Automatic Interpretation of Line Probe Assay Test for Tuberculosis

Jatin Agrawal[1*], Mukul Kumar[1*], Avtansh Tiwari[1*], Sachin Danisetty[1], Soma Dhavala[1], Nakul Jain[1], Prasaanth Balraj[1], Niket Singh[1], Siddhant Shingi[1], Jayakrishna Kurada[1], Raghuram Rao[2], S Anand[2], Nishant Kumar[2]

[1]Wadhwani Institute for Artificial Intelligence
[2]Central Tuberculosis Division, Government of India
{jatin,mukul,sachin,soma,nakul,prasaanth,niket}@wadhwaniai.org,
{raor,anands,kumarn}@rntcp.org

Abstract

Line Probe Assay (LPA) is a widely used method for diagnosing drug-resistant tuberculosis (DRTB), but it is a time-consuming and labor-intensive process that requires expert interpretation. DRTB is a significant threat to global TB control efforts and its prompt diagnosis is critical for initiating appropriate treatment. In this paper, we present an automated LPA test interpretation solution that uses computer vision techniques to extract and analyze strips from LPA sheets and uses machine learning algorithms to produce drug sensitivity and resistivity outcomes with extremely high precision and recall. We also develop OCR models to eliminate manual data entry to further reduce the overall time. Our solution comprises a rejection module that flags ambiguous and novel samples that are then referred to experienced lab technicians. This results in increased trust in the solution. To evaluate our solution, we curate an extensive and diverse dataset of LPA strips annotated by multiple microbiologists across India. Our solution achieves more than 95% accuracy for all drugs on this dataset. The proposed solution has the potential to increase the efficiency, standardization of LPA test interpretation, and fast-tracking the dissemination of results to end-users via a designated Management Information System (MIS).

Introduction

Tuberculosis (TB) is a bacterial infection caused by Mycobacterium tuberculosis and is a major public health concern worldwide. In 2019, there were an estimated 10 million cases of TB, and 1.4 million people died from the disease (WHO, 2020). The LPA testing process involves a complex interaction between a series of probes (reagents) and specific segments of TB bacterial DNA called genotypes (wild types and mutations) which are visible as dark bands of varying intensities against a white background. Each patient's LPA test result is captured over a special 64x3mm paper strip with 27 bands each representing a unique band class, which is interpreted by lab technicians and microbiologists to infer sensitivity to various drugs. LPA test strips for each patient are serially pasted onto an LPA sheet and are interpreted manually in the lab using a reference reading strip. The sample sheet can be seen in Figure 1.

There are two types of LPA tests: i) First Line (FL) and ii) Second Line (SL). FL-LPA picks up drug resistance to First Line Anti TB drugs such as Isoniazid and Rifampicin while SL-LPA picks up drug resistance to Second line Anti TB drugs such as Fluoroquinolones and second line injectables. According to the guidelines of the National Tuberculosis Elimination Programme (NTEP) of India, all micro-biologically confirmed pulmonary TB patients should be tested for drug sensitivity using LPA. Approximately 1,000,000 patients should be tested for LPA. At present, 400,000 tests are performed every year across 64 Culture and drug Sensitivity Test Labs (CDST).

Existing practice requires lab technicians and microbiologists to review each strip, determine the bands present in the strip, and apply a set of rules to infer sensitivity to the corresponding drugs. In our study, we found that it takes up to an hour for a lab technician to read, interpret, and tabulate an LPA sheet containing 24 strips. Through this solution, we show that this process can be automated by using machine learning. We develop a novel system that:

- Automates the interpretation of LPA sheets.
- Improves overall efficiency, standardizes LPA results, and fast-tracks the dissemination of results to the end user via designated MIS.
- Achieves extremely high precision and recall for all classes on a gold standard dataset.
- Mitigates biases and improves trustworthiness by introducing human in the loop.

Related Works

Automation in Tuberculosis Diagnosis

Efforts to automate TB diagnosis have gained traction in recent years. Various automated techniques have been explored to improve the speed, accuracy, and scalability of TB diagnostic processes (Panicker et al. 2015). Automated microscopy techniques, using image processing techniques have been explored to automate various diagnosis tests in TB, and have shown promising results in reducing the subjectivity and time required for TB detection (Lopez-Garnier, Sheen, and Zimic 2019).

*These authors contributed equally.

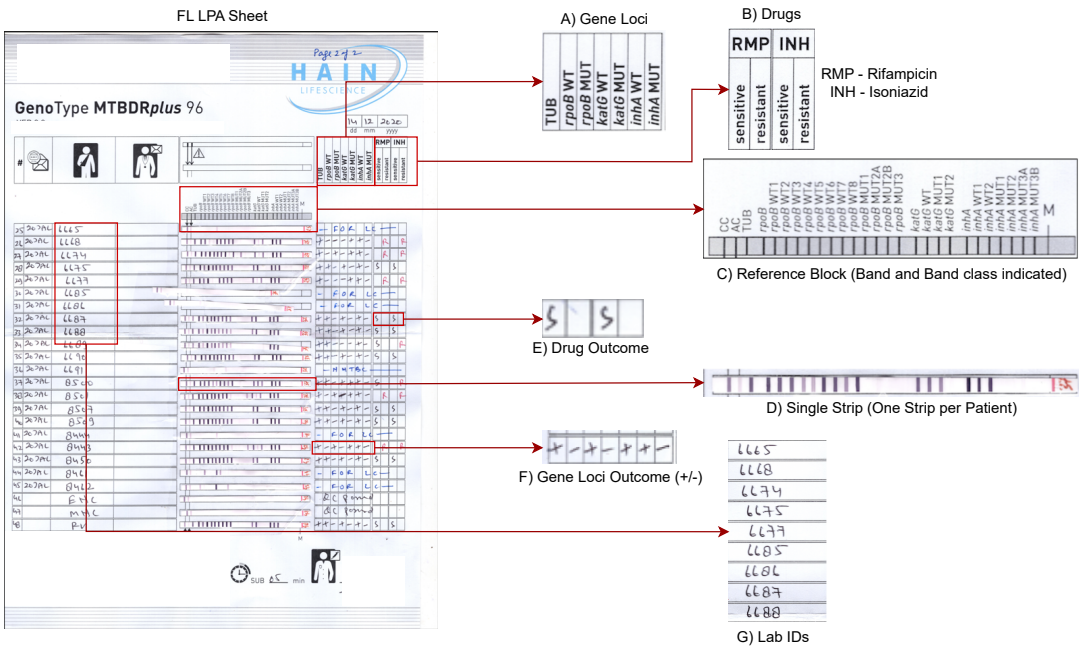

Figure 1: A sample of an FL LPA Sheet. Multiple strips are pasted on a single sheet. We highlight parts such as: A) Gene Loci names, B) Drug names, C) Reference block with reference reading strip, D) A patient strip containing 18 bands, E) Drug Outcomes R/S for Resistive and Sensitive respectively, F) Gene Loci outcomes (+/-) and, G) A column with Lab Ids

AI and Computer Vision in Medical Diagnosis

Integrating Artificial Intelligence (AI) and Computer Vision in medical diagnosis has revolutionized disease detection and treatment. AI techniques, particularly deep learning, have demonstrated remarkable capabilities in analyzing medical images and detecting anomalies (Kvak et al. 2023). For instance in radiology, these algorithms have helped classify and detect various abnormalities in Chest X-rays (Plesner et al. 2023). Object detection, a task of Computer Vision, plays a pivotal role in localizing and classifying objects of interest within images (Yang and Yu 2021), making it highly relevant for automating diagnostic tests like LPA. The application of Object Detection to automate LPA in TB diagnosis remains relatively unexplored. Commonly used object detection models are Faster R-CNN (Ren et al. 2016), YOLO family (Redmon et al. 2016), and SSD (Liu et al. 2016). Recently transformers which were very successful in Natural Language Processing (NLP) found a place in vision tasks and led to the creation of models like Detection Transformers (DETR) (Carion et al. 2020). In this solution, we compared multiple object detection models on the task of band detection in LPA strips.

Human in the Loop in AI applications

To build a reliable AI application, especially in a high-risk domain like Health, it is essential to leverage human expertise and also promote human interaction (Jotterand and Bosco 2020). On the other hand, it is also essential to reduce human effort by referring to only ambiguous or novel samples. (Hendrickx et al. 2023) showcased how different rejection approaches can be used to address this issue. In our system, we take inspiration from this line of work and use a rejection model that sends ambiguous or novel samples to expert lab technicians for further review - empowering them to override erroneous outputs produced by the system.

Dataset

Our dataset comprises scanned images of LPA sheets collected from various TB laboratories across India. The dataset was created by the Central TB Division (CTD) of India. We received a total of 1200 LPA sheets, out of which 810 belonged FL test and 390 belonged to the SL test. Each LPA sheet can have a maximum of 24 strips pasted on it, each strip refers to a test for one patient. We extract the strips from each sheet and our strip dataset consists of 13482 FL strips and 1746 SL strips.

Understanding LPA Sheet and Strips

As shown in Figure 1, each sheet has a few landmarks such as a doctor icon on the top left, a column to write Lab ID, a reference reading strip to identify the band type pasted on the top, and a boxed column to store gene loci and drug-level annotations. These sheets differ in FL and SL since they have different numbers of gene loci and drugs. Each sheet can have a maximum of 24 strips pasted in the designated location as shown in Figure 1. Each strip can have a maximum of 27 bands. Since each band can be present or absent, 2^{27} configurations are possible, but only around 400 of them are seen in practice.

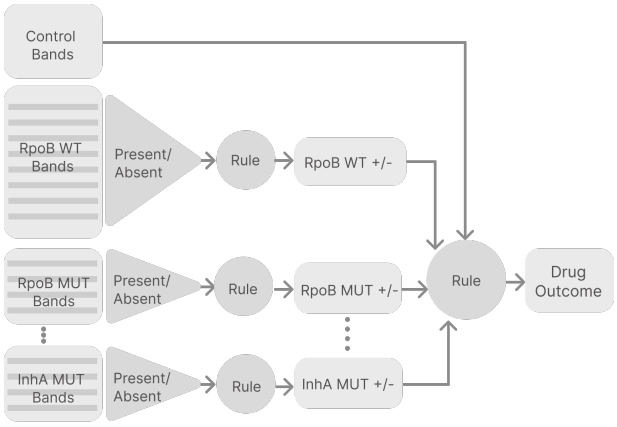

Figure 2: FL LPA Sheet interpretation from Bands to Drug Outcomes. The band's presence or absence leads to loci being positive or negative. That in turn leads to drug outcomes (Resistive/Sensitive/Invalid).

	First Line (FL)		Second Line (SL)	
	Sheets	Strips	Sheets	Strips
Train	650	10841	310	1406
Test	160	2641	80	340

Table 1: Train - Test distribution of the FL and SL Real annotated dataset

Both FL and SL start with CC and AC bands that are used for quality control. The TUB band determines if the person has TB or not. In FL, rpoB, KatG, and InhA bands are the control bands for the Loci RpoB, KatG, and InhA respectively. The reference block in Figure 1 showcases these bands. The remaining bands are used to determine the positive or negative outcome for the gene loci (RpoB WT, RpoB MUT, KatG WT, KatG MUT, InhA WT, and InhA MUT), ultimately determining the resistivity and sensitivity of the drug. The interpretation of strips from bands to drug outcomes can be understood from Figure 2. Additional details can be found at stoptb.org.[1]

Ground Truth Annotations

For all the 1200 LPA sheets we had the following annotations:

- Bounding boxes around each band along with band class for each strip on the sheet.
- Gene Loci (positive/negative) annotations for each strip.
- Drug (resistive/sensitive/NA) annotations for each drug in the strip.

The annotation exercise was extremely tedious as it involved careful examination of each strip. These annotations were done by highly skilled lab technicians across the country. The dataset distribution is shown in the Table 1.

[1]https://stoptb.org/wg/gli/assets/documents/LPA_test_web_ready.pdf

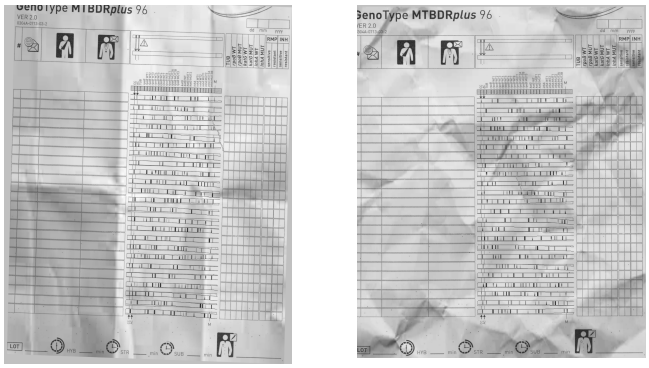

Figure 3: Simulated sheet images. Left sheet (A) is an example of a simulated sheet with folds and right sheet (B) is an example of a simulated sheet with crumples

Simulated Dataset

As explained earlier, developing an accurate benchmark ground truth dataset for this task is both labor-intensive and costly. To reduce the dependency on real datasets, we created a simulated dataset. An LPA sheet without any strips pasted on it always follows a fixed template. As a first step, we got a digital copy of this template. To enhance realism, we incorporated features such as creases, wrinkles, and crumpled textures into the sheet images. We did this by superimposing a wrinkled blank paper image onto the image of the LPA sheet with the help of OpenCV[2] library. We also add a few augmentations like illumination variance, contrast, hue, and brightness shifts.

After creating the simulated sheet, the next step was to simulate strips and place them on the sheet at a specified region. It is easy to imitate bands as they are black bars with varying intensity. We randomly put bands on a blank strip creating a synthetic strip. Then we place this strip on the simulated sheet. Placing multiple strips on a sheet finally gives us a complete simulated sheet. Samples of simulated sheets can be seen in Figure 3.

An important characteristic of the LPA testing that helped us in creating this dataset was that the drug outcomes are dependent on the presence or absence of bands. The rules are deterministic and can be easily codified. Hence, once we fix the band arrangement, we could immediately deduce the corresponding drug and loci outcomes, effectively eliminating the need for manual annotations. There were multiple advantages of having such a dataset. We used it to pretrain our models. It saved us time and money in annotating real datasets and allowed us to create samples for band configurations that may not be present in the real dataset but are possible to occur. We created approximately 10K simulated sheets.

[2]https://docs.opencv.org/4.x/index.htm

Gold Standard Data for evaluation

To rigorously evaluate the robustness of our AI solution, we crafted a gold-standard evaluation dataset. Drawing from a diverse array of LPA Strips collected across various states, we ensured the dataset's representativeness by encompassing real-world complexities. Addressing potential bias, we deliberately over-sampled rare patterns, including instances of complete resistance to drugs, thus affording a balanced evaluation landscape. Expert microbiologists coming from different regions provided their band-level annotations and clinical interpretations regarding drug sensitivity and resistivity. Through these rigorous steps, we established a gold standard evaluation dataset characterized by authenticity and diversity, furnishing a robust benchmark to gauge the performance of our AI solution across realistic conditions and affirming the meticulous methodology of our evaluation approach. The inter-annotator disagreement observed in this dataset allowed us to set the performance upper bound for our system. It also encouraged us to include human in the loop and to give the expert the authority to take the final call.

Methodology

We divide our solution into four major tasks as shown in the Figure 4:

1. Strip extraction: Extract strips from the LPA sheet.
2. Band identification: Locate bands and identify their class.
3. Clinical interpretation: Get drug interpretation from bands.
4. Human in the Loop (HiL): Involve humans (experts) in the loop by flagging novel or ambiguous strips and asking for their review.

Strip extraction

We perform a series of steps to extract strips from the input image of the LPA sheet.

1. Landmark detection. We look for a few landmark objects on the LPA sheet such as the doctor icon, clock, and reference strips. We detect these landmarks on the scanned LPA sheet, against the landmarks on a reference sheet by using ORB KeyPoint detection (Rublee et al. 2011), followed by the RANSAC (Fischler and Bolles 1981) algorithm for registration.

2. Sheet alignment. Subsequently, this registration information is used to estimate the transformation parameters, namely, shift, scale, and rotation, required to transform the scanned sheet into an aligned sheet. Any tilts, or offsets, will be corrected in this step.

3. Strip segmentation. After the sheet is properly aligned, strips are segmented and extracted based on the registered landmarks. The regions of interest (ROIs) are defined around the strips, with enough margin for any errors in detecting the exact bounding boxes that enclose the pasted LPA strips. In addition, lab ID images of each strip are extracted using the landmark objects and tagged with corresponding strips.

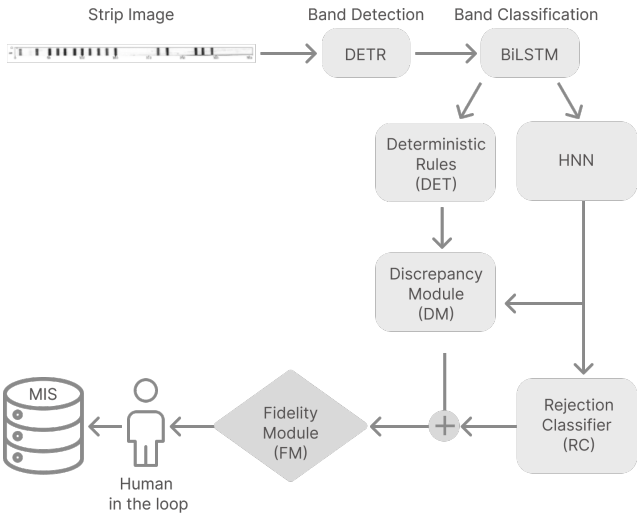

Figure 4: Steps to interpret a strip. Strip once extracted from the sheet is passed to DETR for Band detection and BiLSTM for band classification. Band probabilities are then sent to HNN and deterministic rule (DET) to produce drug interpretations. The rejection classifier rejects some strips based on HNN results. The discrepancy module (DM) rejects some based on the outputs of DET and HNN. The fidelity Module (FM) highlights ambiguous bands and sends the strip for review to the Human in the Loop.

4. Lab ID recognition. Each row in the sheet has a lab ID corresponding to the strip pasted to its right. This lab ID is used to index the patient records in MIS. We used TrOCR (Li et al. 2022), an encoder-decoder model to recognize the lab ID from the images. In our field study, we found that this simple system can save roughly 30% of the total time required for manual data entry.

Band Identification

After the strips are extracted, we perform the Band Identification task to identify the bands that are present in the strip. There are $K = 27$ band classes both in FL and SL. A band corresponds to a particular class based on the alignment between the reference strip bands and the predicted bands (see Figure 5). The reference block can be seen the the Figure 1. We divide the band identification task into two sub-tasks:

1. Band detection. Bands are black bars present on a white strip. Initially, we used an edge detection algorithm by OpenCV to get these bands. It worked well for cases where the band was clearly visible but not for light bands. Also, it picked up extra edges due to shadows and creases in the strip. Next, we used the bounding box annotations for bands present in the strip to train object detection models. We compared FRCNN and DETR on the task of band detection. We first pretrained them on the simulated dataset and finetuned them later on the real dataset. DETR outperformed FRCNN, providing tighter bounding box coordinates resulting in better band classification. Training DETR and FRCNN took approximately 12 hours for 100 epochs on a V100 GPU.

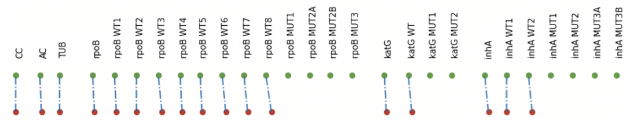

Figure 5: An example of mapping a band to a band class. Red dots denote the centroids of bands present in a strip and green dots denote the centroids of the reference strip bands.

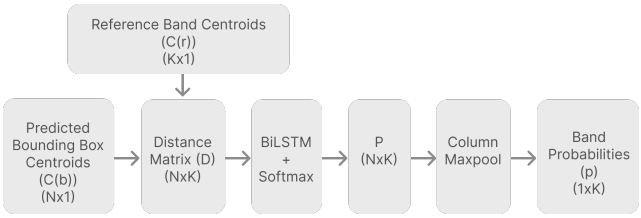

Figure 6: Relative distance between $C(r_k)$ and $C(b_i)$ is calculated to get D. Output of $\mathrm{Softmax_k}(\mathrm{BiLSTM}(D))$ is P. p is final band probability vector, $max_i(P_{ik})$

2. Band classification. Since the band class is dependent upon its distance from the reference strip band and also the class of its neighboring bands, we pose this as a sequence-to-sequence classification problem (see Figure 5). Consider N predicted bounding boxes and $K = 27$ reference bands. Let $\{b_1 \dots b_N\}$ denote predicted bounding boxes. Let $\{r_1 \dots r_K\}$ denote the reference bands. Let $C(\cdot)$ be a function that calculates the x-centroid. We calculate the distance of the centroids of predicted bounding boxes from the centroids of the 27 bands in the reference strip (see in Figure 1). Distance $d_{ik} = |C(b_i) - C(r_k)|$. This gives us a distance matrix $D = (d_{ik})$ of shape $N \times K$. D is now an input to the BiLSTM (Schuster and Paliwal 1997) model. We also apply softmax across K classes to get the P matrix with probability scores.

$$P = \mathrm{Softmax_k}(\mathrm{BiLSTM}(D)) \tag{1}$$

P is of dimension $N \times K$. Each row of P corresponds to a predicted bounding box and each column corresponds to a reference band class. We associate a bounding box with a class by identifying the bounding box with the maximum probability for that class. Specifically, $\hat{b}_k = \arg\max_i(P_{ik})$ where, $\hat{b_k}$ denotes the bounding box for k_{th} class. In the ground truth annotations, we know which band classes are present. Let, y_k be the groundtruth label for k_{th} class (0=absent,1=present). To train the BiLSTM model to map D to P, we first calculate final band probabilities by applying maxpool on the P matrix across all bounding boxes, i.e, $p_k = \max_i(P_{ik})$. We then apply the Binary Cross Entropy (BCE) loss on p_k and y_k. Refer to Figure 6 for more details.

Clinical Interpretation

Under ideal conditions, when the bands are clearly visible, clinical interpretation is straightforward once the bands are identified. Considering each strip to be a 27-dimensional bit array, where 1 represents the presence of a particular band class and 0 represents its absence, a set of deterministic rules can be applied that gives us sensitive (S), resistive (R) or invalid (NA) outcome for each drug. Let $\mathrm{DET}(\cdot)$ be the function that defines the deterministic rules. Using the band probability vector p defined in the previous section, we define the following.

$$\alpha_k = \begin{cases} 1, & \text{if } p_k > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$O_d = \mathrm{DET}(\alpha), \quad O_d \in \{S, R, \mathrm{NA}\} \tag{3}$$

However, we observed that, due to some physical factors, sometimes the bands are not dark enough to mark their presence or absence conclusively. In such cases, neighboring bands play a critical role, leaving room for subjective interpretation. Since in our dataset, we had band-level bounding boxes as well as clinical interpretation of the strips, we trained a model to learn this behavior.

We train a custom Hierarchical Neural Network (HNN) shown in Figure 7. The architecture of HNN replicates the logic encoded in the deterministic rules. The HNN takes band probability and band intensity (75th percentile of pixel values inside the band bounding box) as input for each band. Let $\mathrm{HNN}(\cdot)$ represent HNN, $I(\cdot)$ represent intensity calculator, $\hat{b}_k$, p are from previous section.

$$i_k = I(\hat{b}_k) \tag{4}$$

$$O_h = \mathrm{HNN}(p, i) \quad O_h \in \{S, R, \mathrm{NA}\} \tag{5}$$

We apply Cross Entropy (CE) loss at drug level outputs. Although HNN is modeled to predict the drug outputs, its design allows it to output predictions for each band and loci (see Figure 7). To improve its performance, we apply a BCE loss at the band and loci level as well (see Figure 7).

Human in the Loop

To mitigate model biases and also to leverage expert's opinions, we give a fraction of the strips to the lab technicians for review. We reduce their effort by picking only ambiguous or novel samples. To strike the right balance between effort (number of strips to be reviewed) and efficiency, we created three modules:

1. Rejection classifier. It is a model similar to HNN in the structure that predicts samples that should be rejected i.e., sent for a review. This rejection classifier (RC) takes all the inputs that HNN takes and additionally, it takes HNN outputs as inputs. It is trained to flag failures of HNN. Let Y_h be ground-truth drug labels for HNN and Y_{rc} be ground-truth for the rejection classifier. We define Y_{rc} as follows,

$$Y_{rc} = \begin{cases} 1, & \text{if } Y_h \neq O_h \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Since HNN shows equitable performance on the train and test split, we train RC on the same train split. With improvements in performance of HNN, Y_{rc} becomes skewed, hence we use focal loss (Lin et al. 2020) instead of BCE. For inference, we choose a probability threshold above which the

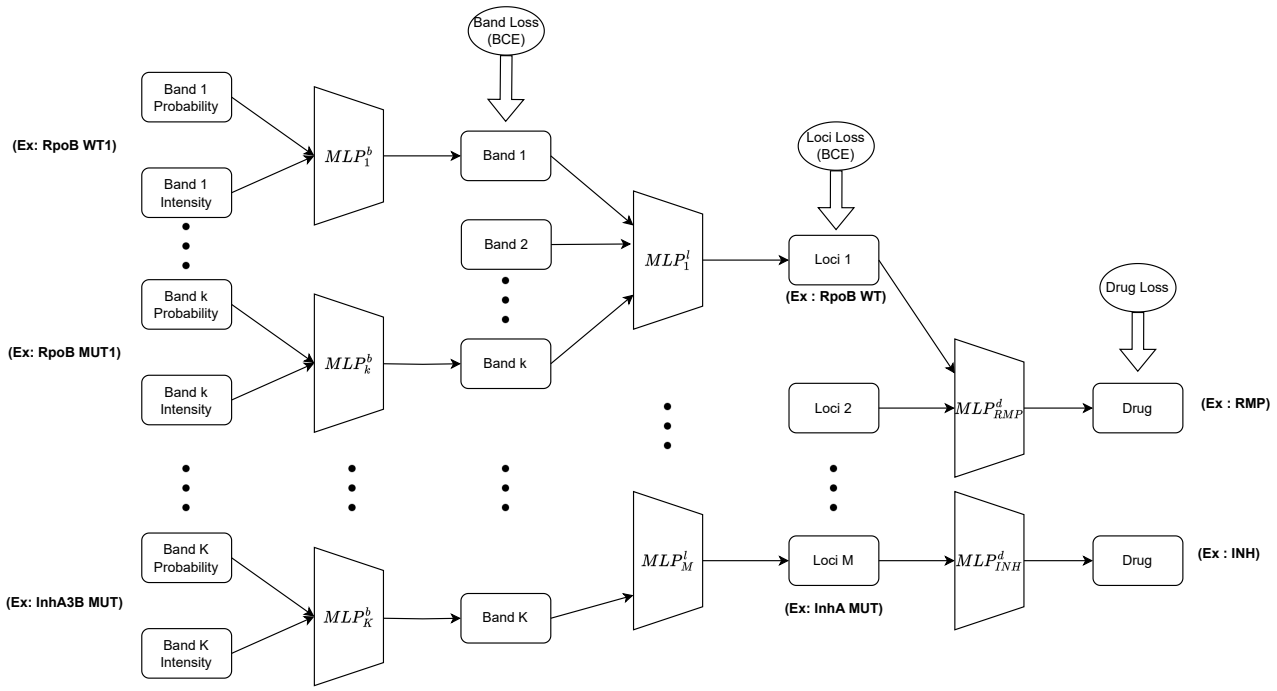

Figure 7: HNN structure. The structure of the HNN is derived from the deterministic rule for loci and drug interpretation from bands. The Bands that affect loci are connected to an MLP, similarly, loci that affect the drug are connected to a MLP that produces drug outputs. There are K bands and M loci. MLP contains 2 layers with $Tanh$ activation.

strip is rejected and sent for review. A lower threshold corresponds to higher effort. We calculate the threshold on the validation set to optimize for the minimum effort to reach 95% accuracy across all drugs.

2. Discrepancy module. For all the strips where DET and HNN differ ($O_d \neq O_h$), the Discrepancy module (DM) sends them for review.

3. Fidelity module. Rejected strips from RC and DM are first gathered in the Fidelity module (FM). FM finds bands that are most likely responsible for this discrepancy and flags them for review. This is done by back-tracking O_d to find bands when flipped will cause O_d to change and also have a low band intensity i_k.

Results and Evaluation

Here we report results on various submodules present in our solution as well as end-to-end system.

Lab ID recognition The lab IDs are handwritten in specified columns. These IDs are 4-5 digit numbers. As a baseline, we trained the CNN-LSTM model on our lab ID dataset. The dataset consists of 5145 train images and 1288 test images of lab IDs cropped from LPA sheets. We also experimented with the TrOCR model, pretrained it on the IAM dataset (Marti and Bunke 2002), and fine-tuned it on our dataset. We evaluated the models on complete lab ID recognition i.e., even if a single digit is incorrectly predicted, lab ID prediction is considered incorrect. We report accuracy numbers for lab ID predictions in Table 2.

Architecture	Accuracy
Custom CNN-LSTM	0.84
TrOCR	0.92
TrOCR pre-trained on IAM dataset	0.943

Table 2: LabID prediction results.

Band Identification As discussed before, the Band Identification task is divided into two subtasks i) Band Detection and ii) Band Classification. We compare DETR and FRCNN models for the Band Detection task. We calculate the Mean Average Precision (mAP) for a single class (Band or no band) at iou $= 0.5$. Scores can be seen in the Table 5. For Band Classification we test our BiLSTM model on the outputs of DETR and FRCNN models respectively. BiLSTM identifies bands present in the strip, hence we track its specificity and sensitivity. The results are mentioned in the Table 6.

Clinical Interpretation We compare HNN and DET performance on the test set for both FL and SL. Table 7 shows the superior performance of HNN compared to the DET. HNN and DET perform similarly on the samples where bands are clearly visible. In samples with light or faint bands, HNN excels. More details on this are mentioned in the supplementary material.

Human in the Loop Since the distribution of classes for each drug is not even, we track the precision and recall for

Drug	Drug Result	Total Strips	Count	Precision (in %)	Recall (in %)	% of strips that have at least one band that require		Accuracy (in %)
						Visual Inspection	Modifying the Band Predictions	
RMP	Sensitive	2641	2155	98	99	37	4.07	98
RMP	Resistive		159	100	91			
RMP	Invalid		327	94	92			
INH	Sensitive		2029	98	98	20.4	3.0	97
INH	Resistive		278	93	99			
INH	Invalid		334	93	88			

Table 3: FL drug interpretation results with Human in the Loop

Drug	Drug Result	Total Strips	Count	Precision (in %)	Recall (in %)	% of strips that have at least one band that require		Accuracy (in %)
						Visual Inspection	Modifying the Band Predictions	
FLQ	Sensitive	340	197	100	97	32	9	98
FLQ	Resistive		66	92	99			
FLQ	Invalid		77	98	100			
KAC	Sensitive		257	98	98	25	2.75	98
KAC	Resistive		6	100	67			
KAC	Invalid		77	94	97			
LLK	Sensitive		261	99	99	6	0.3	99
LLK	Resistive		3	NA	NA			
LLK	Invalid		78	95	100			

Table 4: SL drug interpretation results with Human in the Loop

Architecture	FL mAP	SL mAP
FRCNN	57.5	49.7
DETR	89.27	78.2

Table 5: FRCNN and DETR on Band detection.

Model	FL		SL	
	SP	SN	SP	SN
FRCNN+BiLSTM	0.86	0.87	0.85	0.83
DETR+BiLSTM	0.96	0.97	0.97	0.96

Table 6: Band Classification Results. SP (Specificity), SN (Sensitivity)

each class of each drug. Table 3 and 4 showcase our system performance with human in the loop. Our system achieves near-perfect performance (Accuracy $> 98\%$) on all drugs for both FL and SL with a small fraction of samples requiring human intervention. For HiL, we recalculated the scores considering the samples sent for review to be correct. As seen in Tables 7 and 3, for RMP, accuracy improved from 94% to 98% with 37% samples sent for review. However, only 4% of the total samples required intervention. We strongly believe that this lift due to HiL cannot be achieved only by more data or better models - a plausible reason being that the label noise is incompressible.

Limitations and Ethical Considerations

In theory, a single LPA strip can have 27 bands present or absent leading to 2^{27} configurations. While in practice only about 400 are observed, other rare combinations cannot be ignored. Given the limited diversity in our dataset, we cannot fully guarantee the generalization of our solution on those rare unseen samples. Table 4 showcases model results for SL drugs. The support for resistive KAC and LLK samples is very low (6 and 3 resp.). This is because they occur very

Model	FL (acc %)		SL (acc %)		
	RMP	INH	FLQ	KAC	LLK
DET	76	81	66	62	76
HNN	94	94	88	95	98

Table 7: DET and HNN Drug Interpretation scores

rarely. Model performance estimates are not reliable because the support is very low. We expect this to improve with the availability of more datasets. Light bands are a major source of error in our solution. Strips containing light bands, especially MUT (mutant) type bands make it difficult to interpret the strip. Even a single error in interpreting the band can completely change the drug outcomes. We try to mitigate these problems by establishing Human in the Loop. With the combination of RC, DM, and FM, we flag such cases to be reviewed by the lab technicians.

Conclusion

In this paper, we have described a novel LPA interpretation system. This system has the potential to fast-track LPA tests and also reduce the burden on lab technicians while maintaining the same level of efficacy. Our system uses multiple state-of-the-art (SOTA) models to optimize various subtasks present in the process. We showcase the effectiveness of our system by evaluating it on a gold-standard dataset. To reduce risk and mitigate model biases, we also introduce Human in the loop. This makes our system more transparent and trustworthy.

Acknowledgements

This work is supported by the United States Agency for International Development (USAID) under the TraceTB grant. We thank the contribution of Apoorv Agnihotri, Apoorve

Singhal, and Microbiologists from different IRL labs to contribute in creating a Gold Standard Dataset for evaluation. The microbiologist names are as follows, Dr. M. Hanif (NDTBC, Delhi), Dr. Ritu Singhal (NITRD, New Delhi), Dr. Sarika Jain (NTI, Bangalore), Ms Mamtha HG (NTI, Bangalore), Dr. Soumya Dhawan (IRL Bhopal), Dr. N.Ravi Shankar (Vizag, Andhra Pradesh), Mr. Rooban (IRL Madurai), Dr. Bandana Chaudhary (IRL Assam) and Dr. Sailesh (IRL Nagpur). We also thank, Makarand Tapaswi and Minesh Mathew for their help in writing and reviewing the paper. Dr. Neeraj Aggarwal, Aayushi Bhotica, Dr Aparna, and Dr. Malay Shah for their valuable suggestions and support of the project.

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. arXiv:2005.12872.

Fischler, M. A.; and Bolles, R. C. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM*, 24(6): 381–395.

Hendrickx, K.; Perini, L.; der Plas, D. V.; Meert, W.; and Davis, J. 2023. Machine Learning with a Reject Option: A survey. arXiv:2107.11277.

Jotterand, F.; and Bosco, C. 2020. Keeping the "Human in the Loop" in the Age of Artificial Intelligence: Accompanying Commentary for "Correcting the Brain?" by Rainey and Erden. *Science and Engineering Ethics*, 26.

Kvak, D.; Chromcová, A.; Biroš, M.; Hrubý, R.; Kvaková, K.; Pajdaković, M.; and Ovesná, P. 2023. Chest X-ray Abnormality Detection by Using Artificial Intelligence: A Single-Site Retrospective Study of Deep Learning Model Performance. *BioMedInformatics*, 3(1): 82–101.

Li, M.; Lv, T.; Chen, J.; Cui, L.; Lu, Y.; Florencio, D.; Zhang, C.; Li, Z.; and Wei, F. 2022. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. arXiv:2109.10282.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2): 318–327.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016*, 21–37. Springer International Publishing.

Lopez-Garnier, S.; Sheen, P.; and Zimic, M. 2019. Automatic diagnostics of tuberculosis using convolutional neural networks analysis of MODS digital images. *PLOS ONE*, 14(2): e0212094.

Marti, U.-V.; and Bunke, H. 2002. The IAM-database: An English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5: 39–46.

Panicker, R. O.; Soman, B.; Saini, G.; and Rajan, J. 2015. A review of automatic methods based on image processing techniques for tuberculosis detection from microscopic sputum smear images. *Journal of Medical Systems*, 40(1).

Plesner, L. L.; Müller, F. C.; Nybing, J. D.; Laustrup, L. C.; Rasmussen, F.; Nielsen, O. W.; Boesen, M.; and Andersen, M. B. 2023. Autonomous Chest Radiograph Reporting Using AI: Estimation of Clinical Impact. *Radiology*, 307(3).

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497.

Rublee, E.; Rabaud, V.; Konolige, K.; and Bradski, G. 2011. ORB: an efficient alternative to SIFT or SURF. 2564–2571.

Schuster, M.; and Paliwal, K. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45: 2673 – 2681.

Yang, R.; and Yu, Y. 2021. Artificial Convolutional Neural Network in Object Detection and Semantic Segmentation for Medical Imaging Analysis. *Frontiers in Oncology*, 11.