a result, BI2536, a specific inhibitor of polo-like kinase (PLK) 1/2, was chosen as a potential candidate for the treatment of pulmonary fibrosis. Additionally, there are drugs that have been repurposed for IPF. These drugs include nifuroxazide, niclosamide, dabigatran, and proton pump inhibitors (PPIs). A study done in[22] investigated the repurposing of nifuroxazide, an antidiarrheal drug, for the treatment of pulmonary fibrosis. Using ELISA, an in vitro toolkit, the study found that nifuroxazide was able to ameliorate pulmonary fibrosis in an animal model by blocking the TGF-β/Smad pathway and decreasing the expression of phosphorylated Stat3. Similarly, Niclosamide,[23] an anthelmintic drug, has been reported to reverse fibrosis in the skin and lungs of mice with systemic sclerosis and pulmonary fibrosis. An in vivo method is used for validation of the repurposed drug. In another study, the FDA adverse event reporting system and JMDC Inc. insurance claims were analyzed to predict the dabigatran drug as repurposable for IPF. Predictions were later validated by clinical big data.[24] Few-shot learning, ML, and molecular docking (MD) found that active inhibitors against IPF included Herbacetin, Morusin, Swertiamarin, Vicenin-2, and Vitexin.[25] Combinations of different approaches can be used in DR for Fibrotic diseases to predict the binding of a small molecule to a protein target such as done in.[25] Additionally, a repurposing study[26] identified seven biological pathways that are implicated in all nine fibrotic

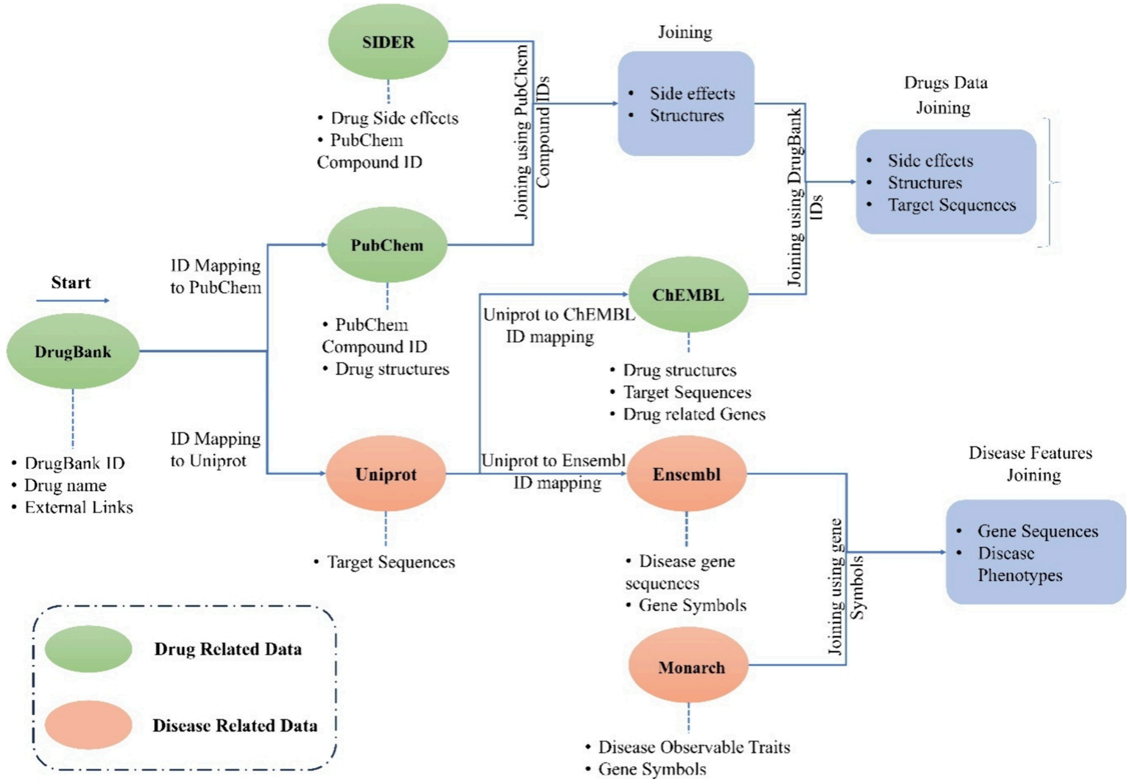


Figure 2. Drug and disease related data extraction and mapping processes followed in this study.

diseases, along with pathways that are specific to certain diseases among them. Captopril and ibuprofen were suggested as anti-IPF drugs, along with nafcillin and memantine drugs recommended for further validation. Thalidomide[27] was also suggested to have the possibility of targeting inflammation for the treatment of IPF. In addition to ML and MD, systems biology can also be used in DR for fibrotic diseases by analyzing the complex interactions between associated proteins and disease host factors and identifying new drug targets. One such study[28] used network-enabled DR for pulmonary fibrosis. Through the measurement of proximity between the drug targets and the signature within the interactome, we have identified numerous potential candidates and furnished a ranked drug list based on their closeness.[29]

However, ML, MD, and systems biology approaches suffer from limitations when individually used for DR. One of the main drawbacks of ML in DR is the potential for overfitting, where the algorithm learns patterns that are specific to the training data but may not generalize well to new data.[30] This can lead to false positive predictions and hinder the identification of promising drug candidates. Similarly, MD relies on accurate structural information for both the small molecule and the protein target, which can be a limitation when information is not available.[31] Additionally, MD does not account for the complex dynamics of protein−ligand interactions, which can affect the drug efficacy and toxicity. Finally, Systems biology approaches require large amounts of data and complex computational models, which can be time-consuming and resource-intensive. Additionally, the biological complexity of the systems being studied can make it difficult to identify key targets and pathways for DR.[32] A possible solution to the mentioned limitations posed by these individual techniques is the combination of ML and MD that has overcome some of these limitations by integrating the strengths of both approaches. ML can help identify potential drug candidates from large data sets, and MD can be used to predict the binding of those candidates to specific protein targets.[33] This can improve the accuracy and efficiency of DR efforts, as demonstrated by several recent studies. One such effort is SperoPredictor[14] which used a combination of ML and MD to identify potential drug candidates for COVID-19.

In this study, a DR strategy that combines multiple ML models has been developed to minimize false positive outcomes (Figure 1). Here, an integrated ML-based DR strategy is developed to significantly reduce the false positive outcomes by introducing ML-based predictions (Figure 1A) followed by literature-assisted validation (Figure 1B) and drug pathway prediction (Figure 1C). The developed strategy is deployed to 1480 FDA-approved drugs and to drugs currently in a clinical trial for IPF to screen them against 'TGFB1', 'TGFB2', “PDGFR-a”, 'SMAD-2/3', 'FGF-2', and more proteins resulting in 247 total and 27 potentially repurposable drugs. Literature and GSEA validation suggested 72 of 247 (29.14%) have been tried for IPF, 13 of 247 (5.2%) drugs have already been used for lung fibrosis, and 20 of 247 (8%) drugs have been tested for other fibrotic conditions such as cystic fibrosis and renal fibrosis. Pathway prediction of the remaining 142 drugs was carried out resulting in 118 distinct pathways. Furthermore, analysis revealed that 29 of 118 pathways were directly or indirectly involved in IPF and 11 of 29 pathways were directly involved. Moreover, 15 potential drug combinations are suggested for showing a strong synergistic effect in IPF. The drug repurposing strategy reported here will be useful for rapidly developing drugs for treating IPF and other related conditions.

2. MATERIALS AND METHODS

2.1. Data Collection and Preparation. The data set prepared contains drug and disease-related feature information

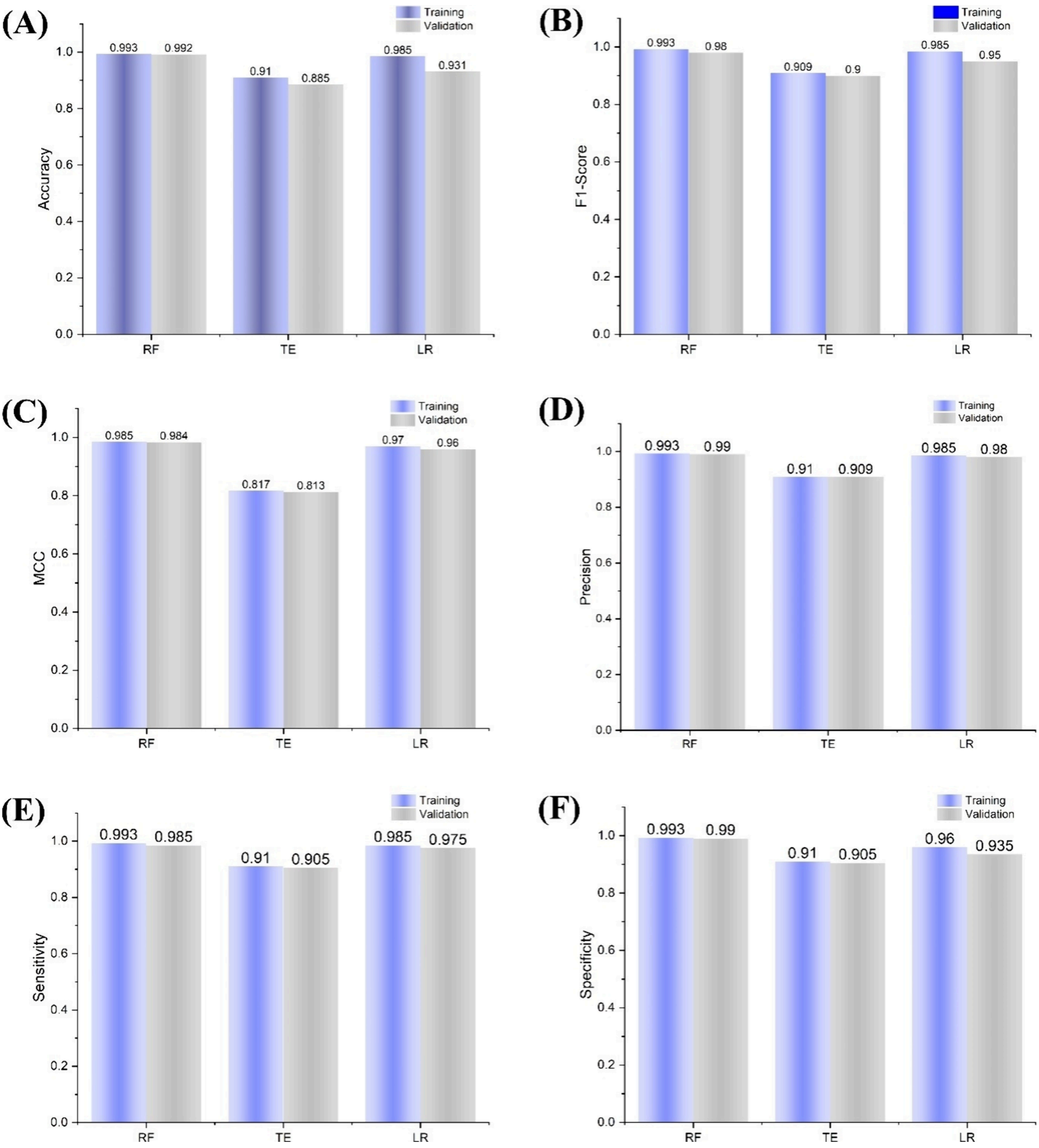


Figure 3. Performance Statistics of the Random Forest (RF), Tree Ensemble (TE), and Logistic Regression (LR). (A) Accuracies of the Models where RF performs as best model, (B) F1-score (C) MCC, (D) precision, (E) sensitivity and (F) specificity. Among all the statistics RF performs best followed by LR, and TE.

obtained from various sources. Chemical structures for the drugs have been used as the main feature of the drugs. Numerous research works have utilized chemical structures presented in SMILES format.[34] SMILES is a simplified line notation utilized to portray the configuration of chemical entities. The SMILES were obtained from DrugBank,[35] PubChem,[36] and ChEMBL[37] between the years 2022 and 2023 (Figure 2). All the drug-related databases are shown in green boxes along with their mapping process in Figure 2. SMILES strings were transformed into a fixed-length vector representation using extended connectivity fingerprints (ECFP) generation methods.[38] These methods encode the chemical structure into a series of binary values representing the presence or absence of specific substructures or molecular properties. Drug target sequences were considered as another feature of the drugs (Figure 2).[39] Each drug targets one or multiple proteins to produce a

https://doi.org/10.1021/acsomega.4c03796

Table 1. Performance Atatistics of the ML Models

	accuracy		F1-score		MCC		precision		sensitivity		pecificity	
	train:	valid:	train:	valid:	train:	valid:	train:	valid:	train:	valid:	train:	valid:
RF	0.993	0.992	0.993	0.98	0.985	0.984	0.993	0.99	0.993	0.985	0.993	0.99
TE	0.91	0.885	0.909	0.9	0.817	0.813	0.91	0.909	0.91	0.905	0.91	0.905
LR	0.985	0.931	0.985	0.95	0.97	0.96	0.985	0.98	0.985	0.975	0.96	0.935

Table 2. IPF Associated Target Gene Information

	Uniprot ID	name	ensemble ID	genes	DSI	DPI
1	P01137	transforming growth factor beta-1 proprotein	ENSG00000105329	TGFB1	0.287	0.962
2	P61812	transforming growth factor beta-2 proprotein	ENSG00000092969	TGFB2	0.433	0.885
3	P16234	″	platelet-derived growth factor receptor alpha″	PDGFR-a	0.415	0.808
4	P09038	fibroblast growth factor 2	ENSG00000138685	FGF-2	0.383	0.923
5	P84022	mothers against decapentaplegic homologue 3	ENSG00000166949	Smad3	0.415	0.923
6	P31751	RAC-beta serine/threonine-protein kinase	ENSG00000105221	Akt-2	0.474	0.769
7	P62736	smooth muscle actinn-alpha	ENSG00000107796	a-SMA	0.54	0.769
8	P02452	″	collagen alpha-1(I) chain″	ENSG00000108821	0.43	0.846
9	Q8QHL3	vascular endothelial growth factor receptor 1		VEGFR	0.419	0.846
10	P08253	72 kDa type IV collagenase	ENSG00000087245	MMP-2	0.333	0.923
11	P29279	connective tissue growth factor	ENSG00000118523	CTGF	0.399	0.846
12	Q5VWK5	interleukin-23 receptor	ENSG00000162594	IL-23	0.415	0.846
13	P05231	interleukin-6	ENSG00000136244	IL-6	0.248	0.962
14	P05121	plasminogen activator inhibitor 1	ENSG00000106366	PAI-1	0.359	0.885
15	P35225	interleukin-13	ENSG00000169194	IL-13	0.846	0.846
16	P24394	interleukin-4 receptor subunit alpha	ENSG00000077238	IL-4Ra	0.474	0.846

therapeutic effect. Here, during the academic years 2023−2024, drug-target sequences were obtained from Uniprot and DrugBank. Protein sequences are represented as a series of amino acids and were preprocessed for nonstandard amino acids or missing data before being transformed into ML-ready format using one-hot encoding. For the Drug Side Effects SIDER database was used to download data on 1,430 FDA-approved drugs.[40] Side effects were mapped to drug structures, protein sequences, and gene data using PubChem IDs, which are provided by SIDER. The process of data mapping is inspired and adopted from our previous study done in ref 14 however, the current study is short of the features.[14]

Additionally, disease information on gene-disease associations (GDA) was gathered from DisGeNET, a comprehensive database containing over 400,000 GDAs from various sources. Gene sequences were then extracted from the Ensemble database using the UniProt IDs and gene symbols from DisGeNET.[41] Data cleaning involving removing or imputing errors, missing data, or redundant information was performed followed by k-mer-encoding-based transformation of sequences into a numerical representation. It captures the common features including.

physicochemical properties, of amino acids, structural properties, and sequence-based properties, such as sequence motifs. Disease-related phenotypes provide important insights into understanding disease mechanisms to light up the road to competitive therapeutic development, due to this disease phenotypes have been utilized for drug discovery and development for long. The Monarch database (Figure 2; purple boxes) was used to obtain information on 10,881 human diseases, along with 8662 phenotypes associated with these diseases. Natural language processing (NLP) preprocessing techniques for text features including tokenization, stemming, and stop word removal were used. Finally, vectorization and embedding were used for processing text features to convert text data into numerical vectors that can be used in ML models. Additionally, for drug and disease-related features, dimensionality reduction techniques, such as principal component analysis (PCA) are used to reduce the number of features while retaining 90% of the important information.

2.2. Model Development, Training, Testing, and Validation (Deep Learning).

After preparing and transforming the data, we considered the data on drug- and disease-related features (DTIs) to be positive samples, while the unknown interactions were treated as negative samples by randomizing the positive samples (Figure 3) (Table 1). To ensure a balanced data set, negative samples were up-sampled. This approach was adopted to avoid bias Such methods may be biased due to undetected interactions between targeted drugs. To prevent duplication between positive and negative samples, we checked that no pairs from the positive samples matched exactly with negative samples. Next three different ML Models including Random Forest (RF),[42] Logistic Regression (LR),[43] and Tree Ensemble (TE)[44] were developed using default parameters initially.

Data were split into the training (80%) and testing data (20%). Initially, the models were trained and tested separately, followed by testing in a similar methodology. The training and testing accuracies were recorded. Additionally, a 10-fold cross-validation was performed to further validate the models using the 20% test data. All the parameters are given in Table S1, and the performance statistics of the models are given in Table 1. Once models are trained, all of the models are deployed separately on the IPF data followed by the prediction fusion of the top-performing models. Only the LR and RF were deployed owning to their higher performance statistics.[45] The predictions of the models were validated from the literature on the individual and combination levels. By leveraging the power of ML, we were able to uncover hidden patterns and relationships within the data sets that would have been challenging to identify

Table 3. Drugs Already Used under IPF and Other Related Conditions

	DrugBank ID	name	ensemble gene ID (ensemble)	disease genes
idiopathic pulmonary fibrosis (IPF)				
1	DB00175	pravastatin	ENSG00000105221	Akt-2
2	DB00184	nicotine	ENSG00000105221	Akt-2
3	DB00201	caffeine	ENSG00000105221	Akt-2
4	DB00248	cabergoline	ENSG00000105221	Akt-2
5	DB00321	amitriptyline	ENSG00000105221	Akt-2
6	DB00333	methadone	ENSG00000105221	Akt-2
7	DB00341	cetirizine	ENSG00000105221	Akt-2
8	DB00370	mirtazapine	ENSG00000105221	Akt-2
9	DB00437	allopurinol	ENSG00000106366	PAI-1
10	DB00458	imipramine	ENSG00000105221	Akt-2
11	DB00468	quinine	ENSG00000105221	Akt-2
12	DB00471	montelukast	ENSG00000087245	MMP-2
13	DB00480	lenalidomide	ENSG00000106366	PAI-1
14	DB00493	cefotaxime	ENSG00000105221	Akt-2
15	DB00502	haloperidol	ENSG00000105221	Akt-2
16	DB00514	dextromethorphan	ENSG00000105221	Akt-2
17	DB00517	anisotropine methyl bromide	ENSG00000105221	Akt-2
18	DB00548	azelaic acid	ENSG00000105221	Akt-2
19	DB00557	hydroxyzine	ENSG00000105221	Akt-2
20	DB00562	benzthiazide	ENSG00000105221	Akt-2
21	DB00565	cisatracurium	ENSG00000105221	Akt-2
22	DB00570	vinblastine	ENSG00000106366	PAI-1
23	DB00590	doxazosin	ENSG00000105221	Akt-2
24	DB00598	labetalol	ENSG00000105221	Akt-2
25	DB00641	simvastatin	ENSG00000105221	Akt-2
26	DB00661	verapamil	ENSG00000105221	Akt-2
27	DB00678	losartan	ENSG00000106366	PAI-1
28	DB00808	indapamide	ENSG00000105221	Akt-2
29	DB00843	donepezil	ENSG00000105221	Akt-2
30	DB00844	nalbuphine	ENSG00000105221	Akt-2
31	DB00850	perphenazine	ENSG00000105221	Akt-2
32	DB00852	pseudoephedrine	ENSG00000105221	Akt-2
33	DB00920	ketotifen	ENSG00000105221	Akt-2
34	DB00924	cyclobenzaprine	ENSG00000105221	Akt-2
35	DB00972	azelastine	ENSG00000105221	Akt-2
36	DB00975	dipyridamole	ENSG00000105221	Akt-2
37	DB00988	dopamine	ENSG00000105221	Akt-2
38	DB01012	cinacalcet	ENSG00000105221	Akt-2
39	DB01039	fenofibrate	ENSG00000105221	Akt-2
40	DB01056	tocainide	ENSG00000105221	Akt-2
41	DB01095	fluvastatin	ENSG00000105221	Akt-2
42	DB01098	rosuvastatin	ENSG00000105221	Akt-2
43	DB01115	nifedipine		VEGFR
44	DB01148	flavoxate	ENSG00000105221	Akt-2
45	DB01198	zopiclone	ENSG00000105221	Akt-2
46	DB01216	finasteride	ENSG00000105221	Akt-2
47	DB01222	budesonide	ENSG00000105221	Akt-2
48	DB01223	aminophylline	ENSG00000106366	PAI-1
49	DB01303	oxtriphylline	ENSG00000106366	PAI-1
50	DB01406	danazol	ENSG00000105221	Akt-2
51	DB01409	tiotropium	ENSG00000105221	Akt-2
52	DB01656	roflumilast	ENSG00000105221	Akt-2
53	DB04843	mepenzolate	ENSG00000105221	Akt-2
54	DB05039	indacaterol	ENSG00000105221	Akt-2
55	DB05154	pretomanid	ENSG00000106366	PAI-1
56	DB05294	vandetanib	ENSG00000106366	PAI-1
57	DB05812	abiraterone	ENSG00000105221	Akt-2
58	DB05990	obeticholic acid	ENSG00000106366	PAI-1
59	DB06410	doxercalciferol	ENSG00000105221	Akt-2
60	DB06616	bosutinib	ENSG00000087245	MMP-2

Table 3. continued

	DrugBank ID	name	ensemble gene ID (ensemble)	disease genes
idiopathic pulmonary fibrosis (IPF)				
61	DB06663	pasireotide	ENSG00000105221	Akt-2
62	DB06772	cabazitaxel	ENSG00000106366	PAI-1
63	DB06800	methylnaltrexone	ENSG00000105221	Akt-2
64	DB08860	pitavastatin	ENSG00000105221	Akt-2
65	DB08896	regorafenib	ENSG00000087245	MMP-2
66	DB08910	pomalidomide	ENSG00000106366	PAI-1
67	DB08916	afatinib	ENSG00000105221	Akt-2
68	DB09053	ibrutinib	ENSG00000105221	Akt-2
69	DB09079	nintedanib	ENSG00000105221	Akt-2
70	DB11217	arbutin	ENSG00000106366	PAI-1
71	DB11619	gestrinone	ENSG00000105221	Akt-2
cystic fibrosis				
72	DB00198	oseltamivir	ENSG00000105221	Akt-2
73	DB00462	methscopolamine bromide	ENSG00000105221	Akt-2
74	DB00487	pefloxacin	ENSG00000105221	Akt-2
75	DB01061	azlocillin	ENSG00000106366	PAI-1
76	DB01165	ofloxacin	ENSG00000105221	Akt-2
77	DB01409	tiotropium	ENSG00000105221	Akt-2
78	DB08897	aclidinium	ENSG00000105221	Akt-2
79	DB09076	umeclidinium	ENSG00000105221	Akt-2
lung fibrosis				
80	DB00474	methohexital	ENSG00000106366	PAI-1
81	DB00700	eplerenone	ENSG00000105221	Akt-2
82	DB00758	clopidogrel	ENSG00000105221	Akt-2
83	DB00843	donepezil	ENSG00000105221	Akt-2
84	DB01240	epoprostenol	ENSG00000105221	Akt-2
85	DB01587	ketazolam	ENSG00000105221	MMP-2
86	DB01591	solifenacin	ENSG00000105221	MMP-2
87	DB02659	cholic acid	ENSG00000105221	TGFB-1/2
88	DB04854	febuxostat	ENSG00000105221	Akt-2
89	DB08881	vemurafenib	ENSG00000106366	PAI-1
90	DB09477	enalaprilat	ENSG00000105221	Akt-2
91	DB14490	ferrous ascorbate	ENSG00000106366	PAI-1
renal fibrosis				
92	DB06212	tolvaptan	ENSG00000105221	Akt-2
93	DB01267	paliperidone	ENSG00000105221	Akt-2
other fibrotic conditions				
92	DB00247	methysergide	ENSG00000105221	Akt-2
93	DB00381	amlodipine	ENSG00000105221	Akt-2
94	DB00421	spironolactone	ENSG00000087245	MMP-2
95	DB00425	zolpidem	ENSG00000105221	Akt-2
96	DB00795	sulfasalazine	ENSG00000105221	Akt-2
97	DB00915	amantadine	ENSG00000105221	Akt-2
98	DB00973	ezetimibe	ENSG00000105221	Akt-2
99	DB02300	calcipotriol	ENSG00000105221	Akt-2
100	DB05271	rotigotine	ENSG00000105221	Akt-2
101	DB06209	prasugrel	ENSG00000105221	Akt-2
102	DB06210	eltrombopag	ENSG00000105221	Akt-2
103	DB08867	ulipristal	ENSG00000105221	Akt-2

through traditional methods alone. Results demonstrated that the combination of models or prediction fusion performs better with an improved hit rate. Moreover, to counter the biases in literature validation, all prediction results were searched which were expected to support the predictions positively and negatively.[46] Additionally, to address the inheritance biases in literature validation, an additional step was taken to enhance the credibility and robustness of the predictions. For each prediction made, an extensive search was conducted to identify relevant sources of information that were anticipated to both support and challenge the predicted outcomes. This meticulous approach ensured a comprehensive and balanced evaluation of the predictions, encompassing a wide range of perspectives and potential outcomes.

2.3. Proteins Involved in IPF. To take predictions by deploying models, we required IPF-related proteins involved in the development of IPF. In this connection, a literature survey of the published articles, patents, and databases was carried out,

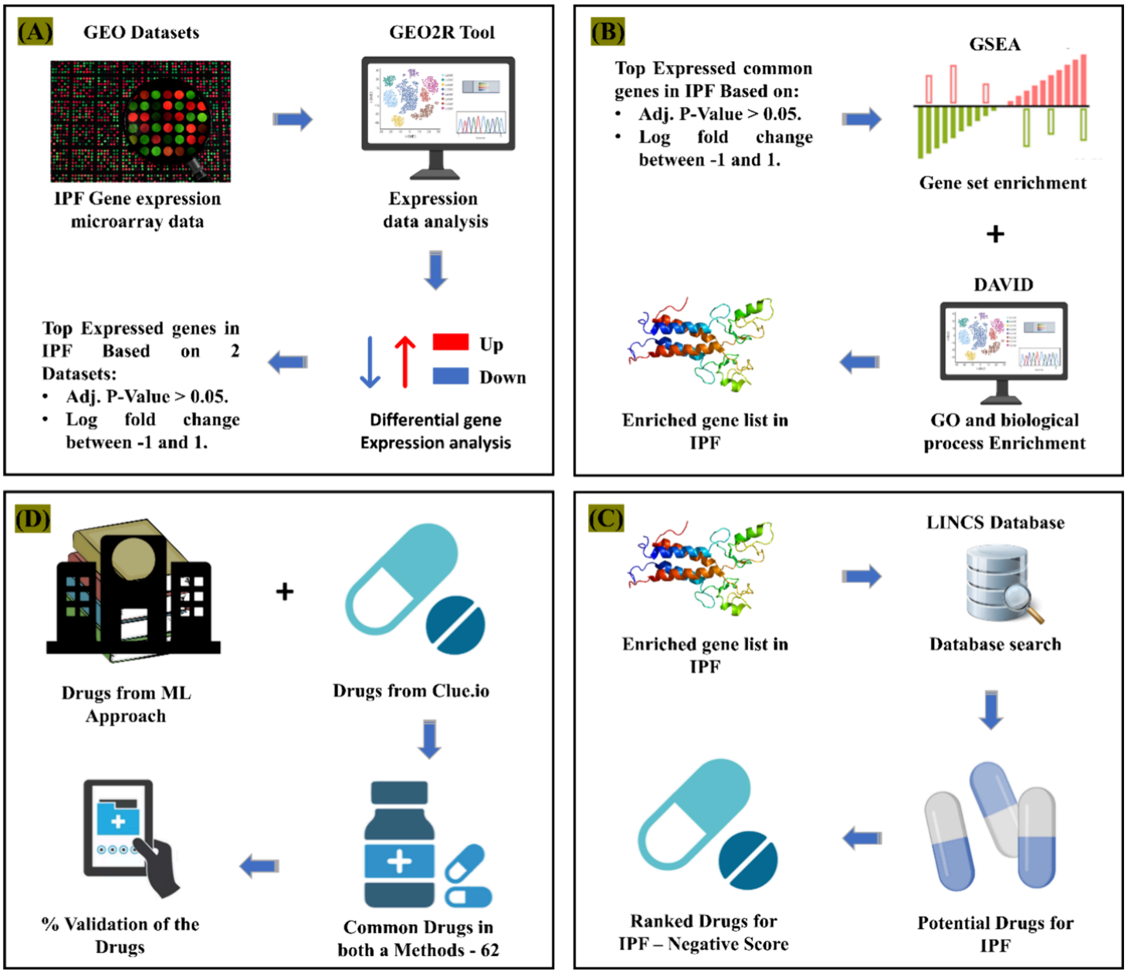

(A)
GEO Datasets
GEO2R Tool
IPF Gene expression microarray data
Expression data analysis
Top Expressed genes in IPF Based on 2 Datasets:
• Adj. P-Value > 0.05.
• Log fold change between -1 and 1.
Up
Down
Differential gene Expression analysis
(B)
Top Expressed common genes in IPF Based on:
• Adj. P-Value > 0.05.
• Log fold change between -1 and 1.
GSEA
Gene set enrichment
+
DAVID
Enriched gene list in IPF
GO and biological process Enrichment
(D)
Drugs from ML Approach
+
Drugs from Clue.io
% Validation of the Drugs
Common Drugs in both a Methods - 62
(C)
Enriched gene list in IPF
LINCS Database
Database search
Ranked Drugs for IPF – Negative Score
Potential Drugs for IPF

Figure 4. Clue-based validation. (A) Data set extraction and analysis, (B) enrichment analysis of the genes, (C) database searches for the drugs using the enriched gene list from GEO, and (D) literature-based validation of the ML-predicted and clue-generated drugs.

and we found 16 related proteins (Table 2). The data for these proteins was prepared and preprocessing steps like main data were performed. Once the data were prepared, models were deployed to get the prediction of the drugs repurposable for IPF, and downstream validation using literature sources was carried out initially, as mentioned in Section 2.2, followed by GSEA-based validation.[47−49]

2.4. External Validation. Here, we conducted an additional analysis using a combination of Gene Expression Omnibus (GEO) data sets and LINCS tools for the purpose of drug repurposing in the context of IPF. The utilization of both GEO data sets and machine learning models allowed us to achieve external validation of our study and validate the effectiveness of our ML approach.[11] To initiate the process, we downloaded IPF-related data sets (GSE2052 and GSE24206) from the GEO database, which provided us with gene expression data associated with IPF. These data sets contained valuable information regarding the gene expression profiles of IPF-affected samples as well as control samples, enabling us to identify potential targets and pathways relevant to IPF.[50]

Clue.io, a comprehensive computational platform, integrates large-scale gene expression data with a vast library of compounds and their known biological effects.[11] By leveraging the vast resources and algorithms available within Clue.io, we were able to explore gene expression data in the context of known drug effects, enabling us to identify novel therapeutic options for IPF-related diseases. One of the key strengths of our study lies in the overlap observed between the drugs identified through our ML models and those derived from the GEO approach.[51] This overlapping subset of repurposed drugs served as a strong validation of our ML approach, demonstrating its reliability and accuracy. The fact that the ML models and GEO approach produced similar results indicates the robustness and consistency of our findings.

2.5. Pathway Prediction. For drugs that were not confirmed through the literature, they were further collected separately. The SMILES data for all of the drugs were collected and prepared as shown in Table S2. Later, the drugs along with SMILES were uploaded on the Netinfer[52] server and pathway predictions were obtained. All of the parameters for drug-pathway association were predicted with default parameters. Predicted results for drug-pathway association are given in Table S2.

Moreover, from the predicted pathways, IPF-associated distinct pathways were separated and drugs associated with these pathways among the unused (in literature) were prioritized. Finally, from the prioritized drugs for preclinical validation, drug combinations based on pathway association were predicted.

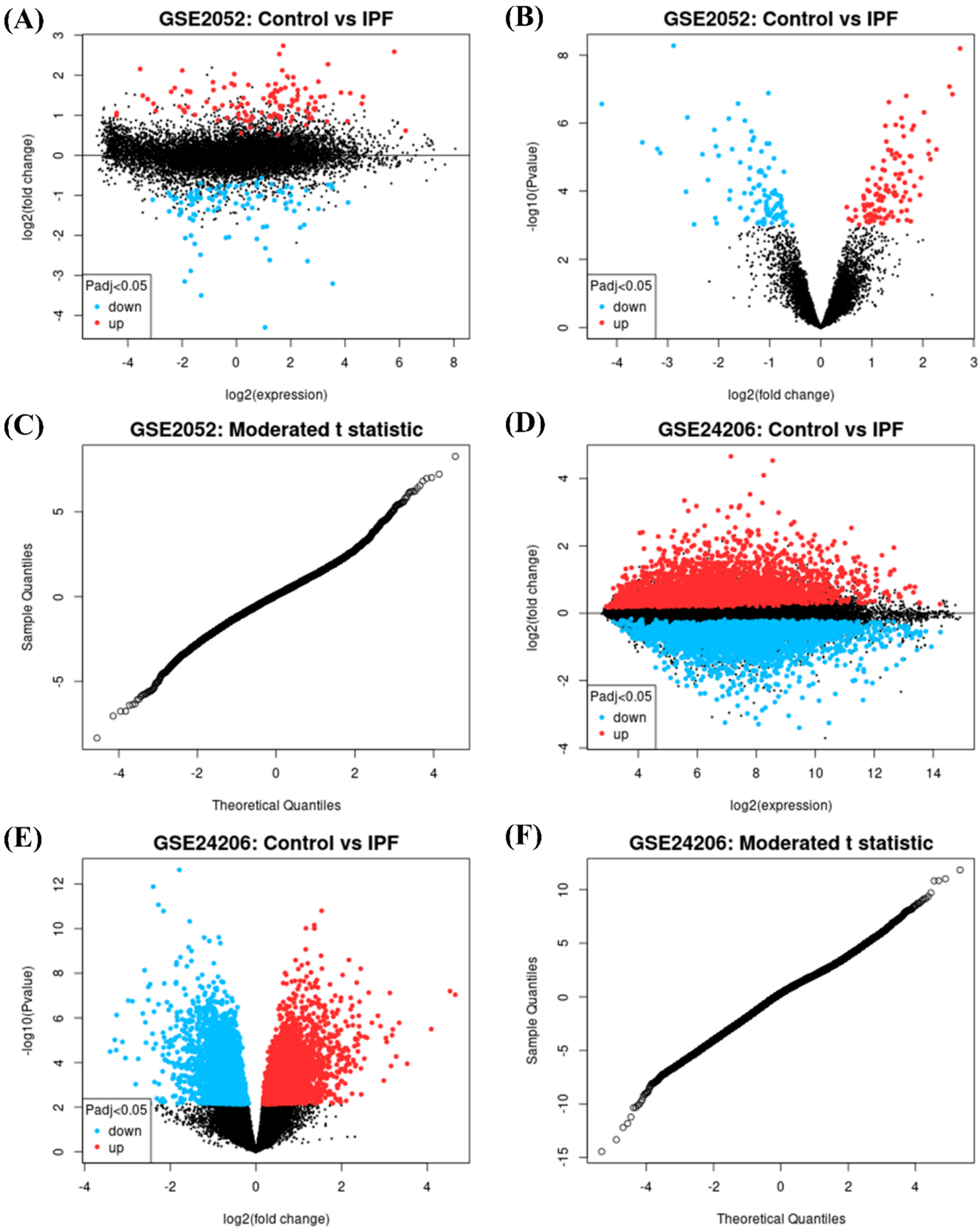


Figure 5. (A−C) Control vs IPF analyzed expression through GEO2R for GSE2052 are given; (D, E) for data set GSE24206 analyzed expression are given. Moreover, Red colored dots depict the overexpression gene expressions, blue colored genes represent the under expressed gene, and black colored dots show the genes with insignificant fold change. (F) Moderated t-statistic score shows the linear relationship between sample quantiles and theoretical quantiles.

3. RESULTS

3.1. Machine Learning Model Training and Validation.

After data collection and preparation of 1480 FDA-approved drugs (Table S2), it was split into training and testing data sets. Training data sets account for 80% while testing data sets account for 20% of the total data. Next, three different ML Models, including RF, LR, and TE were developed and trained with default parameters initially. The training accuracies for RF, LR, and TE were 99.3, 98.5, and 91%, respectively. Similarly, the testing accuracies for RF, LR, and TE were 99.2, 88.5, and 93.1%, respectively, as shown in Figure 3A. Other performance statistics such as f1-score, MCC, precision, sensitivity, and

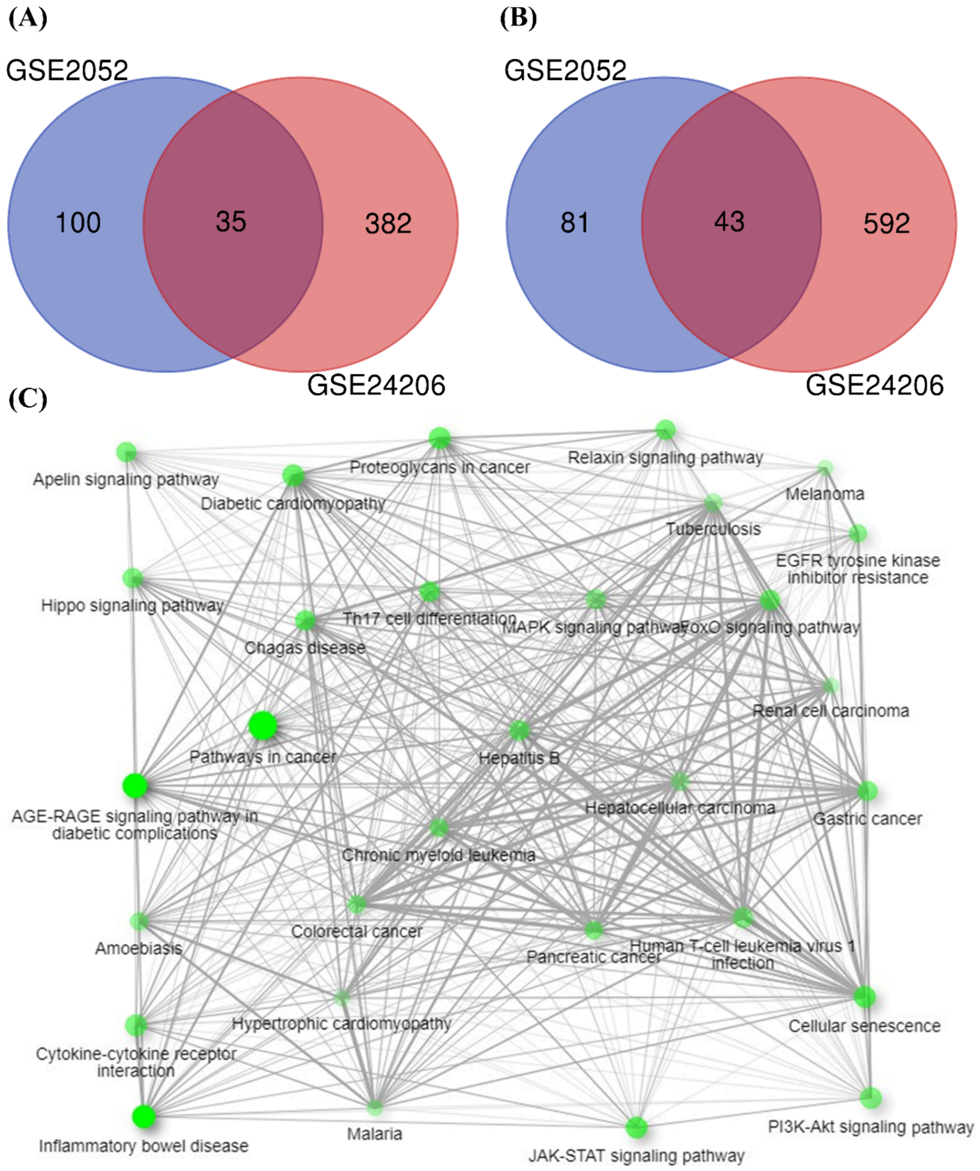


Figure 6. (A) common upregulated genes from both data sets, (B) common downregulated genes from both data sets, and (C) map of pathways and processing involving the common and downregulated genes.

specificity are mentioned in Figure 3B–F, respectively. Additionally, these performance statistics are summarized in Table 1 as well. Based on the given performance parameters, RF and LR were selected for further application in drug repurposing for IPF.

3.2. Data Preparation for IPF. A literature and database search was carried out to find the IPF-related key proteins such as ‘TGFB1′, ‘TGFB2′, “PDGFR-a”, ‘SMAD-2/3′, “FGF-2” and more (Table 2). Later further information related to these proteins was extracted from databases mentioned in the methods. The data was prepared and processed (Figure S1) to get the prediction of repurposable drugs for IPF. Additionally, for the IPF protein targets, STRING enrichment was performed,[53] and protein–protein interaction network[54] results are saved as Figure S2, and the gene coexpression information is also given in Figure S3.

3.3. Predictions from Trained Models. The trained RF and LR regression models were deployed to get the prediction for the IPF-associated protein targets. Both models, RF and LR were deployed, outputs were averaged, and the top 247 drugs (Table S10) against the protein’s targets were shortlisted (Table S2). The detailed procedure for the deployment of the models is given in Section 2.2, and data preparation is given in Figure S4.

3.4. Literature and GSEA-Based Validation of the Prediction. For the predicted drugs, a literature survey was carried out to check if they have already been approved, used, or tried for IPF or not.[55] In the case of positive results, the drug was labeled as validated and not validated otherwise. Among the total 247 drugs, as per the literature, 105 drugs (42.5%) were confirmed from the literature to have been tried for IPF and other fibrotic conditions such as lung and cystic fibrosis (Table 3). Moreover, Clue was used for further validation, as shown in Figure 4. Here, two data sets for IPF-associated disease information were obtained from GEO (GSE2052 and GSE24206) and were analyzed for up and downregulated genes (Figure 4A) with p-value <0.05 and LogFC values in the range of <−1 and >+1 were selected. GSE2052 contains 26 samples which involve normal histology lung tissue samples and IPF lung explant (Table S3).[56] Similarly, GSE24206 consists of a total of 23 samples containing Healthy donor biological replicates, early IPF surgical biopsy upper lobe replicates, and advanced IPF explant upper and lower lobe replicates (Table S4).[57] The data sets were analyzed using the GEO2R tool for differential expression analysis (Figure 5) and the process was repeated for both data sets. Control vs IPF analyzed expressions through GEO2R for GSE2052 are given in Figure 5A−C and those for data set GSE24206 are given in Figure 5D−F. Red-colored dots depict the overexpression of gene expressions, blue-colored genes represent the under-expressed gene, and black-colored dots show the genes with insignificant fold change (Figure 5).

Finally, downloaded data sets were analyzed manually in Excel resulting in 35 positively regulated (Figure 6A) and 43 negatively regulated genes (Figure 6B) common in both data sets (Table S5). Additionally, DAVID, GSEA, GO, biological process enrichment, and pathway enrichment[58] along with ShinyGO 0.77 were performed using a Web server (Figure 4B).[59] Results of pathway and process enrichment from ShinyGo are given in Figure 6C. It shows pathways and processes directly and indirectly involved in IPF, such as. JAK-STAT, PI3K-Akt, EGFR tyrosine kinase, and Relaxin signaling pathways (Figure 6C).[60] Later up-regulated genes were put in the STRING database, the network was exported to Cytoscape, and hub genes were found by setting and sorting the network by degree topology. The degree was set between 1 and 7 for the downregulated network and between 1 and 7 for the upregulated network. The role of these genes/proteins in IPF was confirmed through the literature. Finally, the analyzed genes were used to perform clue for drug repurposing and as a result, the top 1088 drugs with the lowest negative score (Table S6) were selected as candidate drugs (Figure 4C) followed by a literature survey (Figure 4D). By comparing these drugs with ML-based predicted drugs, we found that a total of 62 drugs are overlapped (Figure 7). Of these overlapping drugs, most of them were already confirmed through the literature, and for the remaining drug-pathway association, prediction through NetInfer was carried out to further confirm them and prioritize for validation in vitro.

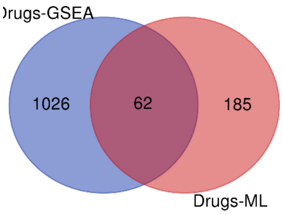


Figure 7. Drugs common between Clue Web server and proposed framework.

3.5. Pathway Prediction and Drug Prioritization. In this study, we employed the NetInfer Web server[52] to explore drug-pathway associations and identify potential repurposable drugs for IPF. The data set comprised 142 drugs (Table S2), and upon analysis, we successfully predicted 118 distinct pathways associated with these compounds (Table S7). Among these pathways, 28 were found to be directly or indirectly involved in the pathogenesis of IPF (Table S8), with 11 core pathways showing direct involvement (Table S9). These findings provide valuable insights into the intricate molecular interactions between drugs and pathways in the context of the IPF. To further prioritize the drug candidates, we considered their associations with the core IPF pathways. Consequently, we identified 21 drugs that demonstrated significant potential for repurposing as therapeutics for IPF (Table 4). The selection of these repurposable drugs was guided by the premise that their targeted pathways have direct relevance to IPF pathogenesis.[61] These promising candidates warrant thorough preclinical validation to assess their safety and efficacy before advancing to clinical trials.

Table 4. Potential Drugs along with Pathways to Be Preclinically Validated

	compound name	pathway ID	description	score
1	escitalopram	hsa04151	PI3K-Akt signaling pathway	0.00637411
2	pirenzepine	hsa04668	TNF signaling pathway	0.00625015
3	disopyramide	hsa04151	PI3K-Akt signaling pathway	0.00624992
4	quetiapine	hsa04151	PI3K-Akt signaling pathway	0.00606185
5	brompheniramine	hsa04151	PI3K-Akt signaling pathway	0.0060605
6	cyproheptadine	hsa04151	PI3K-Akt signaling pathway	0.0060375
7	topiramate	hsa04151	PI3K-Akt signaling pathway	0.0058144
8	chenodeoxycholic acid	hsa04151	PI3K-Akt signaling pathway	0.00570901
9	fluphenazine	hsa04066	HIF-1 signaling pathway	0.00539549
10	trimipramine	hsa04066	HIF-1 signaling pathway	0.00538127
11	quinethazone	hsa04151	PI3K-Akt signaling pathway	0.00625234
12	pirenzepine	hsa04151	PI3K-Akt signaling pathway	0.00622134
13	hyoscyamine	hsa04151	PI3K-Akt signaling pathway	0.00599395
14	lorazepam	hsa04151	PI3K-Akt signaling pathway	0.00596583
15	glimepiride	hsa04151	PI3K-Akt signaling pathway	0.00578804
16	propafenone	hsa04151	PI3K-Akt signaling pathway	0.00577599
17	dimenhydrinate	hsa04151	PI3K-Akt signaling pathway	0.00559847
18	iloperidone	hsa04010	MAPK signaling pathway	0.00525688
19	meclizine	hsa04151	PI3K-Akt signaling pathway	0.00666584
20	flumazenil	hsa04151	PI3K-Akt signaling pathway	0.00591095
21	brompheniramine	hsa04668	TNF signaling pathway	0.00590764
22	dobutamine	hsa04010	MAPK signaling pathway	0.00580683
23	trimipramine	hsa04010	MAPK signaling pathway	0.00527779
24	meclizine	hsa04010	MAPK signaling pathway	0.00664799
25	stiripentol	hsa04151	PI3K-Akt signaling pathway	0.00590532
26	cyproheptadine	hsa04010	MAPK signaling pathway	0.00550832
27	disopyramide	hsa04010	MAPK signaling pathway	0.00531159
28	trimipramine	hsa04668	TNF signaling pathway	0.00519002
29	quetiapine	hsa04010	MAPK signaling pathway	0.00510318

Furthermore, we conducted a drug combination analysis based on pathway overlapping to explore potential synergistic effects. By analyzing the interactions between the 118 pathways associated with the drugs' correlation using their pathways, we identified 15 high-end drug combinations that have the potential to produce enhanced therapeutic outcomes compared to individual drug treatments (Table 5).[62] The synergy score is calculated using the total number of pathways being targeted by combinations.[63,64] Complete information about the drug combinations and pathways targeted is given in Table S11. The concept of pathway overlapping serves as a robust strategy to design novel and efficient drug combinations for complex diseases such as IPF. To ensure the translation of our findings into clinically relevant treatments, we emphasize the importance of rigorous preclinical validation for all prioritized drugs and drug combinations. Preclinical studies are essential to establish the safety profiles, optimal dosages, and efficacy of the identified candidates, laying the groundwork for their eventual evaluation in clinical trials.[65] Overall, our study provides a comprehensive exploration of drug-pathway associations in the context of IPF, identifying promising repurposable drugs and potential drug combinations.[66] These results contribute to the growing field of drug repurposing and offer a potential avenue for the development of more effective and targeted therapies for IPF patients. Further research and validation will be instrumental in advancing these discoveries toward clinical application and ultimately improving the lives of individuals affected by this devastating lung disease.

Table 5. Drug Combination from the Predicted Repurposable Drugs

	drug 1	drug 2	synergy score
1	trimipramine	fluphenazine	23
2	cyproheptadine	fluphenazine	17
3	fluphenazine	iloperidone	17
4	iloperidone	fluphenazine	17
5	fluphenazine	glimepiride	15
6	cyproheptadine	dobutamine	16
7	chenodeoxycholic acid	fluphenazine	15
8	disopyramide	fluphenazine	12
9	disopyramide	trimipramine	12
10	fluphenazine	chenodeoxycholic acid	12
11	brompheniramine	fluphenazine	12
12	brompheniramine	trimipramine	12
13	cyproheptadine	disopyramide	12
14	disopyramide	quetiapine	12
15	dobutamine	disopyramide	12

4. DISCUSSION

In this study, we aimed to identify potential drug candidates for IPF through drug repurposing using ML models and external validation. We collected drug- and disease-related features from various databases and used ML models such as RF and LR for prediction. The trained models were deployed to predict drug–protein interactions relevant to IPF, and the results were validated through a literature review and GEO data sets. Additionally, we employed pathway prediction to prioritize drugs based on their associations with IPF-related pathways.[67,68] The use of ML models allowed us to efficiently analyze complex data sets and uncover hidden patterns and relationships between drugs and proteins, which might have been challenging to identify using traditional methods alone. The combination of RF and LR showed superior performance in predicting drug–protein interactions, ensuring a robust basis for drug repurposing. One limitation of our study is that it relies on the accuracy and completeness of the data obtained from various databases.[69] Although we performed data cleaning and preprocessing steps, there might still be errors or missing information that could affect the accuracy of our predictions. Moreover, while our ML approach demonstrated reliability and consistency, it is essential to acknowledge that predictions are inherently based on associations and might not always reflect direct causation. Future directions for this research include enhancing the data set with more recent and comprehensive data to further improve the accuracy of predictions. Additionally, incorporating additional features, such as drug–drug interactions and drug-target binding affinities, could provide more comprehensive insights into drug repurposing for IPF.[70,71] Validating the predicted drug candidates in preclinical models and clinical trials is crucial to assess their safety, efficacy, and potential for use as IPF treatments.

5. CONCLUSIONS

In conclusion, our study demonstrates that DR is a promising approach for identifying potential therapies for IPF. By employing machine learning models, we successfully predicted drug–protein interactions relevant to IPF, and these predictions were validated using literature sources and GEO data sets. The ML models exhibited robust performance in identifying potential drug candidates, and the overlap between ML-predicted and GEO-derived drugs further reinforced the reliability of our approach. Moreover, pathway prediction allowed us to prioritize drug candidates based on their associations with IPF-related pathways, enhancing the potential for successful repurposing. By focusing on pathways directly implicated in IPF pathogenesis, we identified a subset of drugs with a higher likelihood of efficacy in treating the disease. Overall, this study highlights the importance of drug repurposing as a more efficient and economical approach compared with de novo drug development for IPF. The identified drug candidates offer promising avenues for further preclinical validation and potentially for advancing to clinical trials. Further research and validation of these candidates are essential to bring effective and targeted therapies to IPF patients, ultimately improving their quality of life and outcomes. By leveraging the power of machine learning and pathway analysis, our study contributes to the growing field of drug repurposing and offers hope for the development of more effective treatments for IPF and other challenging diseases.

■ ASSOCIATED CONTENT

Data Availability Statement

Supporting Information including figures and tables are provided as "Supplementary Figures.docx" and Tables from Tables S1–S11. Code will be provided on demand.

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c03796.

Images and tables of the supplementary results generated during this study; four figures from the KNIME Platform for the data analysis followed by figures from the STRING Database; additionally, there are 11 tables; data generated from analysis to provide a detailed overview of the study to the interested researchers (PDF)

Table mentioning the hyper parameters of the classification models used in our study (PDF)
Drugs SMILES (XLSX)
Up- and down-regulated genes from GSE2052 (XLSX)
Up- and down-regulated genes from GSE24206 (XLSX)
Up- and down-regulated genes from Both (XLSX)
Table showing rank, score, type, ID, name, description, and target details (XLSX)
Distinct drug-associated pathways (XLSX)
Directly or indirectly involved pathways (XLSX)
Directly involved pathways (XLSX)
Complete list of predicted drugs (PDF)
Drug pathways (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Kyung Hyun Choi − *Department of Mechatronics Engineering, Jeju National University, Jeju 63243, Republic of Korea;* orcid.org/0000-0002-4503-2458; Email: amm@jejunu.ac.kr

Authors

Faheem Ahmed − *Department of Mechatronics Engineering, Jeju National University, Jeju 63243, Republic of Korea;* orcid.org/0000-0001-8908-2599

Anupama Samantasinghar − *Department of Mechatronics Engineering, Jeju National University, Jeju 63243, Republic of Korea*

Myung Ae Bae − *Therapeutics and Biotechnology Division, Korea Research Institute of Chemical Technology, Daejeon 34114, Korea*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsomega.4c03796

Author Contributions

F.A.: Conceptualization, Formal Analysis, original draft, Writing - review and editing, and images. A.S.: review and editing. M.A.B.: review and editing. K.H.C.: Conceptualization, review and editing, Resources, Supervision, Funding Acquisition.

Funding

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea government (MSIT) (NRF-2022R1A2C2004771).

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Martinez, F. J.; et al. Idiopathic pulmonary fibrosis. *Nat. Rev. Dis. Prim.* **2017**, *3*, 17074.

(2) Onishchenko, D.; et al. Screening for idiopathic pulmonary fibrosis using comorbidity signatures in electronic health records. *Nat. Med.* **2022**, *28* (10), 2107−2116.

(3) Kugeler, K. J.; Schwartz, A. M.; Delorey, M. J.; Mead, P. S.; Hinckley, A. F. Estimating the Frequency of Lyme Disease Diagnoses, United States, 2010−2018. *Emerg. Infect. Dis.* **2021**, *27* (2), 616−619.

(4) Du, K.; et al. Medium-long term prognosis prediction for idiopathic pulmonary fibrosis patients based on quantitative analysis of fibrotic lung volume. *Respir. Res.* **2022**, *23* (1), 372.

(5) Khan, M. A.; et al. Nintedanib and pirfenidone for idiopathic pulmonary fibrosis in King Abdulaziz Medical City, Riyadh: Real-life data. *Ann. Thorac. Med.* **2023**, *18* (1), 45−51.

(6) Ahmed, F.; Samantasinghar, A.; Soomro, A.; Kim, S.; Choi, K. A systematic review of computational approaches to understand cancer biology for informed drug repurposing. *J. Biomed. Inform.* **2023**, *142*, No. 104373.

(7) Ahmed, F.; et al. Drug Repurposing for viral cancers: A paradigm of machine learning, deep learning, and Virtual screening-based approaches. *J. Med. Virol.* **2023**, *95*, No. e28693.

(8) Samantasinghar, A.; et al. A comprehensive review of key factors affecting the efficacy of antibody drug conjugate. *Biomed. Pharmacother.* **2023**, *161*, No. 114408.

(9) Ahmed, F.; et al. Robust ultrasensitive stretchable sensor for wearable and high-end robotics applications. *J. Mater. Sci. Mater. Electron.* **2022**, *33* (35), 26447−26463.

(10) Asif, A.; et al. Microphysiological system with continuous analysis of albumin for hepatotoxicity modeling and drug screening. *J. Ind. Eng. Chem.* **2021**, *98*, 318−326.

(11) Ahmed, F.; et al. Drug repurposing in psoriasis, performed by reversal of disease-associated gene expression profiles. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 6097−6107.

(12) Subbiah, V. The next generation of evidence-based medicine. *Nat. Med.* **2023**, *29* (1), 49−58.

(13) Haleem, A.; Javaid, M.; Singh, R. P.; Suman, R. Telemedicine for healthcare: Capabilities, features, barriers, and applications. *Sensors Int.* **2021**, *2*, No. 100117.

(14) Ahmed, F.; et al. SperoPredictor: An Integrated Machine Learning and Molecular Docking-Based Drug Repurposing Framework With Use Case of COVID-19. *Front. Public Health* **2022**, *10*, No. 902123.

(15) Ahmed, F.; et al. Multi-material Bio-inspired Soft Octopus Robot for Underwater Synchronous Swimming. *J. Bionic Eng.* **2022**, *19*, 1229−1241.

(16) Samantasinghar, A.; Ahmed, F.; Rahim, C. S. A.; Kim, K. H.; Kim, S.; Choi, K. H. Artificial intelligence-assisted repurposing of lubiprostone alleviates tubulointerstitial fibrosis. *Transl. Res.* **2023**, *262*, 75.

(17) Ahmed, F.; Yang, Y. J.; Samantasinghar, A.; Kim, Y. W.; Ko, J. B.; Choi, K. H. Network-based drug repurposing for HPV-associated cervical cancer. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 5186−5200.

(18) Yang, F.; et al. Machine Learning Applications in Drug Repurposing. *Interdiscip. Sci. Comput. Life Sci.* **2022**, *14* (1), 15−21.

(19) Zhao, K.; So, H.-C. Drug Repositioning for Schizophrenia and Depression/Anxiety Disorders: A Machine Learning Approach Leveraging Expression Data. *IEEE J. Biomed. Heal. Informatics* **2019**, *23* (3), 1304−1315.

(20) Zhang, N.; et al. Identification of novel anti-ZIKV drugs from viral-infection temporal gene expression profiles. *Emerg. Microbes Infect.* **2023**, *12* (1), No. 2174777.

(21) Imakura, T.; et al. A polo-like kinase inhibitor identified by computational repositioning attenuates pulmonary fibrosis. *Respir. Res.* **2023**, *24* (1), 148.

(22) Gan, C.; et al. Nifuroxazide ameliorates pulmonary fibrosis by blocking myofibroblast genesis: a drug repurposing study. *Respir. Res.* **2022**, *23* (1), 32.

(23) Boyapally, R.; Pulivendala, G.; Bale, S.; Godugu, C. Niclosamide alleviates pulmonary fibrosis in vitro and in vivo by attenuation of epithelial-to-mesenchymal transition, matrix proteins & Wnt/*β*-catenin signaling: A drug repurposing study. *Life Sci.* **2019**, *220*, 8−20.

(24) Siswanto, S.; et al. Drug Repurposing Prediction and Validation From Clinical Big Data for the Effective Treatment of Interstitial Lung Disease. *Front. Pharmacol.* **2021**, *12*, No. 635293.

(25) Chang, J.; Zou, S.; Xu, S.; Xiao, Y.; Zhu, D. Screening of Inhibitors Against Idiopathic Pulmonary Fibrosis: Few-Shot Machine Learning and Molecule Docking Based Drug Repurposing. *Curr. Comput. Aided. Drug Des.* **2023**, *20*, 134−144.

(26) Karatzas, E.; Kakouri, A. C.; Kolios, G.; Delis, A.; Spyrou, G. M. Fibrotic expression profile analysis reveals repurposed drugs with potential anti-fibrotic mode of action. *PLoS One* **2021**, *16* (4), No. e0249687.

(27) Dsouza, N. N.; Alampady, V.; Baby, K.; Maity, S.; Byregowda, B. H.; Nayak, Y. Thalidomide interaction with inflammation in idiopathic pulmonary fibrosis. *Inflammopharmacology* **2023**, *31* (3), 1167−1182.

(28) Li, A.; Chen, J.-Y.; Hsu, C.-L.; Oyang, Y.-J.; Huang, H.-C.; Juan, H.-F. A Single-Cell Network-Based Drug Repositioning Strategy for Post-COVID-19 Pulmonary Fibrosis. *Pharmaceutics* **2022**, *14* (5), 971.

(29) Islam, M. A.; et al. Bioinformatics-based investigation on the genetic influence between SARS-CoV-2 infections and idiopathic pulmonary fibrosis (IPF) diseases, and drug repurposing. *Sci. Rep.* **2023**, *13* (1), 4685.

(30) Alzubaidi, L.; et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8* (1), 53.

(31) Adeshina, Y. O.; Deeds, E. J.; Karanicolas, J. Machine learning classification can reduce false positives in structure-based virtual screening. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (31), 18477−18488.

(32) Yan, J.; Risacher, S. L.; Shen, L.; Saykin, A. J. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief. Bioinform.* **2017**, *19* (6), 1370−1381.

(33) Dara, S.; Dhamercherla, S.; Jadav, S. S.; Babu, C. M.; Ahsan, M. J. Machine Learning in Drug Discovery: A Review. *Artif. Intell. Rev.* **2022**, *55* (3), 1947−1999.

(34) Mao, F.; et al. Chemical Structure-Related Drug-Like Criteria of Global Approved Drugs. *Molecules* **2016**, *21* (1), 75.

(35) Wishart, D. S.; et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668.

(36) Kim, S.; et al. PubChem 2023 update. *Nucleic Acids Res.* **2023**, *51* (D1), D1373−D1380.

(37) Gaulton, A.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100.

(38) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.

(39) Mahmud, S. M. H.; et al. PreDTIs: prediction of drug-target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques. *Brief. Bioinform.* **2021**, *22* (5), No. bbab046.

(40) Kuhn, M.; Letunic, I.; Jensen, L. J.; Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **2016**, *44* (D1), D1075−D1079.

(41) Piñero, J.; et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **2017**, *45*, D833.

(42) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.

(43) Alzen, J. L.; Langdon, L. S.; Otero, V. K. A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. *Int. J. STEM Educ.* **2018**, *5* (1), 56.

(44) Hazimeh, H.; Ponomareva, N.; Mol, P.; Tan, Z.; Mazumder, R. The tree ensemble layer: Differentiability meets conditional computation. In *37th International Conference on Machine Learning ICML*, 2020, *PartF16814*, pp. 4096-4106.

(45) Sakri, S.; Basheer, S. Fusion Model for Classification Performance Optimization in a Highly Imbalance Breast Cancer Dataset. *Electronics* **2023**, *12* (5), 1168.

(46) Zhou, L.; Pan, S.; Wang, J.; Vasilakos, A. V. Machine learning on big data: Opportunities and challenges. *Neurocomputing* **2017**, *237*, 350−361.

(47) Zhou, Y.; Hou, Y.; Shen, J.; Huang, Y.; Martin, W.; Cheng, F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discovery 2020 61* **2020**, *6* (1), 1−18.

(48) Ahmed, F.; et al. A comprehensive review of artificial intelligence and network based approaches to drug repurposing in Covid-19. *Biomed. Pharmacother.* **2022**, *153*, No. 113350.

(49) Ahmed, F.; et al. Decade of bio-inspired soft robots: a review. *Smart Mater. Struct.* **2022**, *31* (7), No. 073002.

(50) Edgar, R.; Domrachev, M.; Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30* (1), 207−210.

(51) Carracedo-Reboredo, P.; et al. A review on machine learning approaches and trends in drug discovery. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4538−4558.

(52) Wu, Z.; Peng, Y.; Yu, Z.; Li, W.; Liu, G.; Tang, Y. NetInfer: A Web Server for Prediction of Targets and Therapeutic and Adverse Effects via Network-Based Inference Methods. *J. Chem. Inf. Model.* **2020**, *60* (8), 3687−3691.

(53) Karimizadeh, E.; et al. Analysis of gene expression profiles and protein-protein interaction networks in multiple tissues of systemic sclerosis. *BMC Med. Genomics* **2019**, *12* (1), 199.

(54) Szklarczyk, D.; et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47* (D1), D607−D613.

(55) Xie, B.; et al. Idiopathic Pulmonary Fibrosis Registry China study (PORTRAY): protocol for a prospective, multicentre registry study. *BMJ. Open* **2020**, *10* (11), No. e036809.

(56) Liu, Y.; Xu, C.; Gao, W.; Liu, H.; Li, C.; Chen, M. Transcriptome Classification Reveals Molecular Subgroups in Idiopathic Pulmonary Fibrosis. *Genet. Res.* **2022**, *2022*, No. 7448481.

(57) Leng, D.; Yi, J.; Xiang, M.; Zhao, H.; Zhang, Y. Identification of common signatures in idiopathic pulmonary fibrosis and lung cancer using gene expression modeling. *BMC Cancer* **2020**, *20* (1), 986.

(58) Reimand, J.; et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **2019**, *14* (2), 482−517.

(59) Ge, S. X.; Jung, D.; Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **2020**, *36* (8), 2628−2629.

(60) Pan, X.; O'Connor, M. B. Coordination among multiple receptor tyrosine kinase signals controls Drosophila developmental timing and body size. *Cell Rep.* **2021**, *36* (9), No. 109644.

(61) Rangarajan, S.; Locy, M. L.; Luckhardt, T. R.; Thannickal, V. J. Targeted Therapy for Idiopathic Pulmonary Fibrosis: Where To Now? *Drugs* **2016**, *76* (3), 291−300.

(62) Güvenç Paltun, B.; Kaski, S.; Mamitsuka, H. Machine learning approaches for drug combination therapies. *Brief. Bioinform.* **2021**, *22* (6), No. bbab293.

(63) Yang, M.; Jaaks, P.; Dry, J.; Garnett, M.; Menden, M. P.; Saez-Rodriguez, J. Stratification and prediction of drug synergy based on target functional similarity. *NPJ. Syst. Biol. Appl.* **2020**, *6* (1), 16.

(64) Jeon, M.; Kim, S.; Park, S.; Lee, H.; Kang, J. In silico drug combination discovery for personalized cancer therapy. *BMC Syst. Biol.* **2018**, *12* (2), 16.

(65) Steinmetz, K. L.; Spack, E. G. The basics of preclinical drug development for neurodegenerative disease indications. *BMC Neurol.* **2009**, *9* (Suppl 1), S2.

(66) Trachalaki, A.; Sultana, N.; Wells, A. U. An update on current and emerging drug treatments for idiopathic pulmonary fibrosis. *Expert Opin. Pharmacother.* **2023**, *24* (10), 1125−1142.

(67) Kim, S. K.; Jung, S. M.; Park, K.-S.; Kim, K.-J. Integrative analysis of lung molecular signatures reveals key drivers of idiopathic pulmonary fibrosis. *BMC Pulm. Med.* **2021**, *21* (1), 404.

(68) Emig, D.; et al. Drug Target Prediction and Repositioning Using an Integrated Network-Based Approach. *PLoS One* **2013**, *8* (4), No. e60618.

(69) Wang, J.; Liu, Y.; Li, P.; Lin, Z.; Sindakis, S.; Aggarwal, S. Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality. *J. Knowl. Econ.* **2023**, DOI: 10.1007/s13132-022-01096-6.

(70) Amiri Souri, E.; Chenoweth, A.; Karagiannis, S. N.; Tsoka, S. Drug repurposing and prediction of multiple interaction types via graph embedding. *BMC Bioinformatics* **2023**, *24* (1), 202.

(71) Sadybekov, A. V.; Katritch, V. Computational approaches streamlining drug discovery. *Nature* **2023**, *616* (7958), 673−685.