Cluster analysis and ensemble transfer learning for COVID-19 classification from computed tomography scans

Lyubomir Gotsev [a,1,*], Ivan Mitkov [a,2], Eugenia Kovatcheva [a,3], Boyan Jekov [a,4], Roumen Nikolov [a,5], Elena Shoikova [a,6], Milena Petkova [a,7]

[a] State University of Library Studies and Information Technologies, Sofia, Bulgaria
[1] l.gotsev@unibit.bg; [2] ivan.rmitkov@gmail.com; [3] e.kovatcheva@unibit.bg; [4] b.jekov@unibit.bg; [5] r.nikolov@unibit.bg;
[6] e.shoikova@unibit.bg; [7] m.petkova@unibit.bg
* corresponding author

ARTICLE INFO

Article history
Received March 29, 2022
Revised April 22, 2022
Accepted May 13, 2022
Available online July 31, 2022

Keywords
COVID-19
Computed tomography
Binary classification
Clustering
Ensemble learning

ABSTRACT

The paper presents a brief analysis of publications utilizing the public SARS-CoV-2 dataset, consisting of patients' computer tomography scans captured from Brazil hospitals and an experimental setup addressing the found data challenges. The analysis shows that all protocols, with one exception, suffer from data leakage arising from data organization where the patients and their images are not grouped. Each patient is represented with several scans. It can provide misleading results as data of the same individual may occur in both training and test sets. Furthermore, only one paper proposed ensemble learning utilizing as base models VGG-16, ResNet50, and Xception. Therefore, we proposed and experimented with the following strategy to mitigate the found risks of bias: data standardization and normalization to achieve proper contrast and resolution; k-means and group shuffle split to avoid data leakage; augmentation and ensemble transfer learning to deal with limited sample size and over-fitting. Compared with the earlier proposed ensemble approach, the current one stacks VGG-16, Densenet-201, and Inception v3, achieving higher accuracy (99.3 %), second in the related work, and most significantly, it applies augmentation and clustering analysis to avoid overestimation. In contrast, the paper also presented critical metrics in the medical domain: negative prediction value (99.55%), false positive rate (0.89%), false negative rate (0.42%), and false discovery rate (0.83%). The strategy has two main advantages: reducing data pitfalls and decreasing generalization error. It can serve as a baseline to increase the performance quality and mitigate the risk of bias in the field.

1. Introduction

In recent years COVID-19 pandemic has presented an unprecedented challenge to public health with impact and implications in all aspects of human life. The scientific community has focused on understanding the SARS-CoV-2 pathogenesis, developing and improving treatment, prevention, and diagnostic methods. In partnership with policymakers, researchers from various disciplines contribute to controlling the spread and overcoming the consequences of the pandemic. On the other hand, emerging technologies accelerate and optimize the technical and human resources in tackling the global health threat. Applying machine learning-based models to medical imaging problems can support, optimize, and automate the prognostication and diagnosis process. In the Artificial Intelligence (AI) field, computer vision strategies for discovering patterns on CT imaging have been developed, particularly for SARS-CoV-2 infected patients [1], [2]. The most recent systematic literature reviews discuss the findings of the detection and classification of coronavirus images (X-ray and CT) using machine and deep learning techniques from various perspectives.

The entire pipeline of AI-empowered medical imaging, analysis techniques, and applications in COVID-19 is covered in a methodological study [3], including two modalities demonstrating its effectiveness: X-ray and CT. Four main research directions are discussed: AI-empowered contactless imaging workflows, AI-aided image segmentation, AI-assisted differential diagnosis of COVID-19, and AI in follow-up studies. Furthermore, CT-based screening is grouped into classification tasks distinguishing COVID-19: other types, non-COVID-19, and determining the infection severity. According to the review, numerous current AI segmentation and diagnosis studies use small samples, leading to overfitting results. The data quality and quantity should be improved to make the results clinically useful. Although deep learning has emerged as the most effective strategy, imaging may have incomplete and inexact labels, making it complicated to train an accurate network. Labeling is costly and time-consuming, prompting researchers to look into self-supervised deep learning and deep transfer learning methods. It is critical for better COVID-19 screening and diagnosis to merge data from multiple sources such as imaging with clinical manifestations and laboratory examination results. Recommendations for treatment evaluation and follow-up are provided. Medical imaging, natural language processing, and ontology are examples of multidisciplinary integration that enhance the overall COVID-19 measurement. A preference for CT scans directory, recognition of deep transfer learning as an appropriate approach, and attention to the quality of small labeled data are review's contributions to the current research.

The appropriate criteria for the evaluation and the correct benchmark procedure of AI techniques, among the others, are at the focus of a systematic review [4], considering the COVID-19 medical images classification tasks (both for X-ray and CT): binary, multi-class, integrated multi-class and binary, and integrated hierarchical and multi-class. The review highlights the challenges of various evaluation criteria, where its type and number are different within each of the four identified tasks. The meaning and calculation of precision in binary classification differ from precision in multi-class types. Among the others, criteria trade-off and criteria importance is also challenging. The discussed study presents a three-phase methodology for evaluating and benchmarking AI techniques used in all COVID-19 image classification tasks to address the described complexity. The significant steps in the first phase are identifying the dataset and required pre-processing, evaluation criteria, and proper classification techniques (importance of accuracy and loss function to avoid over-under-fitting issues). The decision matrix as an output refers to the intersection (in values) between each AI technique and identified evaluation criteria of each task. The second, namely the development phase, applies the integrated AHP (Analytic Hierarchy Process) [5] and VIKOR (VlseKriterijumska Optimizacija I Kompromisno Resenje) [6] methods. The last phase utilizes objective (the mean ± standard deviation to validate the results) and subjective (models evaluated by experts) validation of the proposed solution. Information obtained directed the current experimental session to the binary classification task with corresponding evaluation metrics.

A thorough review, Roberts *et al.* [7] emphasizes the common pitfalls and recommendations in using machine learning to detect and prognosticate COVID-19 from chest radiographs (CXR) and CT scans. In addition to the datasets considered, several diagnostic domains are presented: CXR, CT and traditional machine learning methods, CT and deep learning. The summary of data extracted includes the type of task (diagnosis/prognosis or both), data used in the model (X-ray/ CT scans or both), predictors, development (training and validation) and test sample size, type of validation (internal/external/both), evaluation (performance of the model), public code (available or not). Participants, predictors, outcomes, and analysis are all used as domains to assess risks of bias (following the PROBAST - Prediction model Risk of Bias Assessment Tool [8]). According to Roberts *et al.* [7] and Wynants *et al.* [9], the identified models' main weakness is their inability to be used in clinical settings due to methodological flaws and/or underlying biases. Many studies suffer from frequently encountered issues: insufficient data (not large enough or of poor quality, a high or unclear risk of bias, poor integration of multistream data, not representative of the target population, poor demographic statistics, including age and sex distribution), deficiencies in methodology and study design, poor reproducibility critical for deployment in clinical practice. Based on the analysis findings, the authors made a series of recommendations organized in domains: data, evaluation, replicability, authors, and reviewers. Wherever possible, the current research takes into account described pitfalls and advice.

Previous studies present a literature review of research in X-ray and CT imaging directories. Hassan *et al.* [10] focus entirely on SARS-CoV-2 diagnostic methods on CT images, categorizing the AI-enabled models per computer vision tasks: classification, segmentation, and detection. Another main contribution is the curation of important information about the 29 most extensively used and essential chest CT datasets utilized for COVID-19 research organized in the following clusters: large and small datasets with supplemental AI-based models, datasets with no supplemental models, datasets with supporting clinical information and data augmentation–based datasets. Released details are utilized to identify relevant and quality data and define the main task in the current experimental session. Backing the critical information about the most widely used and essential COVID-19 chest CT datasets presented in Hassan *et al.* [10], a balance is achieved between the previously discussed data challenges, available data, and desired characteristics to identify the appropriate dataset for the study. Unlike most publications applying transfer learning, Biswas *et al.* [11] propose a classification approach built around prototype-based learning called eXplainable Deep Learning (xDNN). On the other hand, Sonali *et al.* [12] present an optimized convolutional neural network model named ADECO-CNN that achieved the highest results. However, possible data leakage exists using splitting at random. The most extended protocol with three setups, three scenarios, a mix of them, and one of two applied external validation is Silva *et al.* [13]. Silva *et al.* [13] detailed a real cross-dataset analysis, whereas Sonali *et al.* [12] provide less information about external validation. Furthermore, in [13] a discussed series of data challenges are overcome using appropriate pre-processing and two splitting scenarios besides the random one: the "slice" scenario and the voting-based approach. The proposed models are entirely built on the B0 and modified (smaller, deeper) architectures of the EffiecientNet family.

In contrast, Biswas *et al.* [11] proposes applying an ensemble approach that stacks three base models to achieve promising results. Moreover, a gradient-weighted class activation mapping (Grad- CaMS [14]) is utilized. In Biswas *et al.* [11] additional experiments merged the dataset with other public data changing the binary classification (COVID-19/non-COVID-19 but with other pulmonary diseases) to a 3-class (+ Healthy) task. Still, the validation is internal. The [15] study takes an approach to investigate histogram equalization techniques' impact on the transfer learning models' performance. Therefore, the first step obtains the original dataset and two additional copies, then applies Histogram Equalization (HE) [16] to one copy and Contrast Limited Adaptive Histogram Equalization (CLAHE) [12] to the other, resulting in three datasets. The results show that the VGG-19 combined with a dataset using CLAHE achieved the best overall performance, but the highest specificity performed MobileNet-V2 architecture with a dataset using HE. The experiment results do not give conclusive proof or a definitive answer on whether or not histogram equalization techniques have any significant impact on the overall models' performance. However, with one exception ([13]), other protocols suffer from at least one or more data challenges. Furthermore, half of the protocols have no or insufficient information about data pre-processing (Table 1).

Table 1. Overcome Techniques in Protocols

Data Challenge	Overcome technique in classification protocol					
	[17]	[11]	[18]	[19]	[13]	[15]
Image size and contrast	No information	No information	No information	Standardization Normalization Converting the RGB to YUV and YUV back to RGB	Standardization Normalization	Standard Histogram Equalization (HE) Contrast Limited Adaptive HE (CLAHE)
Scans organization by patients	x	x	x	x	Voting-based approach	x
Data splitting Training/validation /testing	x random) 80/20	x (random) 80/20	x (random) 68/17/15	x (random) 70/30	Random/ Slice/Voting	x (random) 60/20/20
Volume	x	TL	aug + TL	aug	TL	aug + TL

This paper reveals the potential for synergetic application of techniques and methods in deep learning to detect COVID-19 pneumonia in chest computed tomography (CT) images. First, the research briefly analyzes recent scientific publications and systematic reviews utilizing the public SARS-CoV-2 dataset to identify the main challenges, gaps, and recommendations. Second, it proposed a strategy addressing the found challenges. Third, it reveals the experimental setup of applying the strategy by combining different techniques with an ensemble method to mitigate bias risks. Finally, the suggested solution is compared with the related protocols.

2. Method

2.1 Dataset

This section points out how far the challenges have been overcome basis on data attributes. The SARS-CoV-2 CT is a publicly available dataset [17] of 2482 CT images retrieved from hospital settings (Public Hospital of the Government Employees and Metropolitan Hospital of Lapa, Sao Paulo, Brazil) from 120 adults with a balanced distribution by infected/non-infected patients and by gender (Table 2).

Some general challenges are overcome by choosing the SARS-CoV-2 dataset, but another arises. Although the images are primarily centered, the scans are not in the original DICOM format suitable for exploring scanner and slice thickness parameters. The pre-processing includes techniques such as standardization and normalization as CT images have different contrast (intensities and grayscales). Data are organized into COVID-19 and non-COVID-19 directories (Fig. 1), but neither has organization by patients. Beneficial, the distribution is known due to the dataset description provided by Biswas *et al.*

[11], such an issue shows a high risk of bias through so-called data leakage. Because several scans present one patient and the patients and their images are not grouped, the data from the same individual can occur in both the training and test sets simultaneously. It may produce misleading outcomes (overestimated results), especially splitting at random. A k-means-based patient clustering method can be helpful [14].

Table 2. Data Characteristics

Characteristic	Description
Access (availability)	Publicly available at https://github.com/Plamen-Eduardo/xDNN-SARS-CoV-2-CT-Scan and https://www.kaggle.com/plameneduardo/sarscov2-ctscan-dataset
Volume	With a limited sample size but relatively large in terms of [8], defined as a large dataset with supplemental AI-based models
Quality	Contrasted, focused images (centered) No missing data No significant artifacts No duplicated items
Format	PNG
Veracity	Hospital databases, Real patients, infected patients confirmed by PCR - test confirmation *Used and presented in scientific papers* [11], [17], [20] *and reviews* [9], [10]
Distribution	Balanced in # of scans by classes and by # of patients in total and per class
Representativity (demographic & target group)	Only Adults (>18), Balanced by gender (male-female) in both classes, Balance by SARS-CoV-2 virus-infected patients and non-infected ones with other pulmonary diseases
Risk of bias	High, further mitigated to low

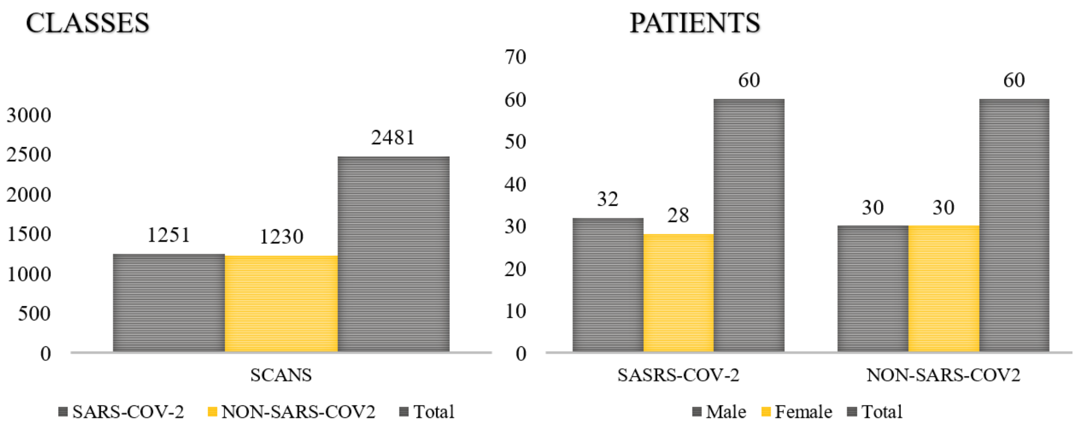


Fig. 1. Data Distribution

The narrower the literature review scope is, the more adaptive the protocol is, and the more risks of bias could be mitigated. The related publications [12], [13], [15]–[17] are beneficial in identifying more specific (data and task-related) challenges and improving the setup and precise results. However, not all of the papers found thus far consider the data gaps entirely. Table 3 summarizes the methods, measures, and highest achieved results in related work applying models on the SARS-CoV-2 CT-scan dataset, the same for the current study. The protocols applying just splitting at random may suffer from overestimated results. The conclusion is confirmed in Silva *et al.* [13].

Table 3. Methods, Metrics, Validation and Results in papers utilizing the SARS-CoV-2 CT-scan dataset

Ref	Method(s)	Metrics	Validation	Results (%)
[17]	Prototype-based deep learning	Precision	Internal	99.16
	eXplainable Deep Learning classification	Recall		95.53
	approach (xDNN)	F Score		97.31
		Accuracy		97.38
		AUC		97.36
[11]	TL	Precision	Internal	98.79
	Stacking ensemble (base models	Recall		98.79
	VGG-16, ResNet50, Xception)	F Score		98.79
	Data merge (binary to the 3-class task)	Accuracy		98.79
	Grad-CaMS	AUC		98.8
[20]	TL	Precision	Internal	96.29
	DenseNet201 (highest results),	Recall		96.29
	VGG16	F score		96.29
	ResNet152V2	Specificity		96.21
	Inception-ResNetV2	Accuracy		96.25
		AUC		97
[19]	optimized CNN model (ADECO-CNN)		Internal	Depending on validation
	TL		External	External to Internal
	VGG-19 & GoogleNet, ResNet	Accuracy		98.2 to 99.99
	compared with ADECO-CNN	Sensitivity		95.7 to 99.96
		Precision		97.9 to 99.92
		Specificity		96.8 to 99.97
[13]	TL	Accuracy (Acc)	Internal	Depending on the mix of
	3 Architectures of EffiecientNet (Baseline	Sensitivity (Se)	(5-fold cross-	setups, scenarios &
	B0, Smaller, Deeper)	Positive	validation)	validation
	3 Setups	Prediction (Pc)	External	External to Internal
	(protocol proposed in Ref. [6] cross-	F score	(cross-dataset	Acc 56.16 to 87.68
	dataset evaluation, impact of input	AUC	analysis)	Se 53.06 to 83.67
	resolution)			Pc 54.74 to 93.98
	3 Scenarios (Random/ Slice/Voting)			
[15]	HE, CLAHE (Histogram Equalization		Internal	Depending on the HE
	techniques) + TL			techniques
	1 original + 2 copied datasets:	Accuracy		95.75 VGG-19 + CLAHE
	1) original - no equalization,	Precision		94.42 VGG-19 + CLAHE
	2) copied +HE,	Recall		97.13 VGG-19 + CLAHE
	3) copied +CLAHE	F score		95.75 VGG-19 + CLAHE
	ResNet-101, VGG-19 + CLAHE (best	ROC-AUC		99.30 VGG-19 + CLAHE
	overall performance),	Specificity		99.60 MobileNet-V2 + HE
	DenseNet201, EfficientNet-B4,			
	MobileNet-V2 + HE (best specificity)			

2.2. The proposed model

The presented analysis shows enough scope to improve the performance and mitigate the risks of bias. The proposed approach involves a strategy to address the challenges by applying all techniques and methods pointed out in Table 4. It contains data standardization and normalization to achieve proper contrast and size unification; k-means (Clustering) and group shuffle split to avoid data leakage; and augmentation and transfer learning to deal with limited sample size, over-fitting, and generalization errors before applying stacking ensemble learning. The pre-trained networks applied to build the base models are DenseNet201, VGG16, and InceptionV3. The first two demonstrate promising results in the discussed publications. InceptionV3 is not mentioned or experimented with so far over the SARS-CoV-2 CT dataset.

Table 4. Data Challenges and Proposed Overcome Data Challenges Techniques

Data Challenge	Overcome in the applied protocol (Risk mitigation)
Image size and contrast	Pre-processing strategy involves standardization and normalization
Data organization	Clustering approach (k-means)
Data splitting	Group shuffle split
Volume	Augmentation (Aug) + Transfer learning (TL)

This section gives detailed information about the proposed approach, data pre-processing applied techniques, methods, and architecture. Fig. 2 illustrates the experimental workflow.

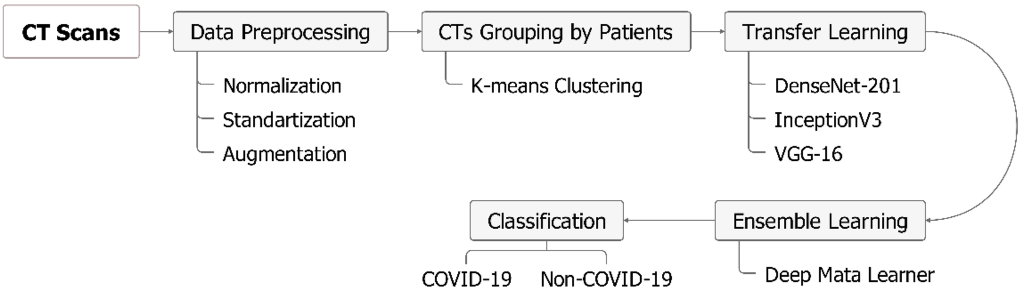


Fig. 2. Workflow

2.2.1 Data Pre-Processing

The data identification, characteristics, distribution, and challenges are already presented. The clinical data is completely anonymized, following all international standards and best practices for data privacy. Fig. 3 shows the examples of images not pre-processed for both classes provided.

Some common image data issues can be noticed: non-standard size and non-standard contrast. The dimensions of the most miniature image are 104 × 153, while the most extensive images are 484× 416 in the SARS-CoV-2 CT dataset. The pre-processing stage consists of three steps before feeding the classification algorithm to the images, as follows:

Step 1: The process of a simple pixel intensity rescaling the data to the range of 0-1 is defined as data normalization. That ensures the exact distribution of the input pixels, making the corresponding convolutional neural networks converge faster during the training phase.

Step 2: The input images have been converted into a fixed size (resized) to maintain compatibility with the preferred network architectures. The current project's input image resolution suitable for the pre-trained networks is 224x224.

Step 3: Adding altered versions of the existing images is known as data augmentation. It is a technique to increase the diversity of training sets by applying transformations, such as image rotation and scaling, thus enlarging the data volume. The purpose is to expose the ML classifier to a wider variety of artificial images generated by rotating the CTs up to 20° and shifting the pixels in height and width (by 20 pixels, while the CTs have also been randomly flipped). That additional pre-processing step is assumed to reduce the overfitting.

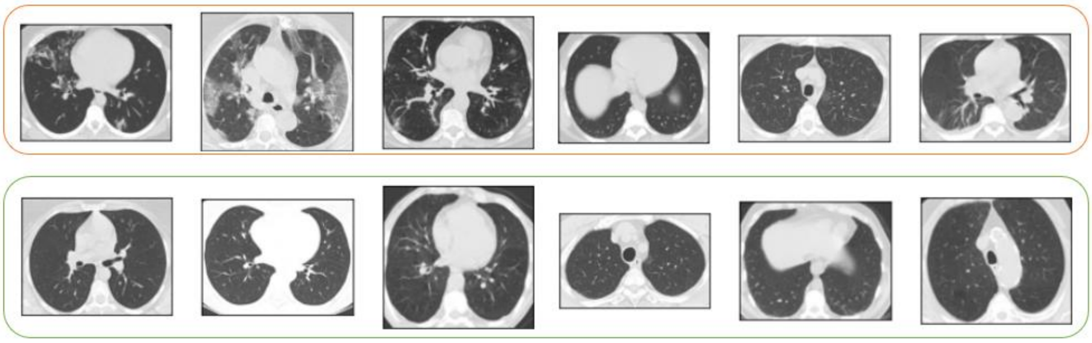


Fig. 3. COVID-19 (1st row) and non-COVID-19 CT Scans (2nd row)

2.2.2. Clustering

Besides the presented points above, the current research uses one more preparation step. The processed data is balanced, representing 60 infected and 60 non-infected patients (Fig. 4). However, the patients and their images are not grouped. The same patient may appear in the training and validation sets simultaneously as one patient is presented by several scan slices. That may produce misleading outcomes, especially in splitting training and validation sets at random.

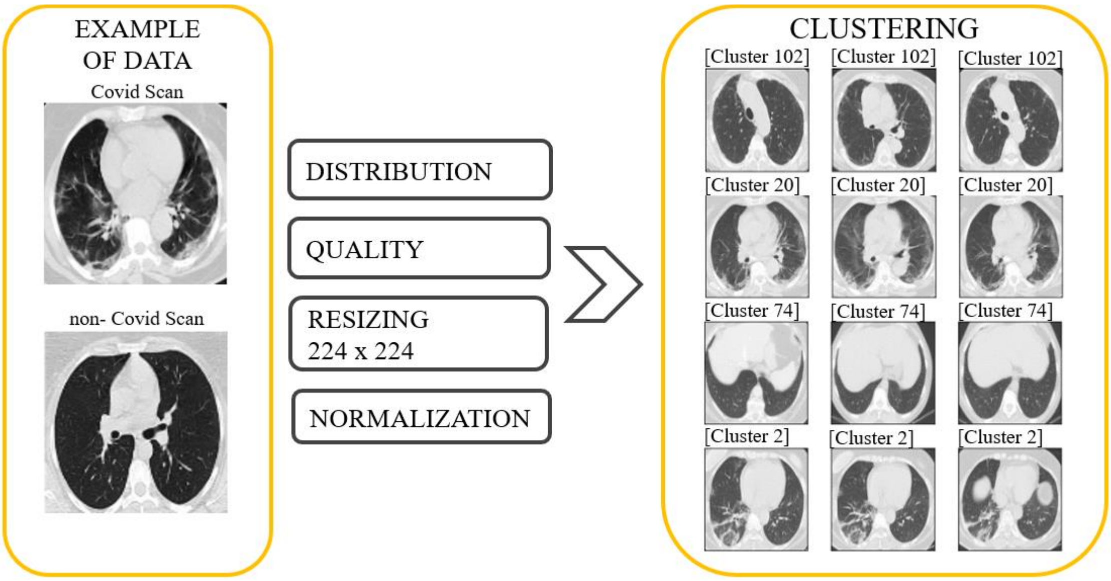


Fig. 4. Pre-processing to Clustering

It can lead to an overestimated result due to data from the same patient/individual in both training and test sets (so-called data leakage). To reduce such cases, the study uses one of the unsupervised learning algorithms, k-means cluster analysis. The data images should be grouped into 120 clusters, the same as the number of patients in the dataset. Then the group shuffle split is applicable to ensure no images from the same group appear in both training and validation sets. As a result, the chance that one patient has slices in both sets is minimized.

2.2.3. Transfer Learning

Since training convolutional neural networks (CNNs) requires massive input data to avoid overfitting, and due to limited computational resources, the current study relies on transfer learning (Fig. 5). It is a common and effective strategy employed when there are insufficient data to train a comprehensive model from scratch. It entails applying features learned from one problem to another [21], [22]. The most common method of transfer learning is as follows:

1) instantiate a base model with pre-trained weights

2) freeze the base model

3) build on top of the output of one (or several) layers from the base model

4) new data can be used to train the model

Different transfer learning approaches depend on the data and the problem. Therefore, modifications are possible. The appropriate workflow for the current issue has to dynamically modify the new model's input data during training, which is required when data augmentation is applied. The last one is a data-space solution to the problem of limited data [23].

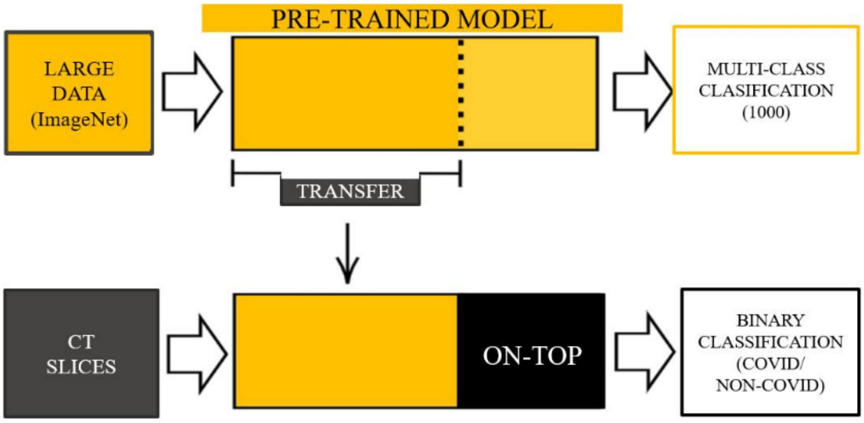

PRE-TRAINED MODEL
LARGE DATA (ImageNet)
MULTI-CLASS CLASIFICATION (1000)
TRANSFER
CT SLICES
ON-TOP
BINARY CLASSIFICATION (COVID/ NON-COVID)

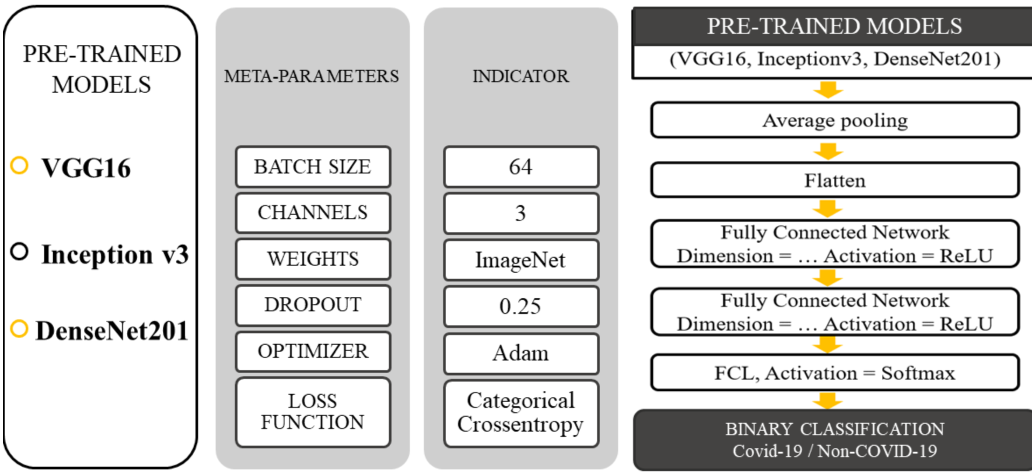
Fig. 5. Transfer Learning

In the experimental session, the knowledge gained by training a series of different CNNs on a large dataset called ImageNet [24] is transferred to the domain of COVID-19. It should be stressed that dealing with pre-trained algorithms based on the ImageNet is suitable for recognizing items from daily life. Therefore, in the approach, unfreezing the final layers of selected algorithms makes it possible to learn some of the characteristics of the training dataset. The final performance of the picked pre-trained networks also depends on their architecture. The three convolutional neural networks, namely DenseNet-201 [25], InceptionV3 [26], and VGG-16 [27], were selected. The first two demonstrate promising results in the related work. InceptionV3 is not mentioned or experimented with so far over the SARS-CoV-2 CT-scan dataset.

2.2.4. Implementation and Meta-Parameters

Before actual implementation (Python), it is essential to import all the relevant libraries: TensorFlow, Keras, Sklearn, Open CV, Matplotlib, Pandas, and NumPy. The main drivers are tensorflow.keras.applications and tensorflow.keras.layers to import pre-trained networks and layers involved in network construction. The ReduceLROnPlateau function reduces the learning rate when validation loss is not changing. The ImageDataGenerator method performs real-time augmentation while the model is still training.

A series of experiments are done with tunning the meta-parameters to avoid overfitting (Fig. 6). A dropout for regularization is applied before the classification layer.



META-PARAMETERS	INDICATOR
BATCH SIZE	64
CHANNELS	3
WEIGHTS	ImageNet
DROPOUT	0.25
OPTIMIZER	Adam
LOSS FUNCTION	Categorical Crossentropy

PRE-TRAINED MODELS
VGG16
Inception v3
DenseNet201
PRE-TRAINED MODELS
(VGG16, Inceptionv3, DenseNet201)
Average pooling
Flatten
Fully Connected Network Dimension = … Activation = ReLU
Fully Connected Network Dimension = … Activation = ReLU
FCL, Activation = Softmax
BINARY CLASSIFICATION
Covid-19 / Non-COVID-19

Fig. 6. Meta-Parameters and Model's Architecture

2.2.5. Ensemble Learning

The basic concept of ensemble learning is to train multiple base learners as ensemble members and combine their predictions into a single output that should have better performance on average than any other member with uncorrelated error on the target data sets [28]. Bagging, boosting, and stacking are transfer learning techniques commonly used for classification tasks [29], [30].

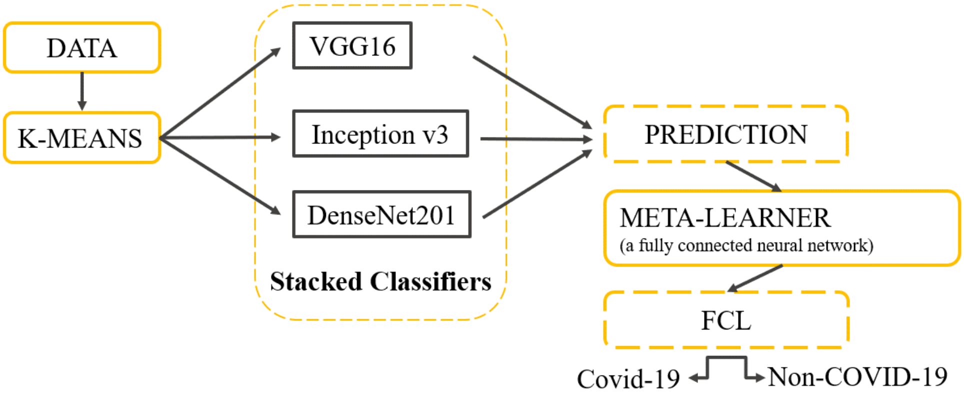


Fig. 7. Stacked Ensemble Approach

The current experimental session explores the stacking ensemble approach, also known as stacked generalization (Fig. 7). It ends with applying output probabilities for every class and weighted voting based on summing each sample and algorithm. Stacking provides a good rank bias for intelligent systems [31]. The input of the meta-learner is a vector of base models' predictions. Only the fully connected neural network (meta-learner) is trained again. The ensemble learning approach has two main advantages. The method can achieve better results than the base models. And it is another strategy for mitigating overfitting and decreasing the generalization error.

3. Results and Discussion

The performance of the proposed models is tested through a validation dataset, covering 20% of the input scans or 24 patients from all available (internal validation). The base models and meta-learner are trained in 50 epochs. Only the meta-learner is trained during ensemble learning, freezing the base models.

Various confusion-matrix-based and weighted evaluation metrics are considered for binary classification tasks, such as accuracy, precision, specificity, recall, and f-measure. Negative Prediction Value, False Positive Rate, False Negative Rate, and False Discovery Rate are evaluated. They are critical measurements, especially in the medical field, where the minimum number of false-negative and false-positive outcomes is preferred to avoid human and social harm. The values of these metrics in the current approach are promising but can't be compared with the related work as they are not presented. The numbers of COVID-19 and non-COVID-19 images correctly classified are referred to as true-positive (TP) and true-negative (TN). False-positive (FP) represents non-COVID-19 patients incorrectly recognized as COVID-19 (Table 5). False-negative (FN) represents COVID-19 patients that are incorrectly recognized as non-COVID-19. Few scans are misclassified (two false positive and one false negative). Thus, the model achieves high accuracy and is sensitive enough, but the most important is that critical metrics commented show a good ability to avoid misclassification.

Table 5. Evaluation metrics

Metrics	Calculation
Sensitivity (Recall, True Positive Rate)	TP/(TP+FN)
Specificity (True Negative Rate)	TN/(TN+FP)
Precision (Positive Prediction Value)	TP/(TP+FP)
Negative Predictive Value	TN/(TN+FN)
False Positive Rate	FP/(FP+TN)
False Negative Rate	FN/(TP+FN)
False Discovery Rate	FP/(TP+FP)
Accuracy	(TP+TN)/(TP+TN+FP+FN)
F1-score	2*(Precision*Recall)/(Precision+Recall)

All proposed learners achieve relatively high performance, with Inception v3 having exceptional values in all observed metrics from transfer learning-based models (Table 6). However, the stacked ensemble classifier is the most efficient.

Table 6. Evaluation of classifiers

	DenseNet 201	*VGG-16*	*Inception v3*	*Ensemble*
Accuracy	94.83	97.20	99.13	99.35
Precision weighted	94.83	97.20	98.71	99.35
Recall weighted	94.83	97.20	98.71	99.35
F1-score weighted	94.83	97.20	98.71	99.35
AUC	95	98	1	1
Specificity	93.75	96.88	98.66	99.11
Negative Predictive Value	94.59	97.31	98.66	99.55
False Positive Rate	6.25	3.13	1.34	0.89
False Negative Rate	5	2.5	1.25	0.42
False Discovery Rate	5.79	2.9	1.25	0.83

In Fig. 8, the test curve stays slightly above train one, explained by the variance in train data added through real-time data augmentation and its absence in the test set. Dropout layers also could infer as they are "on" for training but "off" (skipped) when doing testing. In other worlds in training, due to disabling neurons, some of the information about each sample is lost, and the subsequent layers attempt to construct predictions based on incomplete representations. However, all of the units are available during validation, so the network has its full computational power - and thus, it might perform better than in training. The split rate is 0.2. The learning rate is not too high, so it is ignored as a reason. Nevertheless, the slight difference decreases with the increasing number of epochs. The ensemble classifier does not face such challenges.

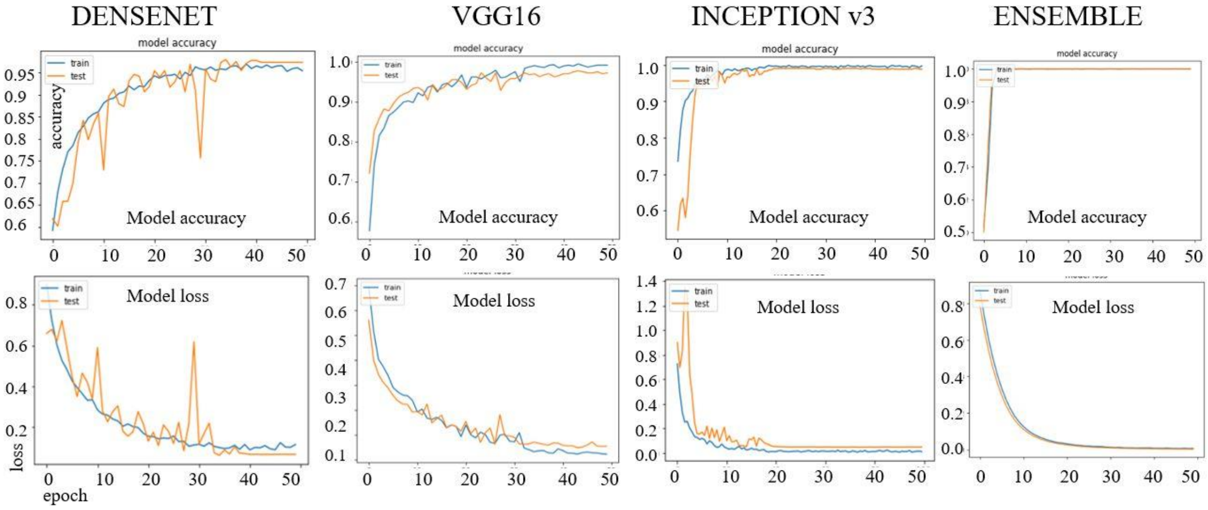


Fig. 8. Model Accuracy & Model Loss

Compared to the related work (Table 7), the data challenges are overcome, especially mitigating data leakage. The last issue is addressed only in Silva *et al.* [13], but the results presented are relatively low. All the other protocols suffer from such a problem. That can lead to overestimated results.

Utilizing the proposed strategy may result in poorer performance of models than those presented in the related work. However, applying a transfer learning ensemble following clustering analysis and augmentation achieves the second-highest results in COVID-19 classification on the SARS-COV2-CT dataset.

Table 7. Related and Current Work's Model Performance Evaluation

Ref #	Data Challenge	Internal Validation					External Validation				
		Accuracy	*Precision*	*Recall*	*F score*	*AUC*	*Accuracy*	*Precision*	*Recall*	*F score*	*AUC*
[17]	PDL[a]	97.38	99.16	95.53	97.31	97.36					
[11]	PDL[a]	98.79	98.79	98.79	98.79	98.8					
[20]	PDL[a]	96.25	96.29	96.29	96.29	97					
[19]	**PDL[a]**	**99.99**	**99.92**	**99.96**			**98.2**	**97.9**	**95.7**		
[13]	DLM[b]	87.68			86.19	90.51	56.16				
[15]	PDL[a]	95.75	94.42	97.13	95.75	99.3					
Proposed	**DLM[b]**	**99.35**	**99.35**	**99.35**	**99.35**	**1**					

[a] Possible Data Leakage, [b] Data Leakage Mitigation

Unlike Silva *et al.* [13], which successfully addressed the data and protocol pitfalls, the current research proposed an ensemble approach to achieving higher performance.

Lawton and Viriri [15] present a stacked ensemble using other base models than proposed, but the possible data leakage is not overcome. Furthermore, the proposed stacking achieves higher results. The reason for that is in applying Inception v3 as a base model. In related work, that pre-trained network is not utilized. However, the current experimental setup demonstrates higher accuracy than the other base models and the base models used in Lawton and Viriri [15]. Another significant difference is Densenet 201 appliance. The last demonstrates promising results in related work, which is the reason to use it as a base model. The earlier paper extends the binary classification to a three-class classification, thus enlarging the discussed dataset. Therefore, we compare the results achieved on the binary task as a primary one in the present research. Lawton and Viriri [15] consolidate the new dataset and then present additional results on the new three-class classification. Sonali *et al.* [12] and Silva *et al.* [13] performed external validation, whereas the current validation is internal - one of the most common protocol challenges.

4. Conclusions

An analysis of related work utilizing the experimented dataset is provided, founding challenges summarized in three groups: general, data and protocol. A strategy addressing the data and protocol challenges is proposed: data standardization and normalization to achieve proper contrast and resolution unification; k-means (Clustering) and group shuffle split to avoid data leakage; augmentation and transfer learning to deal with limited sample size and over-fitting. The presented model design with stacked ensemble learning decreases the generalization error and combines the ensemble members' predictions into a single output outperforming any other member on average. The proposed ensemble transfer learning classifier follows augmentation and cluster analysis, thus making the model more robust and able to handle data pitfalls. Although the results are promising, internal validation is the most significant protocol weakness. The lack of cross-data analysis and the limited amount of training data are the main drawbacks that should be highlighted. An extended experimental session is set up, taking into account these considerations. Currently, the research focuses on retrieving more image data for training and external validation. The following study aims to compare the reported results and the new ones over the vast collection with external validation answering the question about the model's ability for generalization.

Acknowledgment

Project "BG05M2OP001-1.001-0003 "Center of Excellence in Informatics and Information and Communication Technologies" funded by the OP "Science and education for smart growth "2014-2020", EU through the European structural and investment funds.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. The research is funded by Project "BG05M2OP001-1.001-0003 "Center of Excellence in Informatics and Information and Communication Technologies".

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] E. Awulachew, K. Diriba, A. Anja, E. Getu, and F. Belayneh, "Computed Tomography (CT) Imaging Features of Patients with COVID-19: Systematic Review and Meta-Analysis," *Radiol. Res. Pract.*, vol. 2020, pp. 1–8, Jul. 2020, doi: 10.1155/2020/1023506.

[2] I. Soriano Aguadero *et al.*, "Chest computed tomography findings in different phases of SARS-CoV-2 infection," *Radiol. (English Ed.*, vol. 63, no. 3, pp. 218–227, May 2021, doi: 10.1016/j.rxeng.2021.02.003.

[3] O. S. Albahri *et al.*, "Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects," *J. Infect. Public Health*, vol. 13, no. 10, pp. 1381–1396, Oct. 2020, doi: 10.1016/j.jiph.2020.06.028.

[4] F. Shi *et al.*, "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 4–15, 2021, doi: 10.1109/RBME.2020.2987975.

[5] A. U. Khan and Y. Ali, "Analytical Hierarchy Process (AHP) and analytic network process methods and their applications: a twenty year review from 2000-2019," *Int. J. Anal. Hierarchy Process*, vol. 12, no. 3, pp. 369–459, Dec. 2020, doi: 10.13033/ijahp.v12i3.822.

[6] J. H. Kim and B. S. Ahn, "Extended VIKOR method using incomplete criteria weights," *Expert Syst. Appl.*, vol. 126, pp. 124–132, Jul. 2019, doi: 10.1016/j.eswa.2019.02.019.

[7] M. Roberts *et al.*, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nat. Mach. Intell.*, vol. 3, no. 3, pp. 199–217, Mar. 2021, doi: 10.1038/s42256-021-00307-0.

[8] G. S. Collins *et al.*, "Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence," *BMJ Open*, vol. 11, no. 7, pp. 1–7, Jul. 2021, doi: 10.1136/bmjopen-2020-048008.

[9] L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ*, vol. 369, pp. 1–11, Apr. 2020, doi: 10.1136/bmj.m1328.

[10] H. Hassan *et al.*, "Review and classification of AI-enabled COVID-19 CT imaging models based on computer vision tasks," *Comput. Biol. Med.*, vol. 141, pp. 1–21, Feb. 2022, doi: 10.1016/j.compbiomed.2021.105123.

[11] S. Biswas, S. Chatterjee, A. Majee, S. Sen, F. Schwenker, and R. Sarkar, "Prediction of COVID-19 from Chest CT Images Using an Ensemble of Deep Learning Models," *Appl. Sci.*, vol. 11, no. 15, pp. 1–16, Jul. 2021, doi: 10.3390/app11157004.

[12] Sonali, S. Sahu, A. K. Singh, S. P. Ghrera, and M. Elhoseny, "An approach for de-noising and contrast enhancement of retinal fundus image using CLAHE," *Opt. Laser Technol.*, vol. 110, pp. 87–98, Feb. 2019, doi: 10.1016/j.optlastec.2018.06.061.

[13] P. Silva *et al.*, "COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis," *Informatics Med. Unlocked*, vol. 20, pp. 1–9, 2020, doi: 10.1016/j.imu.2020.100427.

[14] 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," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[15] S. Lawton and S. Viriri, "Detection of COVID-19 from CT Lung Scans Using Transfer Learning," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–14, Apr. 2021, doi: 10.1155/2021/5527923.

[16] K. G. Dhal, A. Das, S. Ray, J. Gálvez, and S. Das, "Histogram Equalization Variants as Optimization Problems: A Review," *Arch. Comput. Methods Eng.*, vol. 28, no. 3, pp. 1471–1496, May 2021, doi: 10.1007/s11831-020-09425-1.

[17] E. Soares, P. Angelov, S. Biaso, M. H. Froes, and D. K. Abe, "SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification," *MedRxiv*. Cold Spring Harbor Laboratory Press, pp. 1–8, 2020, doi: 10.1101/2020.04.24.2007858.

[18] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, and M. Kaur, "Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning," *J. Biomol. Struct. Dyn.*, vol. 0, no. 0, pp. 1–8, 2020, doi: 10.1080/07391102.2020.1788642.

[19] A. Castiglione, P. Vijayakumar, M. Nappi, S. Sadiq, and M. Umer, "COVID-19: Automatic Detection of the Novel Coronavirus Disease From CT Images Using an Optimized Convolutional Neural Network," *IEEE Trans. Ind. Informatics*, vol. 17, no. 9, pp. 6480–6488, Sep. 2021, doi: 10.1109/TII.2021.3057524.

[20] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, and M. Kaur, "Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning," *J. Biomol. Struct. Dyn.*, vol. 39, no. 15, pp. 5682–5689, Oct. 2021, doi: 10.1080/07391102.2020.1788642.

[21] F. Chollet, *Deep learning mit python und keras: das praxis-handbuch vom entwickler der keras-bibliothek*. United States of America: MITP-Verlags GmbH & Co. KG, 2018. Available at: Google Books

[22] F. Zhuang *et al.*, "A Comprehensive Survey on Transfer Learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.

[23] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[24] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255, doi: 10.1109/CVPRW.2009.5206848.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[27] S. Tammina, “Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images,” *Int. J. Sci. Res. Publ.*, vol. 9, no. 10, pp. 143–150, Oct. 2019, doi: 10.29322/IJSRP.9.10.2019.p9420.

[28] Y. Yang, “Ensemble Learning,” in *Temporal Data Mining Via Unsupervised Ensemble Learning*, Elsevier, 2017, pp. 35–56. doi: 10.1016/B978-0-12-811654-8.00004-X

[29] S. González, S. García, J. Del Ser, L. Rokach, and F. Herrera, “A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities,” *Inf. Fusion*, vol. 64, pp. 205–237, Dec. 2020, doi: 10.1016/j.inffus.2020.07.007.

[30] A. Telikani, A. Tahmassebi, W. Banzhaf, and A. H. Gandomi, “Evolutionary Machine Learning: A Survey,” *ACM Comput. Surv.*, vol. 54, no. 8, pp. 1–35, Nov. 2022, doi: 10.1145/3467477.

[31] S. Simske, “Introduction, overview, and applications,” in *Meta-Analytics*, Elsevier, 2019, pp. 1–98. doi: 10.1016/B978-0-12-814623-1.00001-0