Supplementary Appendix

AI-based pipeline for early screening of lung cancer: integrating radiology, clinical, and genomics data

Ullas Batra[2*], Shrinidhi Nathany[2*], Swarsat Kaushik Nath[1*], Joslia T Jose[2], Trapti Sharma[1], Preeti P[1], Sunil Pasricha[2], Mansi Sharma[2], Nevidita Arambam[1], Vrinda Khanna[1], Abhishek Bansal[2], Anurag Mehta[2], Kamal Rawal[#1*]

[1]Amity Institute of Biotechnology, Amity University, Noida, Uttar Pradesh, India; [2]Rajiv Gandhi Cancer Institute and Research Centre, New Delhi, India

[*]These authors contributed equally to this work

#Corresponding author:
Dr Kamal Rawal (PhD)
Email ID: krawal@amity.edu
91-120-4735614
HOD and Professor, Center for Computational Biology and Bioinformatics,
Amity Institute of Biotechnology, Amity University, Noida - 201303, India.

1. Supplementary methods

1.1 Additional references related to the prediction of EGFR status using CT images

1. Wang C, Ma J, Shao J, Zhang S, Liu Z, Yu Y, Li W. Predicting EGFR and PD-L1 status in NSCLC patients using multitask AI system based on CT images. Frontiers in immunology. 2022 Feb 18;13:813072.

2. Tan X, Li Y, Wang S, Xia H, Meng R, Xu J, Duan Y, Li Y, Yang G, Ma Y, Jin Y. Predicting EGFR mutation, ALK rearrangement, and uncommon EGFR mutation in NSCLC patients by driverless artificial intelligence: a cohort study. Respiratory Research. 2022 Dec;23(1):1-3.

3. Nguyen HS, Ho DK, Nguyen NN, Tran HM, Tam KW, Le NQ. Predicting EGFR Mutation Status in Non–Small Cell Lung Cancer Using Artificial Intelligence: A Systematic Review and Meta-Analysis. Academic Radiology. 2023 Apr 28.

4. Silva F, Pereira T, Morgado J, Frade J, Mendes J, Freitas C, Negrao E, De Lima BF, Da Silva MC, Madureira AJ, Ramos I. EGFR assessment in lung cancer CT images: analysis of local and holistic regions of interest using deep unsupervised transfer learning. IEEE Access. 2021 Apr 2;9:58667-76.

1.2 Inclusion criteria, data collection timeframe, data sources, and CT scanner information for each cohort

COHORT 1, 2, and 3

Cohorts 1-3 were consecutively included between Jan. 1st, 2015, and Dec. 31, 2021. To ensure a robust and reliable analysis, we divided the data into three distinct cohorts based on the time periods of data collection.

Cohort 1 was primarily utilised to train the AIPS-M machine learning and deep learning models, and it served as the foundation for our research study. Cohort 2 and Cohort 3 were reserved exclusively for independent testing of the trained models. This approach allowed us to evaluate the performance and generalizability of the models in an unbiased manner, specifically within the Indian population.

Regarding the differences between the cohorts, we would like to provide more clarity. Although the data were collected from the same institution, there were certain variations in demographic characteristics, disease stages, and treatment approaches over the years. These differences may have influenced the outcomes and model performance, and by separating the cohorts, we aimed to capture the potential temporal changes in cancer-related patterns and account for them during model evaluation.

The utilisation of independent testing cohorts (Cohorts 2 and 3) was essential to ensure that our machine learning and deep learning models could reliably predict cancer-related outcomes in a real-world setting. It helped validate the models' performance, minimise overfitting, and establish their robustness for use in the Indian clinical context.

All the patients in Cohorts 1, 2 and 3 satisfied the following inclusion criteria:

1. histologically confirmed primary lung cancer
2. pathologic examination of tumour specimens has been carried out with proven records of *EGFR* mutation status
3. diagnostic CT data obtained

Patients were excluded if:

1. received treatment before the CT examination
2. the duration between the CT examination and subsequent gene sequencing exceeded one-month

Technical information:

1. *Total number of subjects:* 2066 (Cohort 1 = 1379, Cohort 2 = 591, Cohort 3 = 96)

2. *Data collection timeframe:* Retrospectively collected between January 2015 till December 2021

3. *Source of data:* Rajiv Gandhi Cancer Institute and Research Centre, New Delhi

4. *Scanner manufacturer and model:*

Manufacturer Name	**Model Name**
GE Medical Systems	Discovery 600, Discovery 610, Discovery MI, Discovery ST, Discovery STE
Philips	GEMINI TF TOF 64, TruFlight Select, Ingenuity TF PET/CT
Siemens	1080, 1093, Biograph 16, Biograph 20, Biograph 20_mCT, Biograph 40, Biograph 64, Biograph 64_mCT, Biograph16_TruePoint, Biograph20, Biograph20_mCT, Biograph40_TruePoint

5. *CT scanning parameters:*
    a. *Slice thickness:* CT images with slice thicknesses equal to 5mm and 1mm were included
    b. *Peak voltage:* Information not available in the DICOM metadata.
    c. *X-ray tube current:* Information not available in the DICOM metadata.

6. *Mutational Testing:* The subjects diagnosed with lung carcinoma were subjected to *EGFR* testing by the therascreen *EGFR* Mutation detection kit as per manufacturer recommendations for tissue genotyping, plasma-based genotyping by Roche cobas V2 and bioRAD droplet digital PCR.

7. *Clinical data:* The hospital's medical records were used to determine the patient's clinical data such as age, gender, smoking status, and histology.

COHORT 4

All the patients in Cohort 4 satisfied the following inclusion criteria:

1. Data in the TCIA cohort 4 (Stanford Hospital) is publicly available. All 211 patients who have both *EGFR* gene detection and thick CT images were included (168 *EGFR*-wild type patients and 43 *EGFR*-mutant patients).

Technical information:

1. *Total number of subjects:* 211

2. *Data collection timeframe:* Subjects were recruited between April 7th, 2008 and September 15th,

2012 (**Bakr et al., 2018**).

3. *Source of data:* Stanford University School of Medicine, Stanford, California, United States and Veterans Affairs Healthcare System, Palo Alto, California, United States

4. *Scanner manufacturer and model:* The choice of scanners varied depending on the institution and the preferences of the physicians, and scanning protocols also differed among them (**Bakr et al., 2018**).

5. *CT scanning parameters:*
    a. *Slice thickness:* 0.625–3 mm (median: 1.5 mm)
    b. *Peak voltage:* 80–140 kVp (mean 120 kVp).
    c. *X-ray tube current:* 124–699 mA (mean 220 mA)

6. *Mutational Testing:* Single nucleotide mutation detection was performed using SNaPshot technology based on a dideoxy single-base extension of oligonucleotide primers after multiplex polymerase chain reaction (PCR). Exons 18, 19, 20 and 21 were tested for *EGFR* mutations.

COHORT 5

All the patients in Cohort 5 satisfied the following inclusion criteria:

1. Both standard-dose diagnostic CT scans and lower-dose CT scans from lung cancer screening examinations were considered acceptable for inclusion in the dataset.
2. For each scan included in the Database, a crucial requirement was that the collimation and reconstruction interval should not exceed 3 mm.

Technical Information:

1. *Total number of patients:* 1018

2. *Data collection timeframe: NA* (Not provided by the authors)

3. *Source of data:* A collaborative effort involving seven academic centres and eight medical imaging companies was undertaken to identify, address, and resolve complex organisational, technical, and clinical issues. The objective was to establish a strong foundation for a comprehensive database (**Armato et al., 2011**).

4. *Scanner manufacturer and model:* The dataset comprised scans from various scanner manufacturers and models. Specifically, it included 670 scans from seven different GE Medical Systems LightSpeed scanner models, 74 scans from four different Philips Brilliance scanner models, 205 scans from five different Siemens Definition, Emotion, and Sensation scanner models, and 69 scans from Toshiba Aquilion scanners (**Armato et al., 2011**).

5. *CT scanning parameters:*

a. *Slice thickness:* 0.6 mm (n=7), 0.75 mm (n=30), 0.9 mm (n=2), 1.0 mm (n=58), 1.25 mm (n=349), 1.5 mm (n=5), 2.0 mm (n=124), 2.5 mm (n=322), 3.0 mm (n=117), 4.0 mm (n=1), and 5.0 mm (n=3).
b. *Peak voltage:* 120 kV (n=818), 130 kV (n=31), 135 kV (n=69), and 140 kV (n=100)
c. *X-ray tube current:* 40–627 mA (mean 222.1 mA)

6. *Estimated size range of the lung nodules:* As per the annotated Cohort 5 [n = 1010] received from the expert radiologists, the estimated range of the nodules is 3 mm to 30 mm (**Armato et al., 2011**).

1.3 Potential solutions to tackle systematic differences due to site, scanner, and scanning parameters

The following are the potential solutions to tackle systemic differences due to site, scanner, and scanning parameters:

1. **Standardisation of Acquisition Parameters:** Ensuring that the acquisition parameters, such as voxel size, field of view, and slice thickness, are standardised across all sites. This step helps to minimise variations that can arise due to differences in scanning protocols.
2. **Image Preprocessing:** Preprocessing techniques like intensity normalisation and image registration may have been used to align the images and make them consistent across cohorts. Intensity normalisation ensures that pixel values have a common scale, while image registration corrects for spatial misalignments between images.
3. **Feature Extraction:** Instead of using raw pixel values directly, features may have been extracted from the images to capture relevant information while reducing the impact of site-specific variations. Common feature extraction techniques include texture analysis, shape analysis, and radiomic features.
4. **Domain Adaptation:** Advanced techniques like domain adaptation or transfer learning might have been applied to make the model more robust to domain shifts between different sites. These methods enable the model to learn from both source and target domains, effectively mitigating the effects of site-related differences.
5. **Data Augmentation:** Data augmentation techniques may have been used to artificially increase the size of the dataset, ensuring a more diverse representation of imaging data from different sites. This helps the model generalise better to unseen data.
6. **Multi-Site Validation:** To assess the model's performance on unseen data from different sites, a multi-site validation approach might have been employed. This involves training the model on data from one or more sites and testing it on data from a different site, allowing the evaluation of cross-site generalisation.
7. **Cohort-Specific Model Fine-Tuning:** Depending on the extent of site-specific variations, cohort-specific fine-tuning of the model might have been performed. This process adapts the model's parameters to the peculiarities of each cohort while ensuring a certain level of generalizability.

We applied rescaling and windowing techniques to DICOM images before training the object detection models. The sequence involved the following steps:

1. *Rescaling*: We normalised the intensity values of the DICOM images to a standardised range (between 0 and 255). This step ensured that the object detection models received input with consistent intensity ranges across images in different cohorts and avoided any biases resulting from varying intensity scales.
2. *Windowing*: We applied windowing to adjust the intensity ranges of the DICOM images based on the specific visualisation requirements. We selected an appropriate window width and level so that we can emphasise certain structures or pathologies while suppressing others, enhancing the visibility of the relevant information for the object detection model.

By applying rescaling followed by windowing, we optimised the data representation for training an object detection model.

1.4 Enhancing Cohort 5 representativeness for AIPS-N model generalisation: standardisation, diversity, and harmonisation strategies

Steps taken to ensure the representativeness of Cohort 5 for training the AIPS-N model and its subsequent generalisation to Cohorts 1-4:

1. **Image Format and Consistency:** All images within Cohorts 1-5 adhere to a uniform CT format and are stored in 3D DICOM format. This standardisation guarantees that input for the AIPS-N model remains consistent, mitigating any potential disparities in image data.

2. **Nature of Images:** Confirming the reviewer's observation, we affirm that all images across these cohorts are exclusively focused on lung cancer cases. This singular emphasis ensures direct relevance to the training process for the AIPS-N model.

3. **Comprehensive Selection Process:** The selection of Cohort 5 was meticulously undertaken to encompass a diverse spectrum of lung cancer scenarios, mirroring those present in Cohorts 1-4. The selection process factored in various clinical aspects, lesion attributes, and patient demographics. This approach aimed to minimise bias resulting from concentrating solely on specific subsets of cases.

4. **Global Diversity:** Importantly, Cohort 5's images were collected from diverse geographical locations worldwide. This global inclusion introduces ethnic diversity, enriching the dataset's representation.

5. **Data Harmonization Techniques:** It is noteworthy that data harmonisation techniques were employed across all cohorts to enhance consistency. Rescaling normalised intensity values to a standardised range (0 to 255), ensuring uniformity, while windowing adjusted intensity ranges based on visualisation needs. This technique emphasised specific structures or pathologies, enhancing information visibility for object detection models.

In summary, the careful curation of Cohort 5, standardised image format, consistent focus on lung cancer, global diversity, and data harmonisation collectively underscore its representativeness. These measures synergistically contribute to the AIPS-N model's capacity for effective generalisation across diverse cohorts, aligning with our aim to ensure accuracy and applicability in various lung cancer scenarios.

1.5 Comprehensive 3D Image Annotation for Lung Cancer Nodules by Radiologists: Blinded-Read and Unblinded-Read Phases

Each subject in the study was associated with both a clinical thoracic CT scan and an XML file containing the results of a comprehensive two-phase image annotation process. This annotation process involved the participation of four experienced thoracic radiologists.

During the initial blinded-read phase, each radiologist independently reviewed every CT scan in a blind manner. They carefully examined the scans and marked any identified lesions belonging to one of three specific categories: "nodule greater than or equal to 3 mm," "nodule less than 3 mm," and "non-nodule greater than or equal to 3 mm." The purpose of this phase was to capture the individual radiologists' independent interpretations and annotations without any influence or bias from other radiologists.

Following the blinded-read phase, the study proceeded to the subsequent unblinded-read phase. In this phase, each radiologist independently reviewed their own initial marks alongside the anonymized annotations made by the other three radiologists. The radiologists had access to all the marks, allowing them to consider multiple perspectives and opinions.

During the unblinded-read phase, each radiologist carefully reviewed and assessed the combined annotations to render their final opinion on the presence and classification of lung nodules in each CT scan. The goal of this phase was to leverage the collective expertise and insights of the four radiologists to identify lung nodules as comprehensively as possible. The process aimed to capture a diverse range of opinions without enforcing a forced consensus among the radiologists.

By conducting this two-phase image annotation process, the study aimed to maximise the identification of lung nodules while accommodating the inherent variability in radiologists' interpretations and opinions. This approach allowed for a comprehensive evaluation of the lung nodules present in each CT scan while acknowledging the importance of individual expertise and diverse perspectives within the radiology community.

1.6 Hyperparameters tuned while training the Faster R-CNN model.

- Number of GPUs
  cfg.SOLVER.REFERENCE WORLD SIZE = 2

- Number of data loading threads
  cfg.DATALOADER.NUM_WORKERS = 4

- Number of images per batch
  cfg.SOLVER.IMS_PER_BATCH = 4

- Base learning rate
  cfg.SOLVER.BASE_LR = 0.0125

- Number of iterations
  cfg.SOLVER.MAX_ITER = 100 #1500 No. of iterations

- Number of ROI heads batches per image
  cfg.MODEL.ROI_HEADS.BATCH_SIZE_PER_IMAGE = 256

- Number of classes
  cfg.MODEL.ROI_HEADS.NUM_CLASSES corresponds to the number of classes in each property (Table 2)

- No. of iterations after which the validation set is evaluated
  cfg.TEST.EVAL_PERIOD = 100

All other configurations are kept as default from Detectron2. Interested readers can refer to Detectron2's documentation page for further details about these default configurations.

1.7 Merging AIPS-N Scores with clinical factors

The AIPS-N scores are generated for the five nodule features including malignancy, margin, sphericity, spiculation, and texture. It is important to note that different nodules have different AIPS-N scores assigned to each feature, for example, a nodule could be assigned a score of 2 (moderately unlikely) for malignancy as shown in the following table.

Feature name	**Score = 1**	**Score = 2**	**Score = 3**	**Score = 4**	**Score = 5**
Sphericity	Linear	Ovoid/Linear	Ovoid	Ovoid/Round	NA
Margin	Poorly Defined	Near Poorly Defined	Medium Margin	Near Sharp	Sharp
Texture	Non-Solid/GGO	Non-Solid/Mixed	Part Solid/Mixed	Solid/Mixed	Solid
Malignancy	Highly Unlikely	Moderately Unlikely	Indeterminate	Moderately Suspicious	Highly Suspicious
Spiculation	No Spiculation	Nearly No Spiculation	Medium Spiculation	Near Marked Spiculation	Marked Spiculation

1.8 Imputation of missing categorical and numerical data

The AIPS-N scores of each nodule are combined with the patient's clinical factors which include age, gender, the status of smoking, and histology. The AIPS-N scores combined with the clinical factors of each patient, result in merged data with 9 input features as depicted below:

Nodule	Age#	Gender*	Smoking*	Histology*	Sphericity*	Margin*	Texture*	Malignancy*	Spiculation*
1	57	M	Non-smoker	Adenocarcinoma	2	5	2	1	5
2	66	F	Smoker	Squamous Cell Carcinoma	4	4	4	5	3

#Numerical variable
*Categorical variable

In our case, the missingness is not systematic or related to certain data characteristics (e.g., certain categories in the categorical variables or specific ranges in the numerical variables). Plus, the missingness appears to be random and not related to any observable factors, therefore it is reasonable to consider it as Missing Completely At Random (MCAR).

For MCAR data, using the mean value as an imputation strategy to impute numerical variables is reasonable because it preserves the central tendency of the variable. By filling in missing values with the mean, we maintain the overall average value, which helps to minimise the potential bias in subsequent analyses. For categorical data, imputing missing values with the value that appears most frequently (mode) is a logical choice. This is because the mode represents the most common category in the dataset. By assigning the mode to the missing values, we maintain the distribution and frequency of the categories, thus preserving the overall pattern and preventing any undue influence on subsequent analyses **(Haukoos et al., 2007)**.

1.9 Optimum value of deep learning parameters

- Number of hidden layers = 2
- Number of units = 1900
- Weights regularisation with the ‘l2’ regularisation function
- Bias regularisation with the 'l1_l2' regularisation function
- Weights constraint with max-norm having a value of 2
- Bias constraint with max-norm having a value of 1
- Weights initialization with ‘glorot_uniform'
- Learning rate = 0.01
- Stepsize = 150
- Decay rate = 0.2
- Stair = False
- Epsilon = 1e-07
- Tuner/epochs = 2
- Tuner/initial epoch = 0
- Tuner/bracket = 6
- Tuner/round = 0

1.10 Training, validation, and testing using Cohort 4

Merged data from 211 patients in Cohort 4 (White population) including the number of features, and the number of wild-type and mutant samples:

Total number of patients	**Number of input features**	**Wild-type samples**	**Mutant samples**
211	9	168	43

Cohort 4 with balanced classes after re-sampling of data using RandomOversampler **(Ghorbani et al., 2020):**

Before random oversampling		
Wild-type samples	Mutant samples	Total
168	43	211
After random oversampling		
168	168	336

Cohort 4 split into training, validation, and testing subsets:

Training subset		**Validation subset**		**Testing subset**	
Wild-type samples	Mutant samples	Wild-type samples	Mutant samples	Wild-type samples	Mutant samples
117	118	31	20	20	30

1.11 Techniques to avoid data leakage

Data leakage is a critical issue in machine learning training that can lead to overestimation of model performance and invalid results. To avoid data leakage, we used the following techniques:

1. Train-Validation Split: We ensured a proper separation between the training and validation datasets before any preprocessing or feature engineering. The validation set was only used for evaluation, and no information from it was used during training. Additionally, a testing set (holdout set)[1] was kept separate from the rest of the data and was only used to evaluate the model's performance after it had been trained.
2. Cross-Validation: We used K-fold cross-validation **(Xiong et al., 2020)** for evaluating the performance of the models by training them on multiple subsets of the data.
3. Feature Engineering: We conducted feature scaling and normalisation on the training set and then applied to the test set.
4. Data Imputation: If missing data imputation seemed necessary, we filled in missing values based on information from the training set only **(Haukoos et al., 2007)**.

1.12 Factors that may have contributed to the comparatively poorer precision and F1 score in Cohort 3, as compared to Cohort 2, for the AIPS-M ML model trained using Cohort 1

The cause could be due to data imbalance, data distribution shift, feature relevance, data quality, model, and domain shift. As per our analysis, there are two major differences between Cohort 2 (C2) [n = 591] and Cohort 3 (C3) [n = 96]. First, the total number of samples in Cohort 3 is very less than the number of samples in Cohort 2. Second, the presence of a class imbalance in Cohort 3, with unequal representation of positive and negative cases, can affect the precision and F1 score. A disproportionate number of instances in one class could lead to biased predictions and lower performance metrics.

Cohort	Number of positive samples	Number of negative samples	The ratio of positive and negative samples
Cohort 2	305	286	51.6 : 48.4
Cohort 3	27	69	28.13 : 71.87
C2 + C3	332	355	48.23 : 51.77

In a separate experiment, we combined validation Cohort 2 and Cohort 3 to form a balanced validation cohort, ensuring an equal representation of samples from both cohorts. The trained ML models were then validated using this merged cohort, and the obtained results are as follows:

Algorithm	AUC	Accuracy	F1 score	Precision	Recall
SVM	0.69	0.67	0.7	0.63	0.8
Random Forest	0.87	0.86	0.87	0.82	0.92
Decision Tree Classifier	0.84	0.84	0.84	0.79	0.91
XGB	0.86	0.86	0.86	0.8	0.92
Randomised Search CV	0.87	0.86	0.87	0.82	0.92
GridSearchCV	0.87	0.86	0.87	0.82	0.92

In the merged cohort, both the F1 score and precision exhibit notably higher values when compared to Cohort 3 and are comparable to the values observed in Cohort 2.

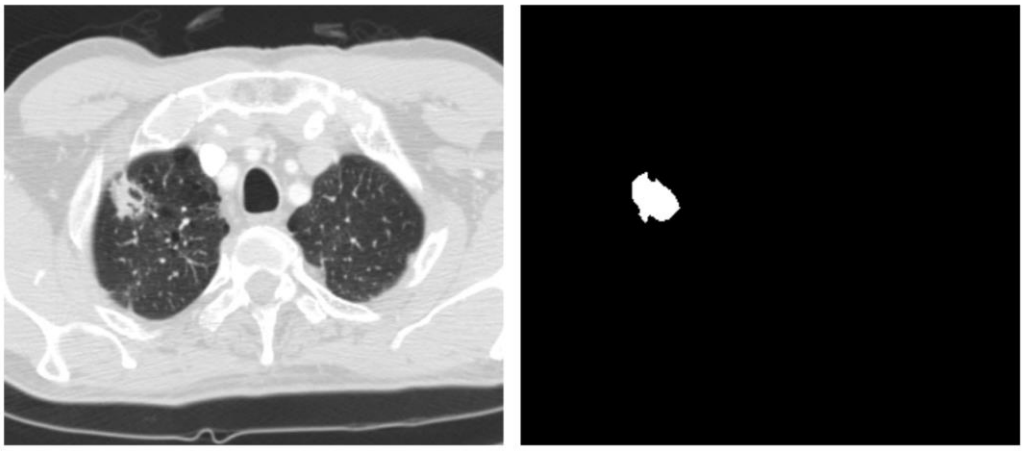
1.13 Rationale for developing AIPS-M models for the Indian and White populations separately

Developing AIPS-M models separately for the Indian and White populations is based on the understanding that population-specific characteristics and variations can influence model performance and generalizability. This approach ensures that the models are tailored to each population's unique features and improves accuracy. The decision to develop them separately for the Indian and White populations depends on factors such as data availability and potential population-specific differences in imaging features.

However, the AIPS-N was not developed separately for the Indian and White populations. The AIPS-Nodule (AIPS-N) model was developed using the LIDC-IDRI CT image dataset (Cohort 5) from TCIA (The Cancer Imaging Archive). This dataset consisted of 1010 patients and a total of 244,527 images. If there are sufficient annotated data and indications of population-specific variations, developing separate AIPS-N models can capture these nuances and enhance performance.

2. Supplementary figures

2.1 CT slice, corresponding mask and annotation



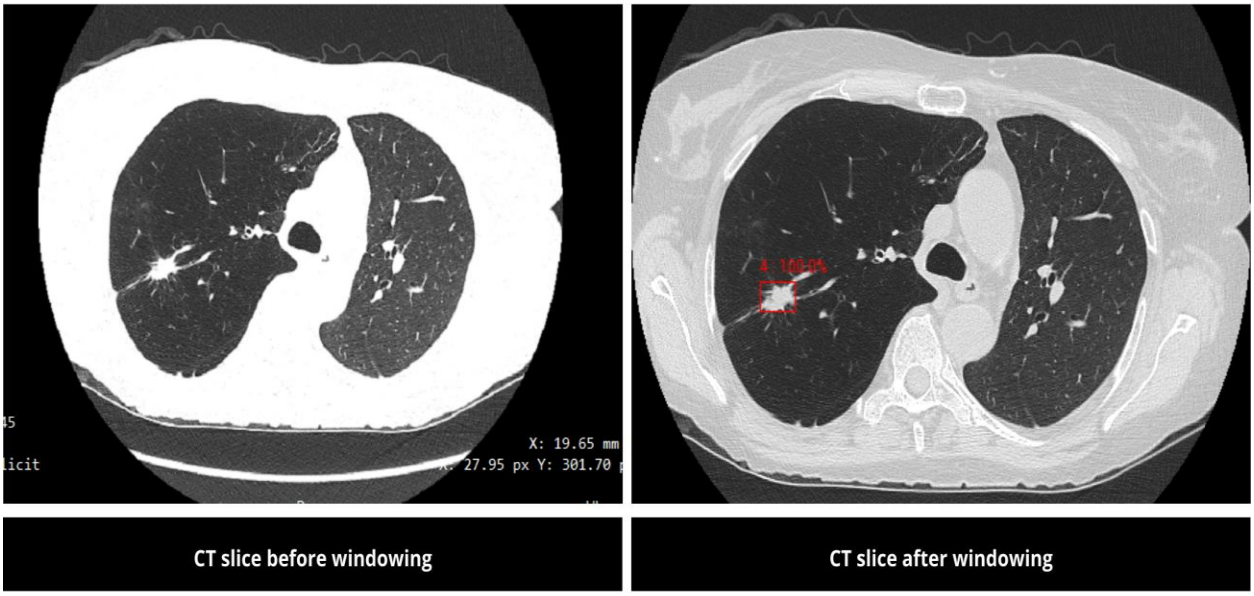
A B


```
{
    "sphericity": {"score": 4},"margin": {"score": 4},
    "spiculation": {"score": 4},"texture": {"score": 4},
    "malignancy": {"score": 3},
    "mask_px": [174,173,172,171,170,169,168,167,166,166,166,
        166,166,166,166,166,166,166,166,167,167,168,169,170,
        170,171,171,171,170,171,172,172,173,174,175,175,175,
        176,177,178,179,180,181,182,183,184,185,186,187,188,],
    "mask_py": [250,251,251,252,252,253,254,255,256,257,258,
        259,260,261,262,263,264,265,266,267,268,269,270,271,
        272,273,274,275,276,277,278,279,280,281,280,279,278,],
    "area": 573.0
}
```


C

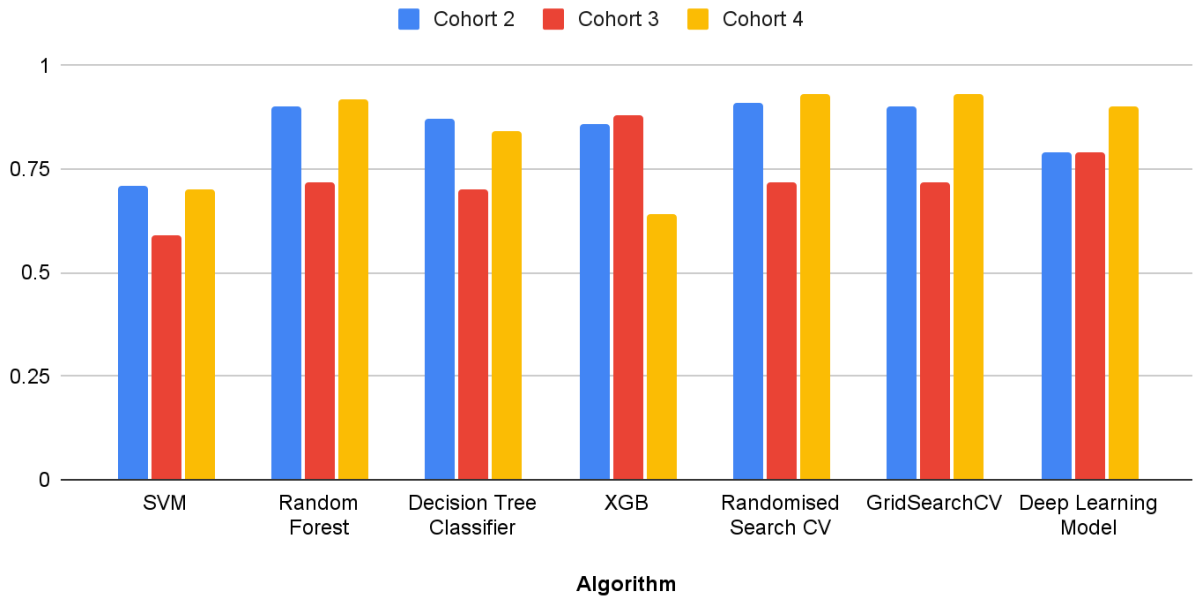
Supplementary Figure S1: The figure depicts the corresponding mask (B) and annotation (C) (JSON format) of a CT slice (A)

2.2 Contrast between a CT slice before and after preprocessing (windowing)



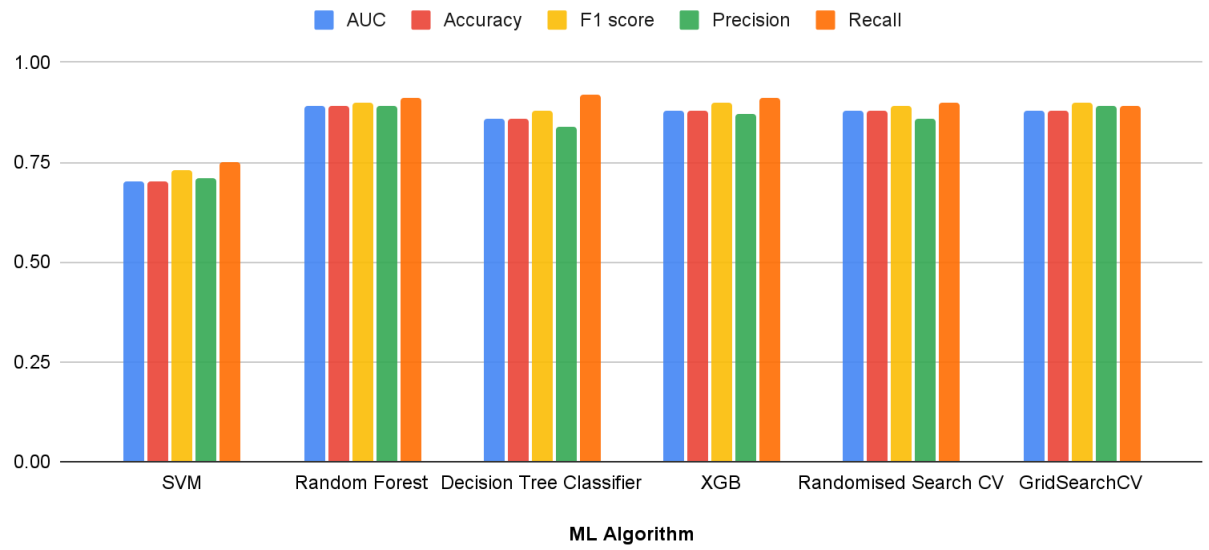
Supplementary Figure S2: The contrast between a CT image before preprocessing and after setting the window width to 1500 HU and window level to -500 HU for the evaluation of lung parenchyma

2.3 Diagram with combined results (AUC) of ML models and the DL model trained using Cohort 1



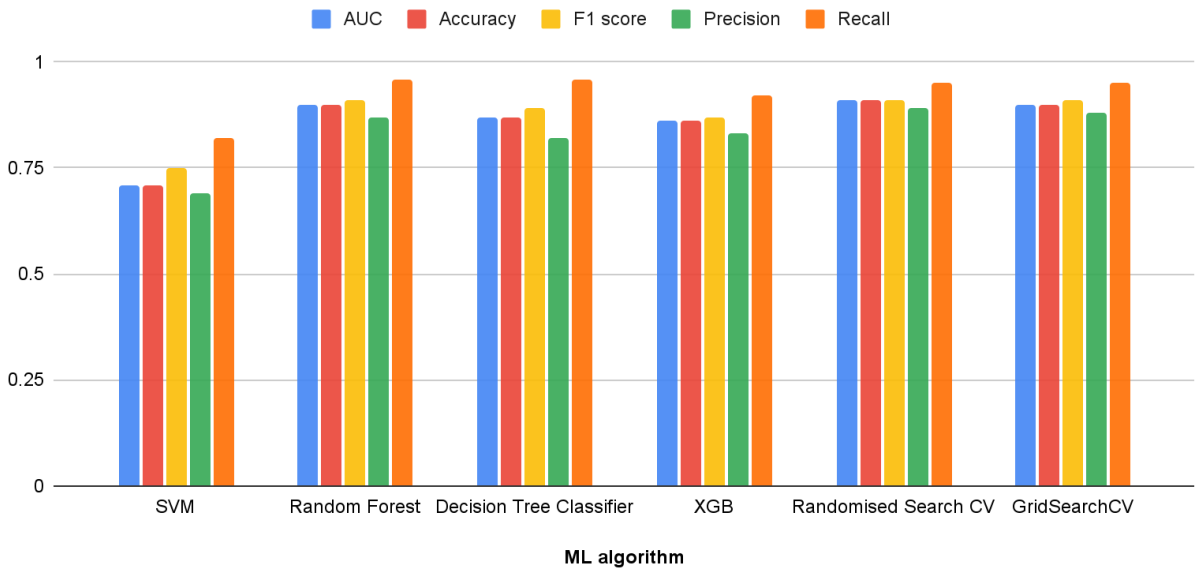
Supplementary Figure S3: Combined results (AUC metric) of ML models and the DL model trained using Cohort 1

2.4 Diagram with performance metrics obtained from all the machine learning algorithms trained using Cohort 1 (Indian population) on the validation subset



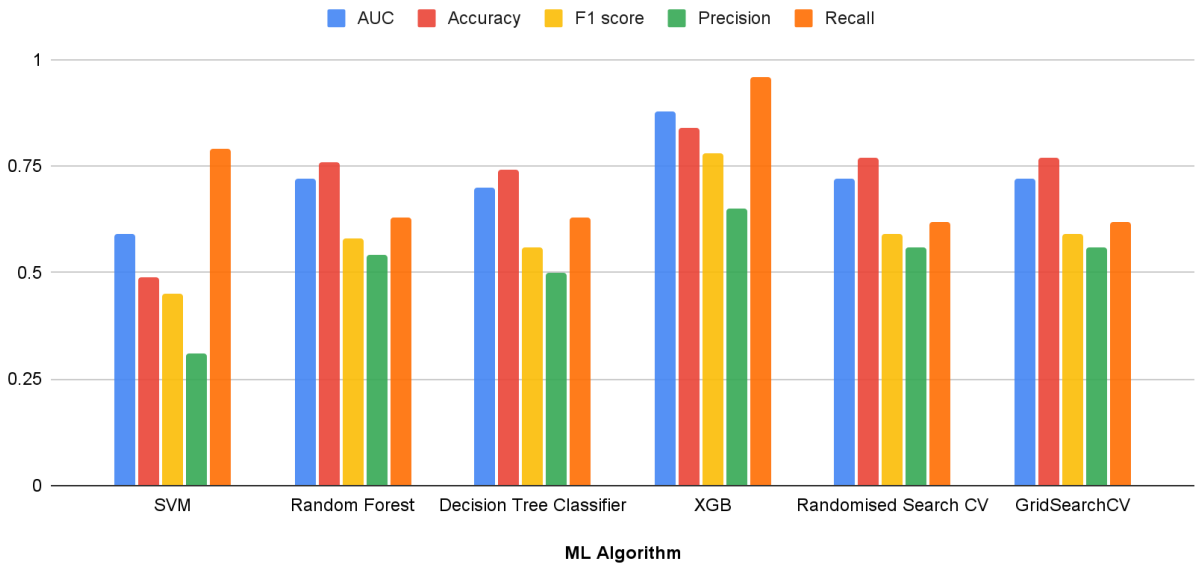
Supplementary Figure S4: Performance metrics obtained from all the machine learning algorithms trained using Cohort 1 (Indian population) on the validation subset.

2.5 Diagram with performance metrics obtained from all the machine learning (ML) algorithms on the testing Cohort 2, consisting of the Indian population



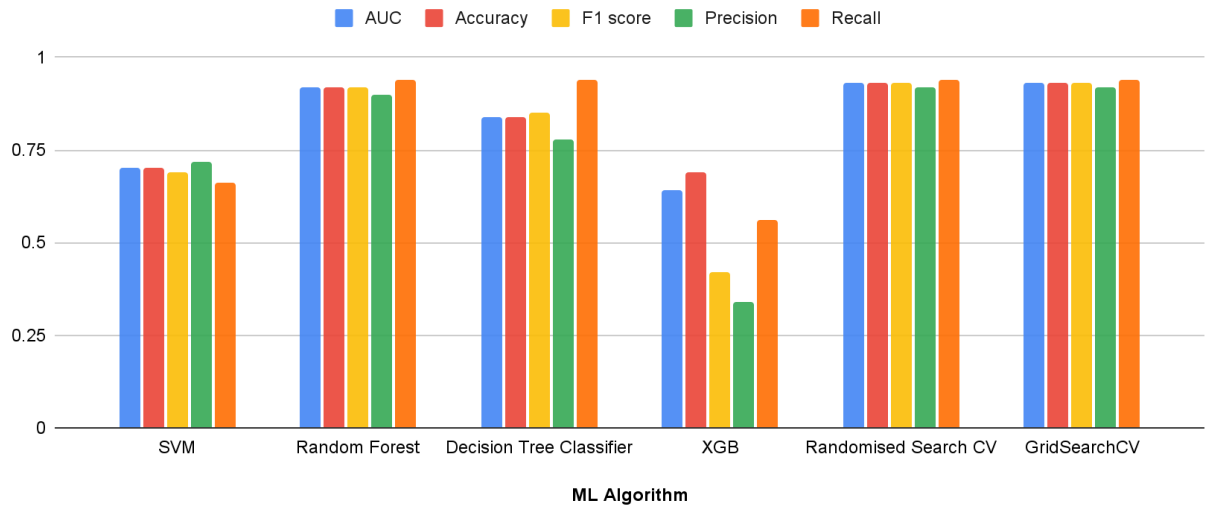
Supplementary Figure S5: Performance metrics obtained from all the machine learning (ML) algorithms on the testing Cohort 2, consisting of the Indian population.

2.6 Diagram with performance metrics obtained from all the machine learning (ML) algorithms applied to the testing Cohort 3, consisting of the Indian population.



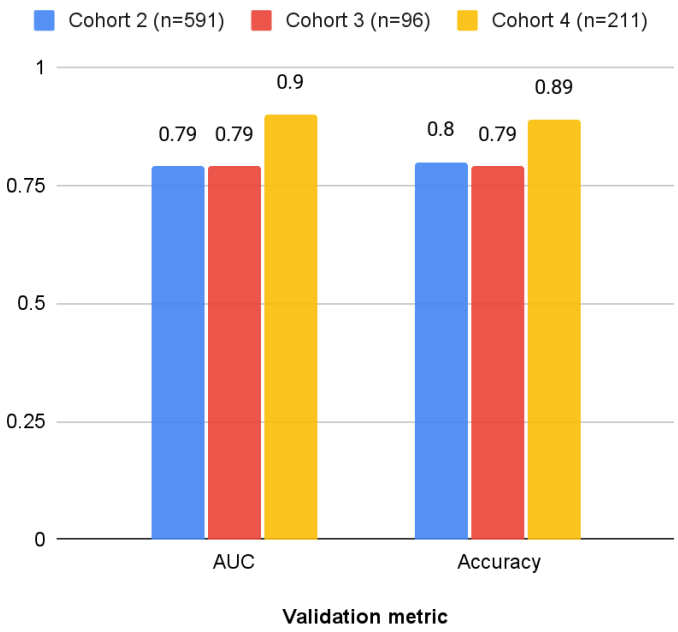
Supplementary Figure S6: Performance metrics obtained from all the machine learning (ML) algorithms applied to the testing Cohort 3, consisting of the Indian population.

2.7 Diagram with testing metrics obtained from all the machine learning (ML) models applied to the validation Cohort 4, consisting of the White population



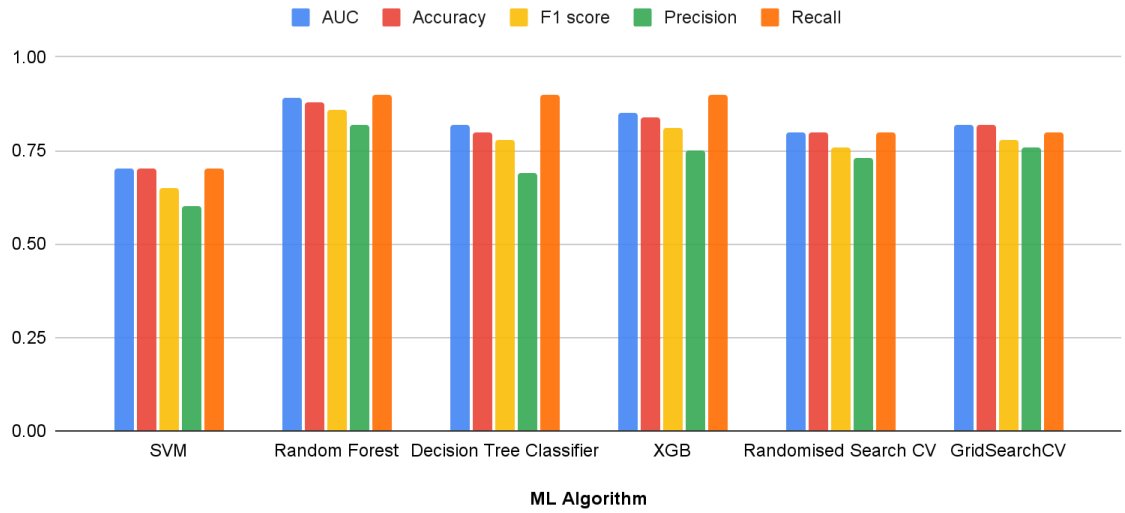
Supplementary Figure S7: Testing metrics obtained from all the machine learning (ML) models applied to the validation Cohort 4, consisting of the White population. This evaluation allowed us to assess and compare the performance of the ML algorithms in predicting outcomes specifically within the White population.

2.8 Diagram with performance metrics obtained from the DL model applied to testing Cohort 2 and Cohort 3, consisting of the Indian population and Cohort 4, consisting of the White population



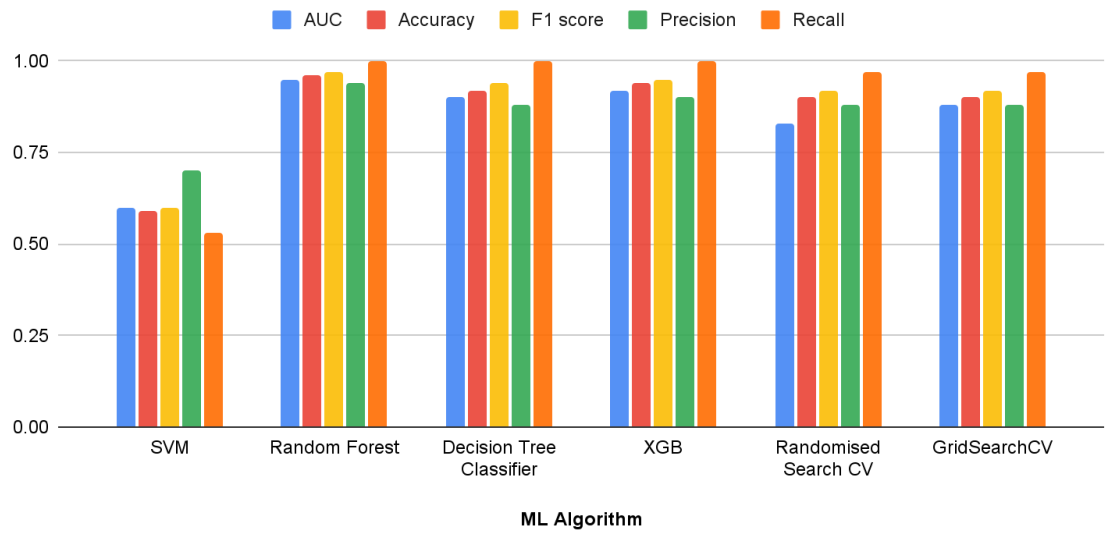
Supplementary Figure S8: Performance metrics obtained from the DL model applied to testing Cohort 2 and Cohort 3, consisting of the Indian population and Cohort 4, consisting of the White population.

2.9 Diagram with performance metrics obtained from all the machine learning (ML) algorithms trained using Cohort 4 (White population) on the validation subset



Supplementary Figure S9A: Performance metrics obtained from all the machine learning (ML) algorithms trained using Cohort 4 (White population) on the validation subset

2.10 Diagram with performance metrics obtained from all the machine learning (ML) algorithms trained using Cohort 4 (White population) on the testing subset

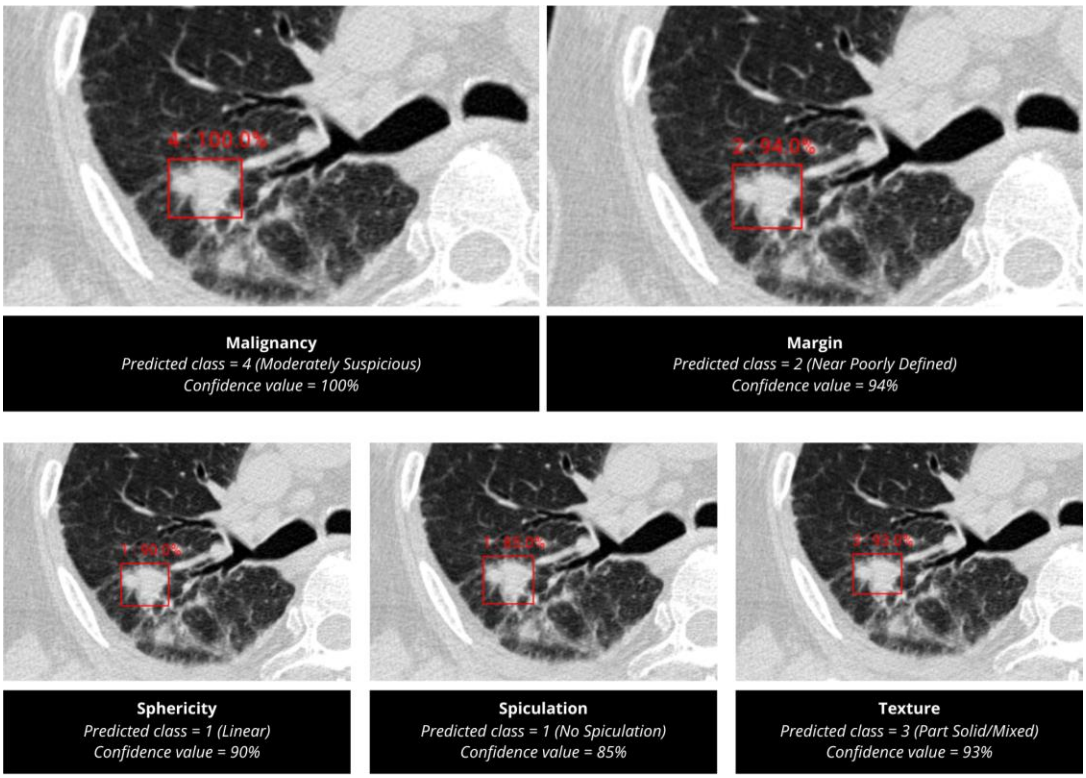


Supplementary Figure S9B: Performance metrics obtained from all the machine learning (ML) algorithms trained using Cohort 4 (White population) on the testing subset

2.11 Diagram with performance metrics obtained from the DL algorithm trained using Cohort 4 (White population) on the validation and testing subsets

Supplementary Figure S10: Performance metrics obtained from the DL algorithm trained using Cohort 4 (White population) on the validation and testing subsets

2.12 A magnified version of the predictions made by the AIPS-N model



Supplementary Figure S11: A magnified version of the predictions made by the AIPS-N model. The zoom level, which refers to the degree to which an image or object is enlarged compared to its original size, is set at 2x.

3. Supplementary Tables

3.1 Division of images and annotations into training, validation, and testing subsets for each feature.

Sr. No.	Feature	Total images	Images in the training subset	Images in the validation subset	Images in the testing subset
1.	Malignancy	3817	2671	573	573
2.	Margin	3817	2671	573	573
3.	Sphericity	3817	2671	573	573
4.	Spiculation	3817	2671	573	573
5.	Texture	3817	2671	573	573

3.2 The number of images in the training subset were balanced according to the class with the fewest images to mitigate the influence of class imbalance on the model's performance

Feature name	Total images	Images in the training subset	Images in the validation subset	Images in the testing subset
Malignancy	1501	355	573	573
Margin	1626	480	573	573
Spiculation	1451	305	573	573
Sphericity	1978	832	573	573
Texture	1596	450	573	573

3.3 Merged data from 1379 Indian patients in Cohort 1 including the number of features, and the number of wild-type and mutant samples.

Total number of patients	Number of input features	Wild-type samples	Mutant samples
1379	9	699	680

3.4 Cohort 1 with balanced classes after re-sampling of data using RandomOversampler

Before using RandomOversampler (Ghorbani et al., 2020)		
Wild-type samples	Mutant samples	Total
699	680	1379
After using RandomOversampler (Ghorbani et al., 2020)		
699	699	1398

3.5 Cohort 1 split into training and validation subsets

Training subset		**Validation subset**	
Wild-type samples	Mutant samples	Wild-type samples	Mutant samples
489	489	210	210

3.6 Performance metrics of the AIPS-M ML models trained using Cohort 1 (Indian population) in predicting *EGFR* genotype in the validation subset

Algorithm	AUC	AUC 95% CI*	Accuracy	Accuracy 95% CI*	F1 score	F1 score 95% CI*	Precision	Precision 95% CI*	Recall	Recall 95% CI*
SVM **(Morgado et al., 2021)**	0.7	0.64 to 0.76	0.7	0.64 to 0.76	0.73	0.67 to 0.79	0.71	0.64 to 0.78	0.75	0.69 to 0.81
Random Forest **(Jia et al., 2019)**	0.89	0.85 to 0.92	0.89	0.85 to 0.92	0.9	0.88 to 0.92	0.89	0.87 to 0.91	0.91	0.89 to 0.93
Decision Tree Classifier **(Anzar et al., 2019)**	0.86	0.82 to 0.90	0.86	0.82 to 0.90	0.88	0.86 to 0.90	0.84	0.81 to 0.87	0.92	0.90 to 0.94
XGB **(Morgado et al., 2021)**	0.88	0.85 to 0.91	0.88	0.85 to 0.91	0.9	0.88 to 0.92	0.87	0.84 to 0.90	0.91	0.89 to 0.93
Randomised Search CV	0.88	0.85 to 0.91	0.88	0.85 to 0.91	0.89	0.88 to 0.91	0.86	0.84 to 0.88	0.9	0.89 to 0.91
GridSearchCV **(Ventura et al., 2021)**	0.88	0.85 to 0.91	0.88	0.85 to 0.91	0.9	0.88 to 0.92	0.89	0.87 to 0.91	0.89	0.88 to 0.90

*Confidence Interval

3.7 Performance metrics of the AIPS-M ML in predicting *EGFR* genotype in the testing Cohort 2 (Indian population)

Algorithm	AUC	AUC 95% CI*	Accuracy	Accuracy 95% CI*	F1 score	F1 score 95% CI*	Precision	Precision 95% CI*	Recall	Recall 95% CI*
SVM **(Morgado et al., 2021)**	0.71	0.67 to 0.74	0.71	0.67 to 0.74	0.75	0.72 to 0.78	0.69	0.64 to 0.74	0.82	0.77 to 0.87
Random Forest **(Jia et al., 2019)**	0.9	0.84 to 0.94	0.9	0.84 to 0.94	0.91	0.88 to 0.94	0.87	0.84 to 0.90	0.96	0.93 to 0.99
Decision Tree Classifier **(Anzar et al., 2019)**	0.87	0.84 to 0.89	0.87	0.84 to 0.89	0.89	0.86 to 0.92	0.82	0.78 to 0.86	0.96	0.93 to 0.99
XGB **(Morgado et al., 2021)**	0.86	0.83 to 0.89	0.86	0.83 to 0.89	0.87	0.84 to 0.90	0.83	0.81 to 0.85	0.92	0.90 to 0.94
Randomised Search CV	0.91	0.88 to 0.94	0.91	0.88 to 0.94	0.91	0.88 to 0.94	0.89	0.86 to 0.92	0.95	0.93 to 0.97
GridSearchCV **(Ventura et al., 2021)**	0.90	0.87 to 0.93	0.90	0.87 to 0.93	0.91	0.88 to 0.94	0.88	0.85 to 0.91	0.95	0.93 to 0.97

*Confidence Interval

3.8 Performance metrics of the AIPS-M ML in predicting *EGFR* genotype in the testing Cohort 3 (Indian population).

Algorithm	AUC	AUC 95% CI*	Accuracy	Accuracy 95% CI*	F1 score	F1 score 95% CI*	Precision	Precision 95% CI*	Recall	Recall 95% CI*
SVM **(Morgado et al., 2021)**	0.59	0.52 to 0.66	0.49	0.40 to 0.58	0.45	0.36 to 0.54	0.31	0.23 to 0.39	0.79	0.70 to 0.88
Random Forest **(Jia et al., 2019)**	0.72	0.67 to 0.77	0.76	0.72 to 0.80	0.58	0.53 to 0.63	0.54	0.48 to 0.59	0.63	0.57 to 0.69
Decision Tree Classifier **(Anzar et al., 2019)**	0.7	0.60 to 0.70	0.74	0.68 to 0.79	0.56	0.51 to 0.61	0.5	0.44 to 0.56	0.63	0.57 to 0.69
XGB **(Morgado et al., 2021)**	0.88	0.85 to 0.91	0.84	0.81 to 0.87	0.78	0.75 to 0.81	0.65	0.62 to 0.68	0.96	0.92 to 0.99
Randomised Search CV	0.72	0.67 to 0.77	0.77	0.73 to 0.81	0.59	0.54 to 0.64	0.56	0.50 to 0.62	0.62	0.58 to 0.66
GridSearchCV **(Ventura et al., 2021)**	0.72	0.67 to 0.77	0.77	0.73 to 0.81	0.59	0.54 to 0.64	0.56	0.50 to 0.62	0.62	0.58 to 0.66

*Confidence Interval

3.9 Performance metrics of the AIPS-M ML in predicting *EGFR* genotype in the testing Cohort 4 (White population)

Algorithm	AUC	AUC 95% CI*	Accuracy	Accuracy 95% CI*	F1 score	F1 score 95% CI*	Precision	Precision 95% CI*	Recall	Recall 95% CI*
SVM **(Morgado et al., 2021)**	0.70	0.62 to 0.78	0.70	0.63 to 0.77	0.69	0.61 to 0.77	0.72	0.64 to 0.79	0.66	0.59 to 0.73
Random Forest **(Jia et al., 2019)**	0.92	0.88 to 0.96	0.92	0.89 to 0.95	0.92	0.89 to 0.95	0.90	0.87 to 0.93	0.94	0.91 to 0.97
Decision Tree Classifier **(Anzar et al., 2019)**	0.84	0.80 to 0.88	0.84	0.81 to 0.87	0.85	0.82 to 0.88	0.78	0.74 to 0.82	0.94	0.91 to 0.97
XGB **(Morgado et al., 2021)**	0.64	0.59 to 0.69	0.69	0.66 to 0.72	0.42	0.37 to 0.47	0.34	0.31 to 0.37	0.56	0.52 to 0.60
Randomised Search CV	0.93	0.90 to 0.96	0.93	0.90 to 0.96	0.93	0.90 to 0.96	0.92	0.89 to 0.95	0.94	0.91 to 0.97
GridSearchCV **(Ventura et al., 2021)**	0.93	0.90 to 0.96	0.93	0.90 to 0.96	0.93	0.90 to 0.96	0.92	0.89 to 0.95	0.94	0.91 to 0.97

*Confidence Interval

3.10 Performance metrics of the AIPS-M deep learning (DL) model in predicting *EGFR* genotype in Cohorts 2 and 3 (Indian population) and Cohort 4 (White population)

	AUC	Accuracy
Cohort 2 (n=591)	0.79	0.8
Cohort 3 (n=96)	0.79	0.79
Cohort 4 (n=211)	0.9	0.89

3.11 Training and testing of ML and DL algorithms using only the clinical factors to evaluate their performance compared to models trained with both clinical factors and AIPS-N scores

Performance metric of models trained using Cohort 1 (Indian population)	With clinical factors only	With clinical factors and AIPS-N scores
Average AUC value of ML models on the validation subset	0.73	0.85
Average AUC value of ML models on the testing Cohort 2 (Indian population)	0.72	0.86
Average AUC value of ML models on the testing Cohort 3 (Indian population)	0.64	0.72
Average AUC value of ML models on the testing Cohort 4 (White population)	0.6	0.83
AUC value of the DL model on the validation subset	0.79	0.86
AUC value of the DL model on the testing Cohort 2 (Indian population)	0.7	0.79
AUC value of the DL model on the testing Cohort 3 (Indian population)	0.57	0.79
AUC value of the DL model on the testing Cohort 4 (White population)	0.61	0.9

3.12 Performance metrics obtained from all the machine learning (ML) algorithms trained using Cohort 4 (White population) on the validation subset

Algorithm	AUC	AUC 95% CI*	Accuracy	Accuracy 95% CI*	F1 score	F1 score 95% CI*	Precision	Precision 95% CI*	Recall	Recall 95% CI*
SVM **(Morgado et al., 2021)**	0.7	0.63 to 0.77	0.7	0.57 to 0.74	0.65	0.51 to 0.70	0.6	0.44 to 0.65	0.7	0.57 to 0.73
Random Forest **(Jia et al., 2019)**	0.89	0.86 to 0.92	0.88	0.84 to 0.91	0.86	0.82 to 0.90	0.82	0.78 to 0.86	0.9	0.87 to 0.93
Decision Tree Classifier **(Anzar et al., 2019)**	0.82	0.78 to 0.86	0.8	0.75 to 0.85	0.78	0.73 to 0.83	0.69	0.60 to 0.77	0.9	0.87 to 0.93
XGB **(Morgado et al., 2021)**	0.85	0.80 to 0.90	0.84	0.78 to 0.90	0.81	0.76 to 0.86	0.75	0.70 to 0.80	0.9	0.86 to 0.94
Randomised Search CV	0.8	0.77 to 0.83	0.8	0.76 to 0.84	0.76	0.73 to 0.80	0.73	0.71 to 0.76	0.8	0.81 to 0.89
GridSearchCV **(Ventura et al., 2021)**	0.82	0.78 to 0.86	0.82	0.77 to 0.86	0.78	0.75 to 0.83	0.76	0.73 to 0.79	0.8	0.82 to 0.89

*Confidence Interval

3.13 Performance metrics obtained from all the machine learning (ML) algorithms trained using Cohort 4 (White population) on the testing subset

Algorithm	AUC	AUC 95% CI*	Accuracy	Accuracy 95% CI*	F1 score	F1 score 95% CI*	Precision	Precision 95% CI*	Recall	Recall 95% CI*
SVM **(Morgado et al., 2021)**	0.6	0.43 to 0.57	0.59	0.49 to 0.69	0.6	0.49 to 0.61	0.7	0.57 to 0.73	0.53	0.38 to 0.68
Random Forest **(Jia et al., 2019)**	0.95	0.92 to 0.98	0.96	0.93 to 0.99	0.97	0.94 to 0.99	0.94	0.91 to 0.97	1	0.97 to 1.00
Decision Tree Classifier **(Anzar et al., 2019)**	0.9	0.85 to 0.95	0.92	0.88 to 0.96	0.94	0.90 to 0.98	0.88	0.83 to 0.93	1	0.97 to 1.00
XGB **(Morgado et al., 2021)**	0.92	0.88 to 0.96	0.94	0.90 to 0.98	0.95	0.92 to 0.98	0.9	0.86 to 0.94	1	0.97 to 1.00
Randomised Search CV	0.83	0.81 to 0.87	0.9	0.86 to 0.94	0.92	0.89 to 0.95	0.88	0.84 to 0.92	0.97	0.94 to 0.99
GridSearchCV **(Ventura et al., 2021)**	0.88	0.83 to 0.91	0.9	0.87 to 0.93	0.92	0.89 to 0.95	0.88	0.84 to 0.92	0.97	0.94 to 0.99

3.14 Performance metrics obtained from the deep learning (DL) algorithm trained using Cohort 4 (White population) on the validation and testing subsets

	AUC	Accuracy
Validation subset	0.9	0.82
Testing subset	0.86	0.88

Table 3.15: Predictions of the AIPS-M model

Sr. No.	Algorithm	Prediction
1.	Support Vector Machine (SVM)	True Positive
2.	Random Forest	True Positive
3.	Decision Tree Algorithm	True Positive
4.	Grid Search Cross-Validation	True Positive
5.	Randomised Search Cross-Validation	True Positive

4. Supplementary references

1. Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, Zheng H, Benson JA, Zhang W, Leung AN, Kadoch M. A radiogenomic dataset of non-small cell lung cancer. Scientific data. 2018 Oct 16;5(1):1-9.

2. Armato III SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics. 2011 Feb;38(2):915-31.

3. Haukoos JS, Newgard CD. Advanced statistics: missing data in clinical research—part 1: an introduction and conceptual framework. Academic Emergency Medicine. 2007 Jul;14(7):662-8.

4. Ghorbani R, Ghousi R. Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. IEEE Access. 2020;8:67899–911.

5. Xiong Z, Cui Y, Liu Z, Zhao Y, Hu M, Hu J. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. Computational Materials Science. 2020 Jan 1;171:109203.

6. Morgado J, Pereira T, Silva F, Freitas C, Negrão E, de Lima BF, et al. Machine Learning and Feature Selection Methods for EGFR Mutation Status Prediction in Lung Cancer. Applied Sciences. 2021 Apr 6;11(7):3273.

7. Jia TY, Xiong JF, Li XY, Yu W, Xu ZY, Cai XW, et al. Identifying EGFR mutations in lung adenocarcinoma by noninvasive imaging using radiomics features and random forest modeling. Eur Radiol. 2019 Sep;29(9):4742–50.

8. Anzar I, Sverchkova A, Stratford R, Clancy T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. BMC Med Genomics. 2019 Dec;12(1):63.

9. Ventura A, Pereira T, Silva F, Freitas C, Cunha A, Oliveira HP. Stacking Approach for Lung Cancer EGFR Mutation Status Prediction from CT Scans. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) [Internet]. Houston, TX, USA: IEEE; 2021 [cited 2022 Oct 13]. p. 3099–105. Available from: https://ieeexplore.ieee.org/document/9669429/

10. Warrens MJ. Five ways to look at Cohen's kappa. Journal of Psychology & Psychotherapy. 2015 Jul 28;5.