

Urdu Handwriting Recognition Using Machine Learning

Advisor: Dr. Ubaid Ullah Fiaz
Co-Advisor: Dr. Kashif Javed

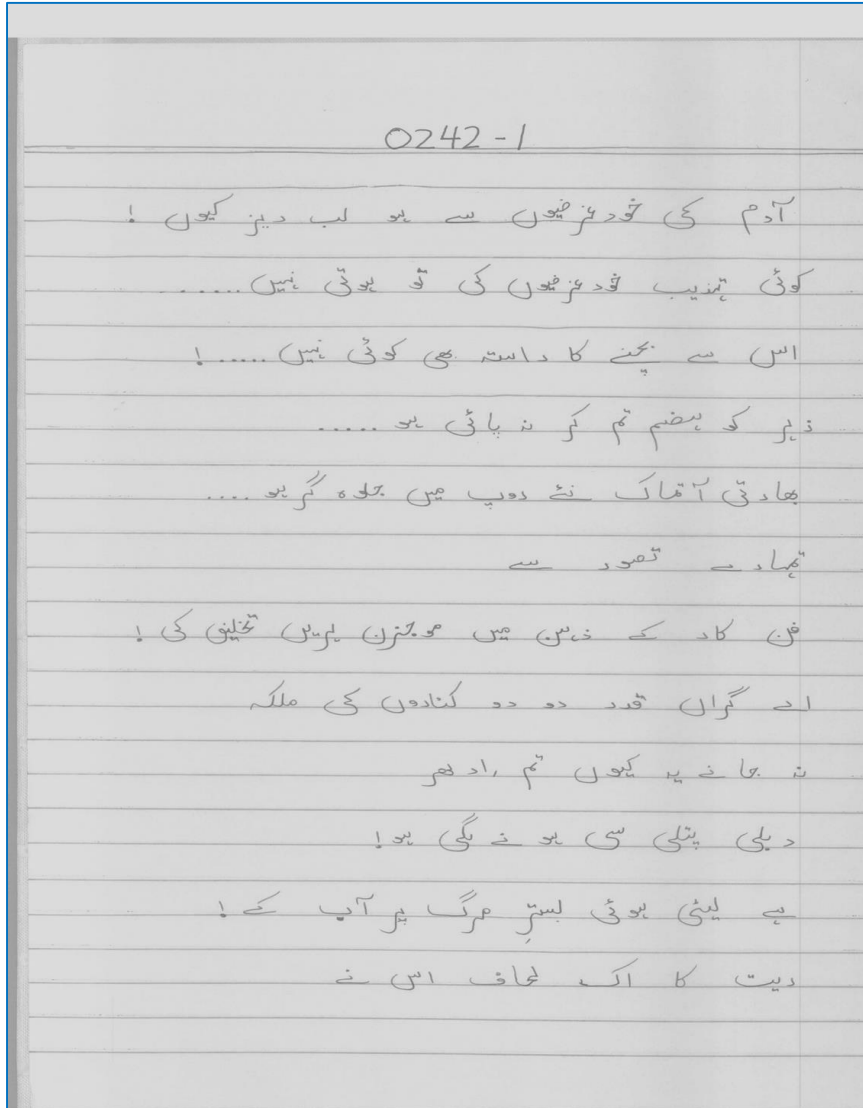
Group Members

2015-EE-167	Shehryar Malik
2015-EE-168	M. Naeem Maqsood
2015-EE-188	Abdur Rehman Ali

Presentation Outline

- ☐ **Problem Statement**
- ☐ **Proposed Solution and Analytical Analysis**
- ☐ **Methodology – Flow Chart**
- ☐ **Block Diagram and Details of Sub Systems**
- ☐ **Design of sub systems**
- ☐ **Results**
- ☐ **Benchmarking**
- ☐ **Modern Tools**
- ☐ **Project Management and Team Work**

Problem Statement

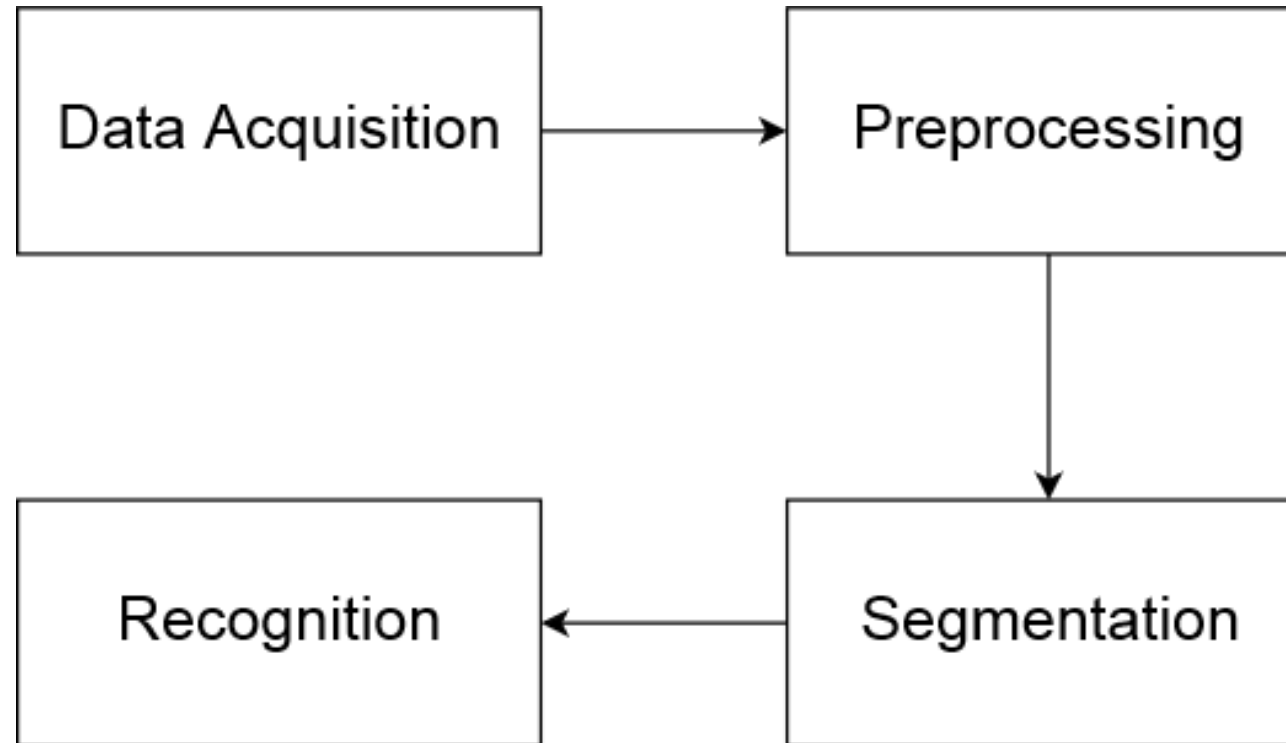


آدم کی خود غرضیوں سے بولب ریز کیوں !
 کوئی تہذیب خود غرضیوں کی تو ہوتی نہیں
 اس سے بچنے کا رستہ بھی کوئی نہیں !
 زہر کو بھم تم کر نہ پائی ہو
 بھارتی انما اک نئے روپ میں جلوہ گر ہو
 تمہارے تصور سے
 فن کار کے ذہن میں موجزن لہریں تخلیق کیا
 اے گراں قدر دو دو کناروں کی ملکہ
 نہ جانے یہ کیوں تم ادھر
 دہلی پتلی سی ہوئے لگی ہوا
 بے لیٹی ہوئی بستر مرگ پر آب کے
 ریت کا اک لحاف اس نے
 گنگوتری تک مرض اس کا پھیلا ہوا
 اسی وجہ سے تو
 تو بتاؤ کہ تم
 کیوں جٹاؤں سے نیچے گریں؟

Problem Statement

- Recently, state institutions have been digitizing handwritten records (such as old land records) and FIRs.
- This project may be extended to be used in these initiatives.
- Deep Learning algorithms have been applied for this purpose to other languages and have resulted in surprisingly high accuracies.
- Goal of this project is to apply deep learning algorithms for Urdu handwriting recognition.

Proposed Solution



Data Acquisition

- Selected 15,000 lines from approximately 1 million lines taken from Urdu books
- Selection Criteria: Most frequent trigrams, bigrams and unigrams
- Have collected data samples of approximately 5000 lines from 160 writers
- Will extend this to 500 writers

Preprocessing – Binarization & Noise Removal

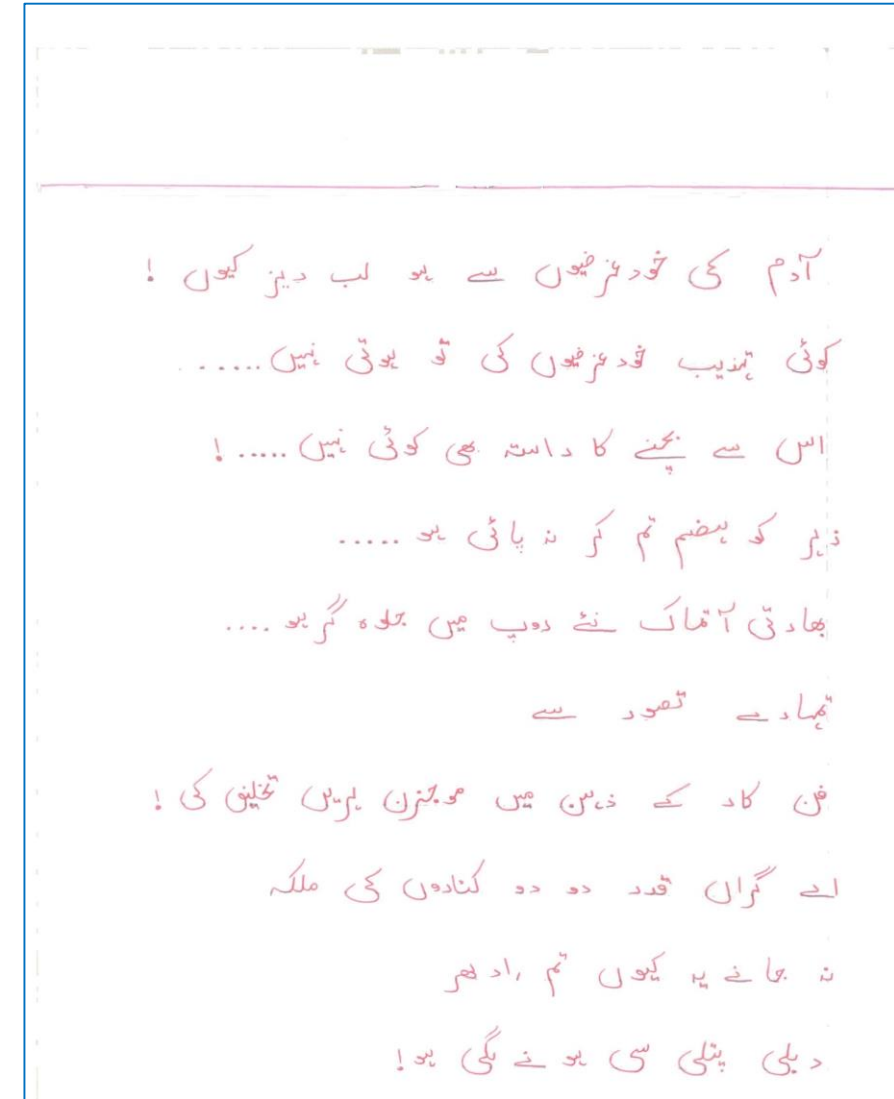
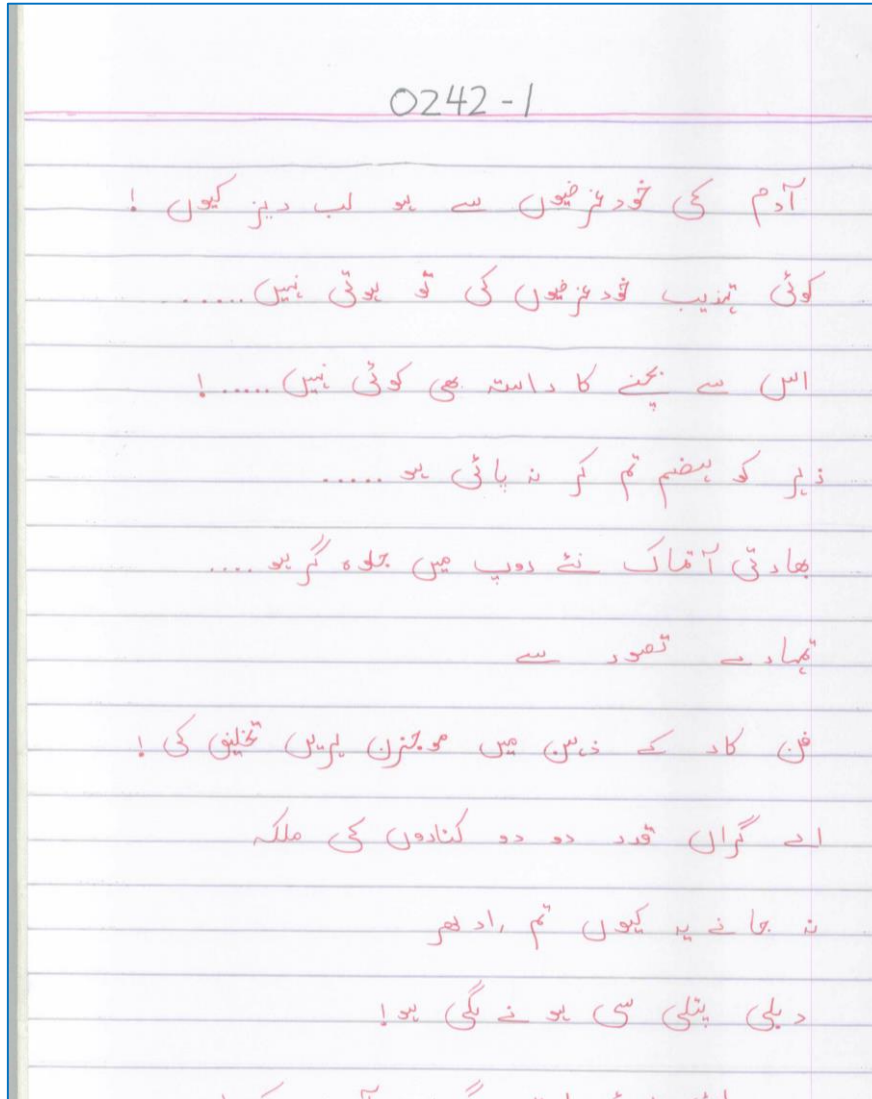


Image Segmentation – Horizontal Projection

ہر قسم کا صفحہ ادا کرنے پر قادر ہو جائے۔

صورت حالات کا صحیح جائزہ لینے کے بعد سرسید احمد خان کو ایسے رفقاء مل گادی

تلاش ہوئی جن پر وہ اعتماد کر

سزاوار بھی تھا، انہیں ایسے رفقاء مل گئے۔ ان رفیقوں میں حالی،

سکس اور جوان کی حکومت علی کے رموز سے آگاہ ہوں۔ کچھ اتفاق کی

ہاں ہے، کچھ یہ کہ سرسید کا فلو ص اس کا

ہر قسم کا صفحہ ادا کرنے پر قادر ہو جائے۔

صورت حالات کا صحیح جائزہ لینے کے بعد سرسید احمد خان کو ایسے رفقاء مل گادی

تلاش ہوئی جن پر وہ اعتماد کر

سکس اور جوان کی حکومت علی کے رموز سے آگاہ ہوں۔ کچھ اتفاق کی

ہاں ہے، کچھ یہ کہ سرسید کا فلو ص اس کا

سزاوار بھی تھا، انہیں ایسے رفقاء مل گئے۔ ان رفیقوں میں حالی،

شبلی اور نذیر احمد بہت مشہور ہیں۔

صوم، مرد مسلمان، انسان کا مل 219

اقبال کے شعور تخلیق کا ابلاغ و اظہار 181

”پڑھ بابا کو کیا ہوا؟ لوگ جمع تھے اور وہ صبح رہا تھا، اب کیس نظر نہیں

”ڈاکٹر نے اسے پاگل خانے بھیج دیا ہے، سنا ہے کل رات وہ سب کو تنگ

”تو بابا کو پاگل خانے بھیج دیا گیا۔“ لوگ اسے زحمت کرنے کے لئے جمع

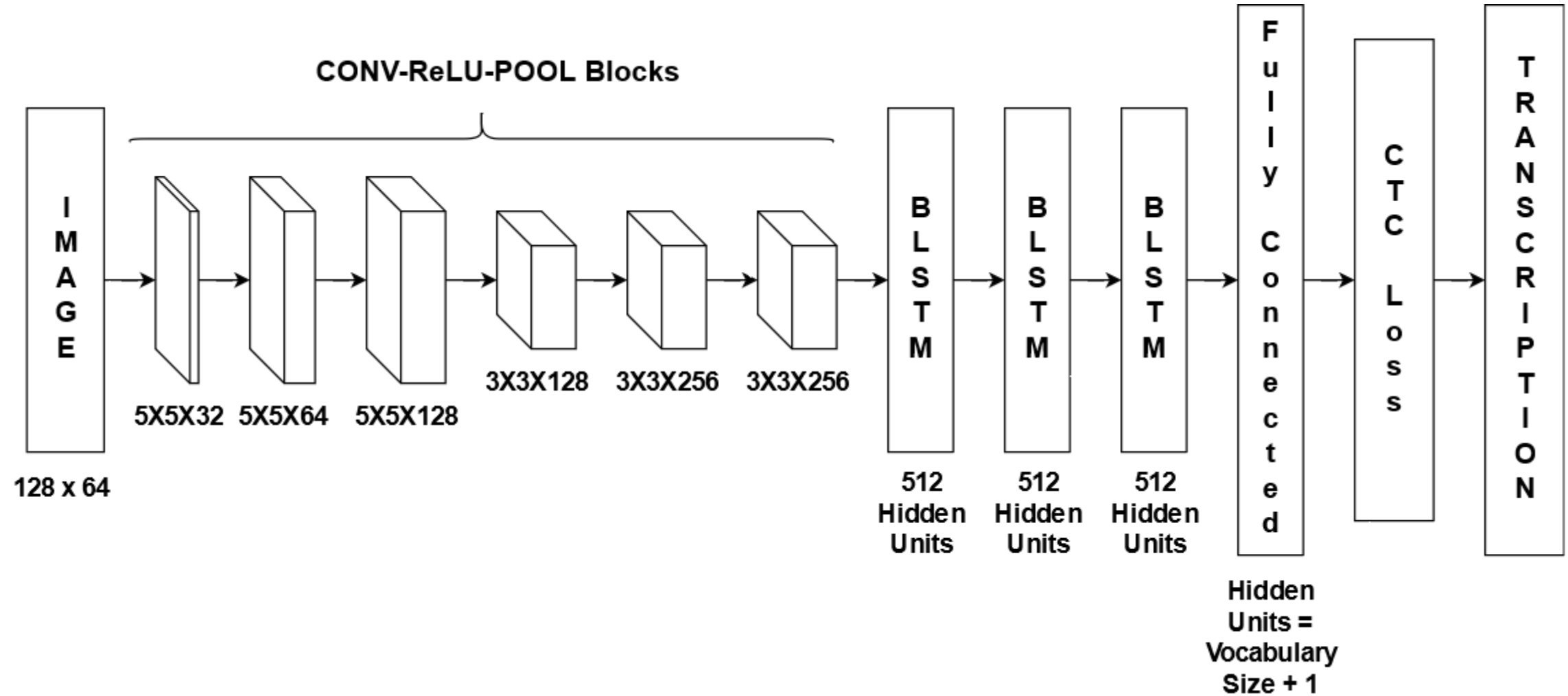
Image Segmentation – Connected Components

صورتِ حالات کا صحیح جائزہ لینے کے بعد سرسید احمد خان کو ایسے رفقاء ملارہی

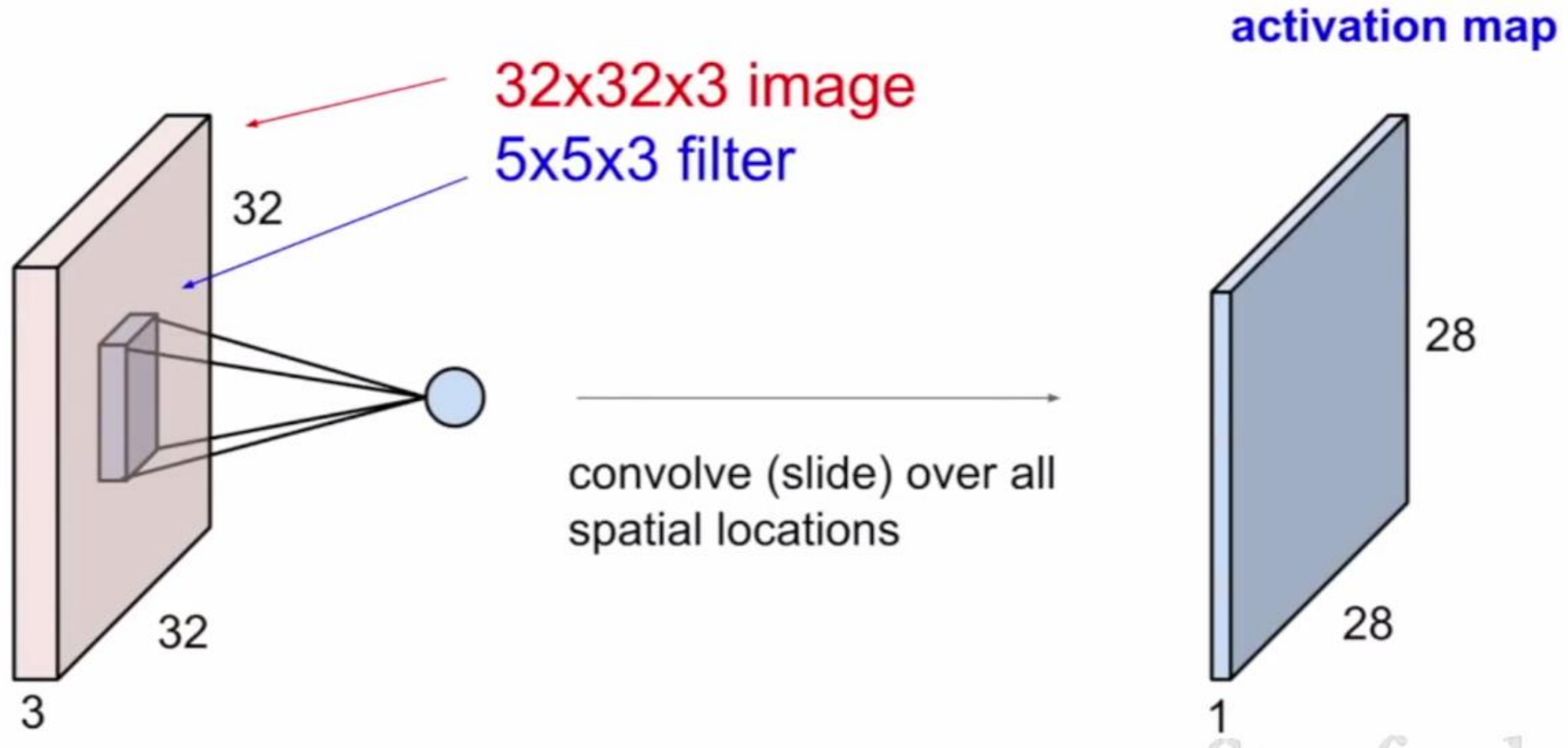


صورتِ حالات کا صحیح جائزہ لینے کے بعد سرسید احمد

Recognition Model



CONV-ReLU-POOL Blocks: Convolution



CONV-ReLU-POOL Blocks: ReLU

- ReLU = Rectified Linear Unit
- Given some input matrix/tensor X :

$$\text{ReLU}(X)_{ij} = \max(0, X_{ij})$$

- Main Idea: Each convolutional filter should learn to recognize some 'feature' of the image. Output of filter should be > 0 when it sees that 'feature' in the image and < 0 when it doesn't. Output is proportional to how sure the filter is of that 'feature' being present in the image

CONV-ReLU-POOL Blocks: Pool

- Subsampling layer: reduces dimension of data
- E.g. 2X2 pooling layer

5	2	0	1
4	7	1	3
9	4	2	6
2	5	8	3

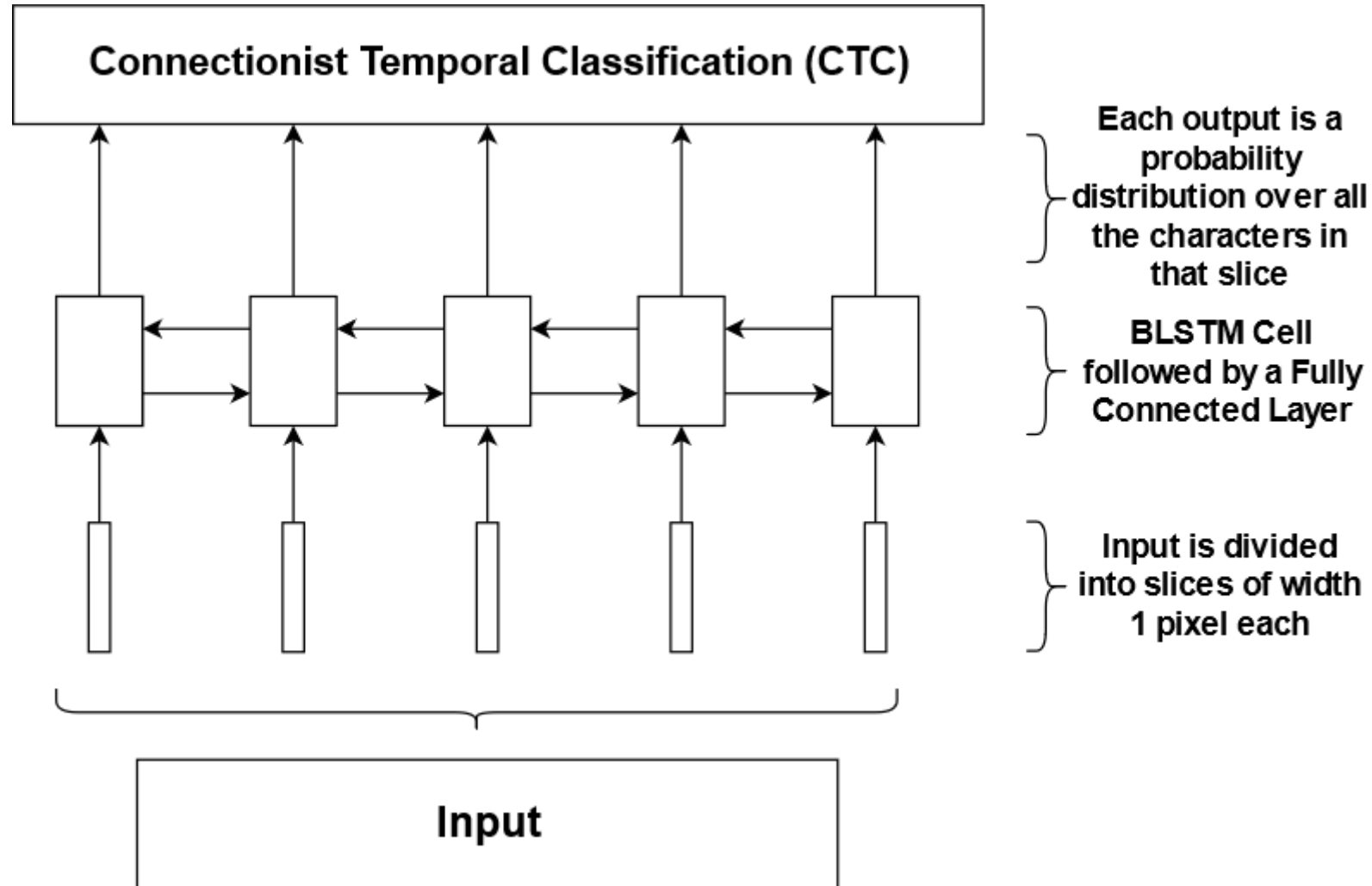
Input



7	3
9	8

Output

Bidirectional LSTM & CTC Loss



Connectionist Temporal Classification (CTC)

- Suppose transcription, t , is:

n - e e u - r a - l l -

where '-' indicates a blank in the sequence

- β function merges repeated labels and deletes blanks. Output, o , is:

n e u r a l

- Let l be the true label and x be the image. Then loss is defined as:

$$-\log \sum_{l \in \beta(t)} p(t | x)$$

- Minimize this loss with respect to network's parameters

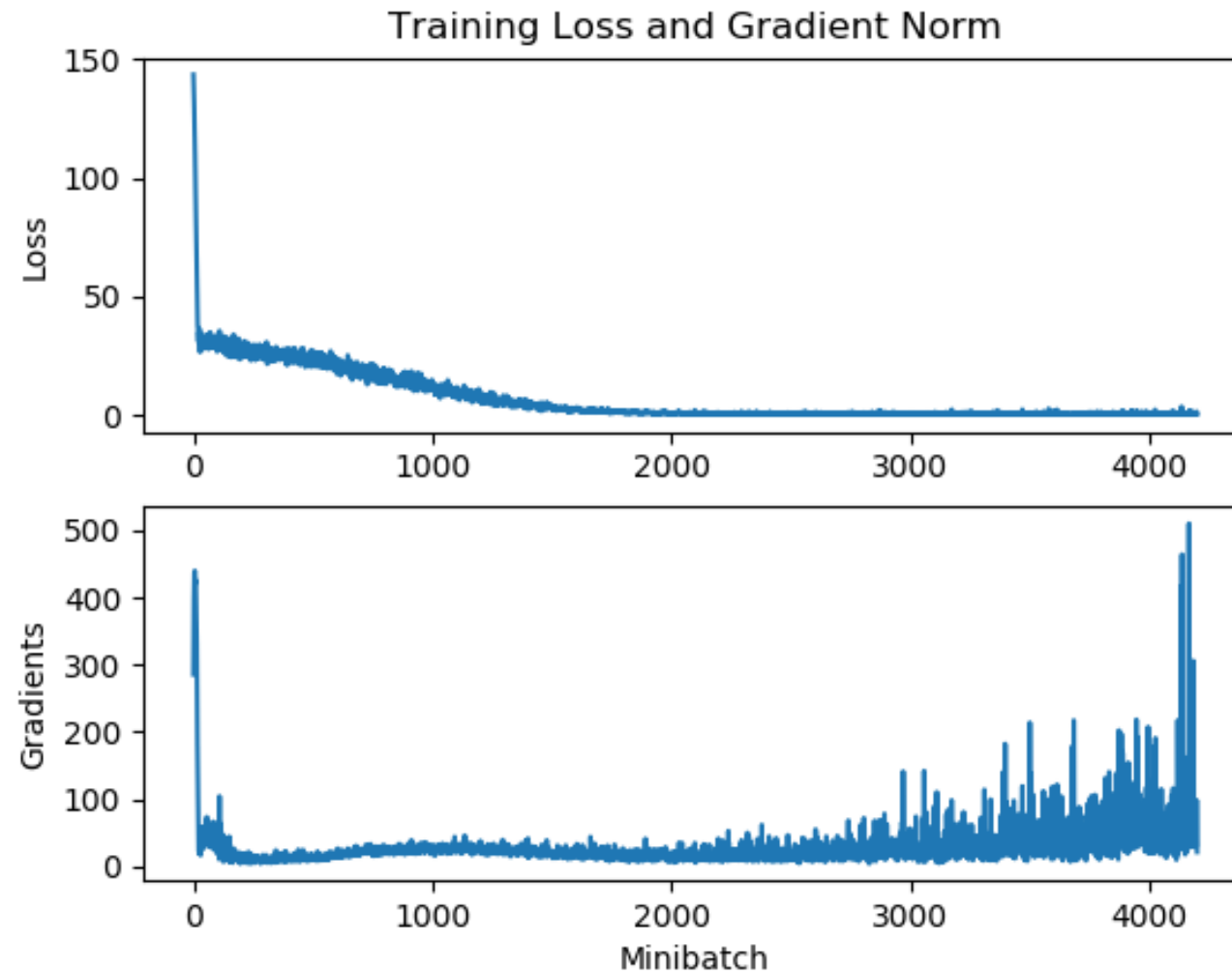
Accuracy Metrics

- Loss
 - Equal to negative log probability of true label
 - So if probability of true class is 0.6, then loss = $-\log(0.6) = 0.7$
- Character Recognition Rate (CRR):
 - Let o be the output of system and l be the correct label. Then:

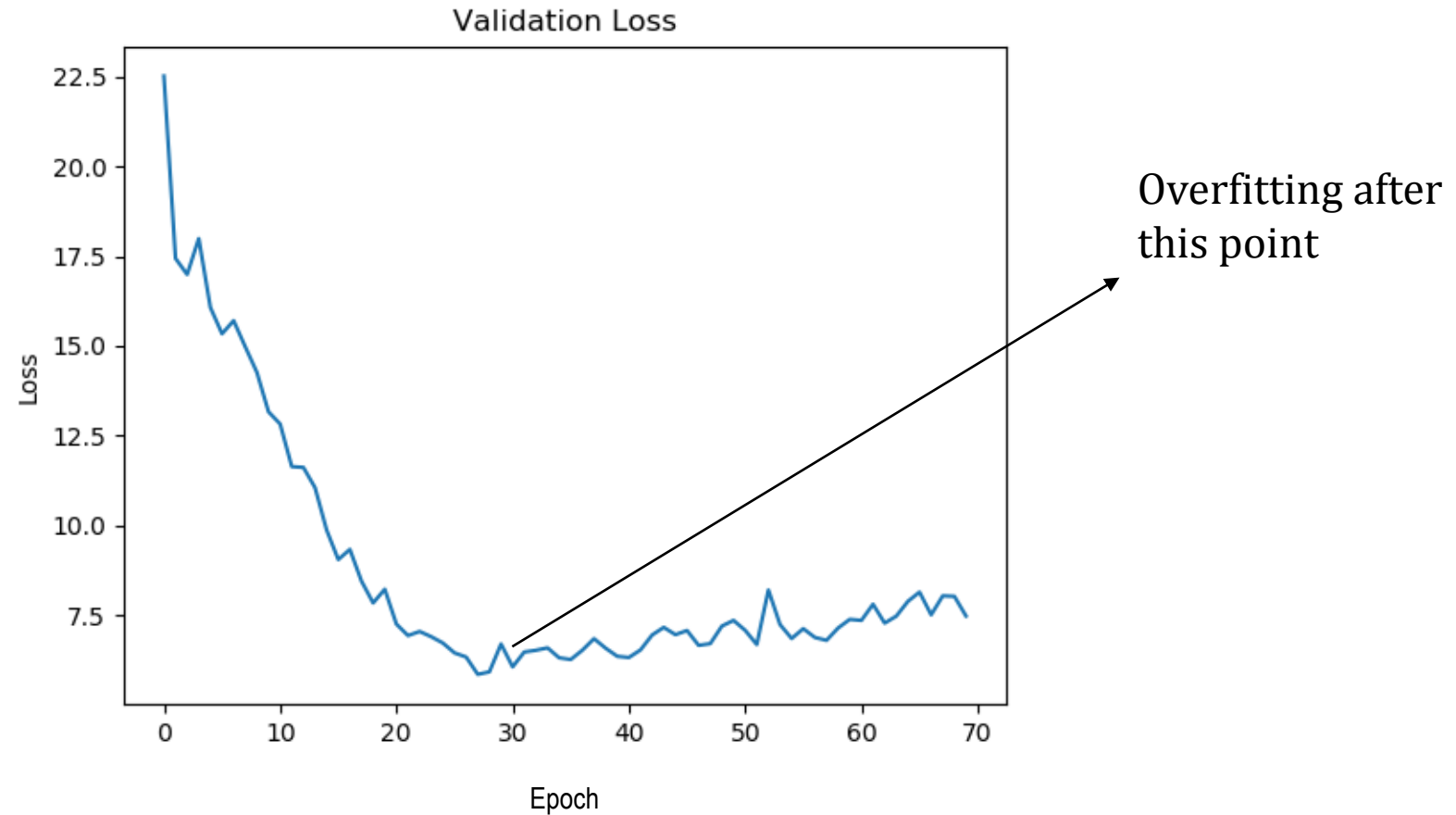
$$\% \text{ CRR} = \text{Levenshtein distance}(o,l) / \text{Length of } l$$

where Levenshtein distance(o,l) is the number of insertions, deletions and substitutions required to make o and l equal

Results

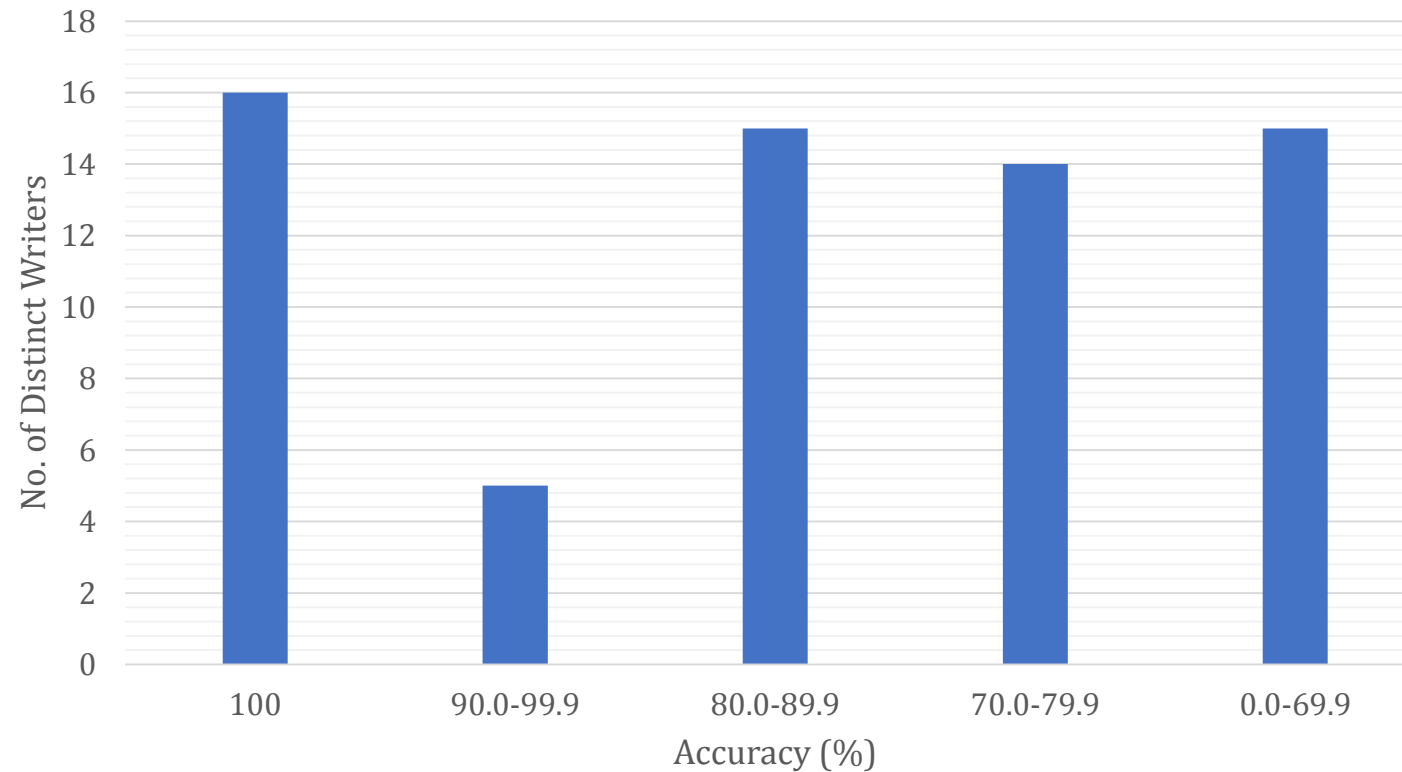


Results

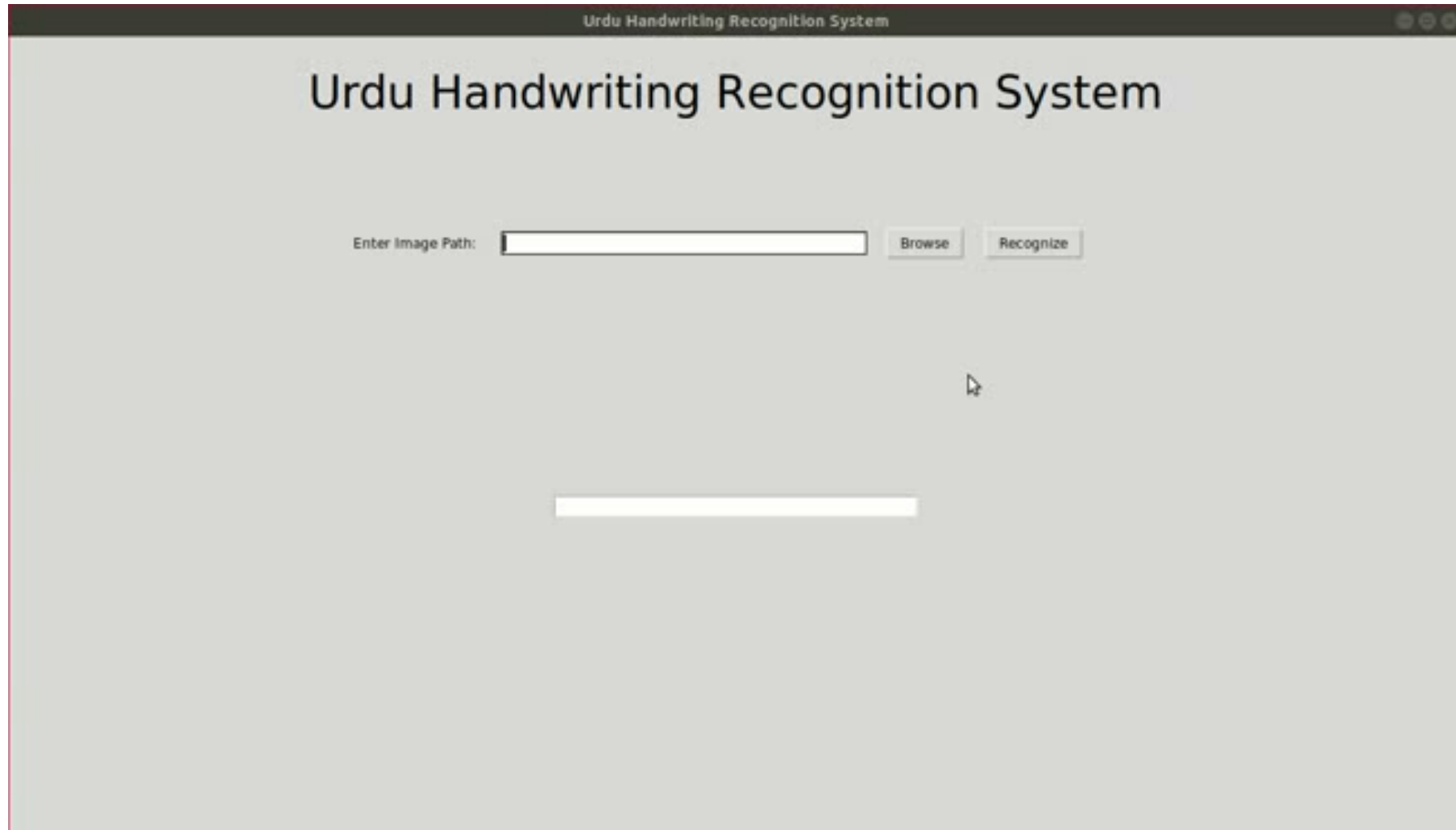


Results

- Validation set accuracy = 72.2%
- Model captures variation between writers



Prototyping



Benchmarks For Other Recognition Tasks

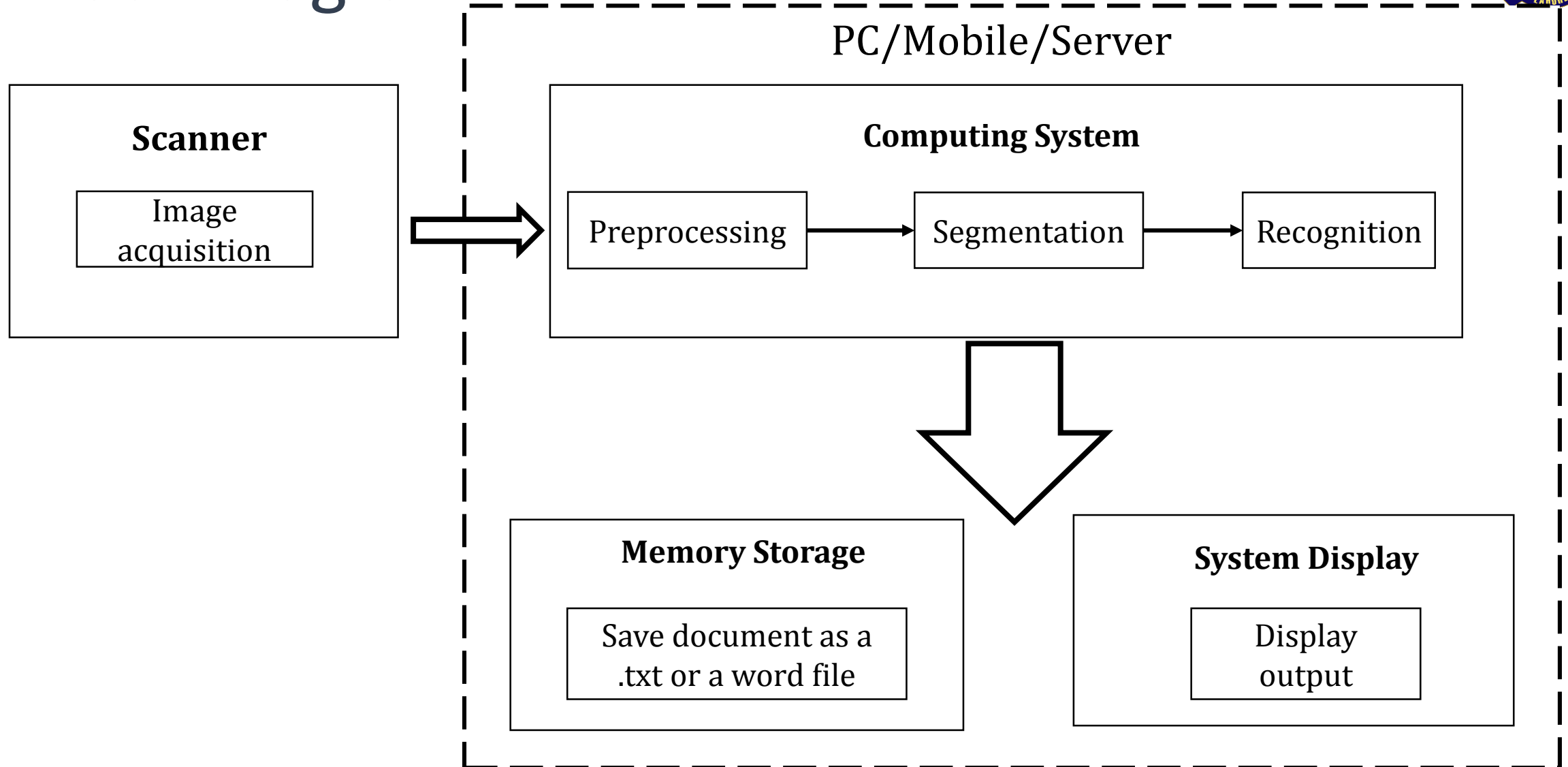
Language	Dataset	Dataset Size	Character Recognition Rate (%)
English Handwritten Text	IAM	13,533	93.4 ¹
Arabic Handwritten Text	KHATT	9,000	75.8 ²
Urdu Printed Text	UPTI	10,000	98.0 \pm 0.25 ³

¹ Bluche et al. *Scan, Attend and Read: End-to-End Handwritten Paragraph Recognition with MDLSTM Attention*. Preprint: arXiv:1604.03286v3. August 23, 2016

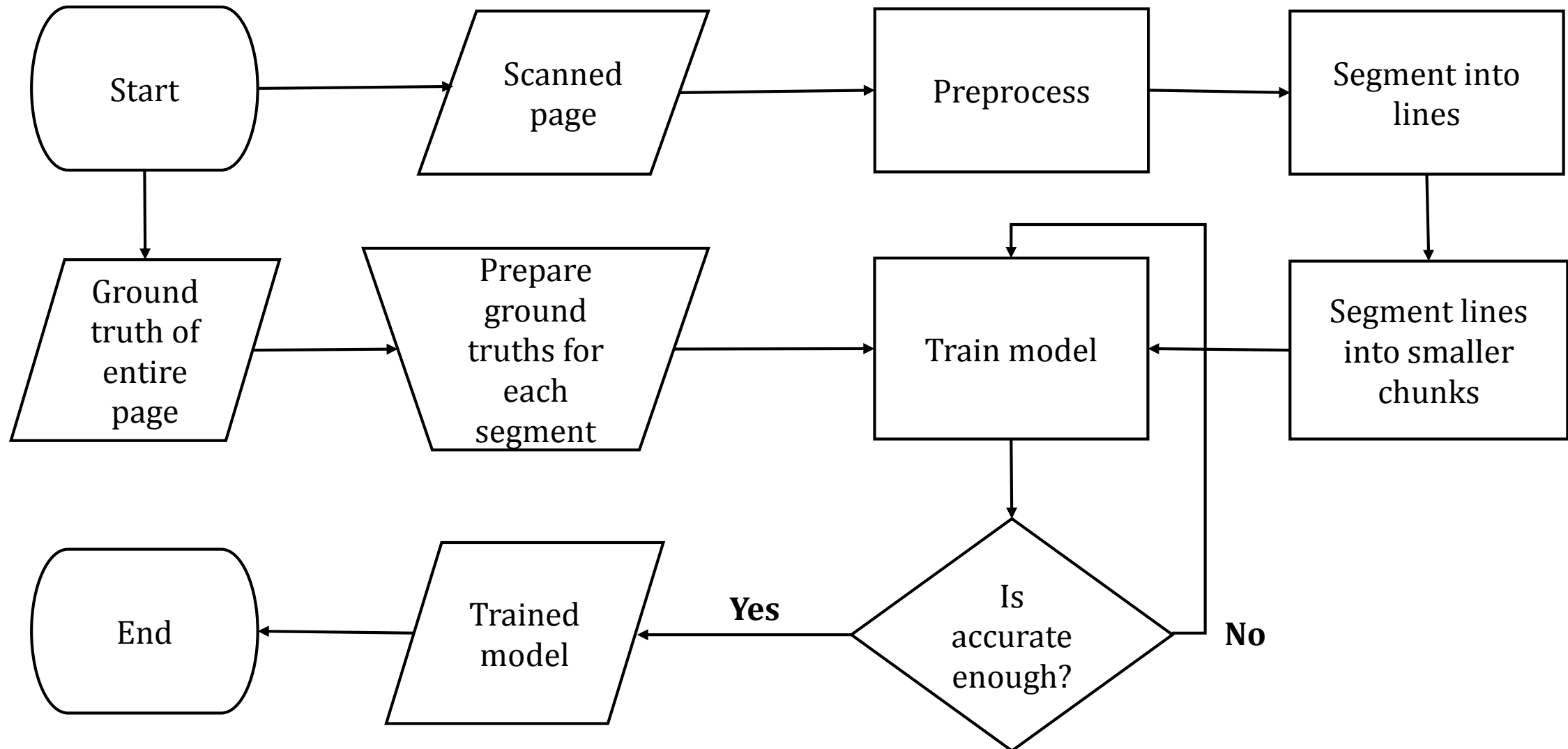
² Ahmed et al. *Deep KHATT: A Deep Learning Benchmark on Arabic Script*. Document Analysis and Recognition (ICDAR), IAPR International Conference, 2017.

³ Naz et al. *Urdu Nasta'liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks*. SpringerPlus. November, 2016.

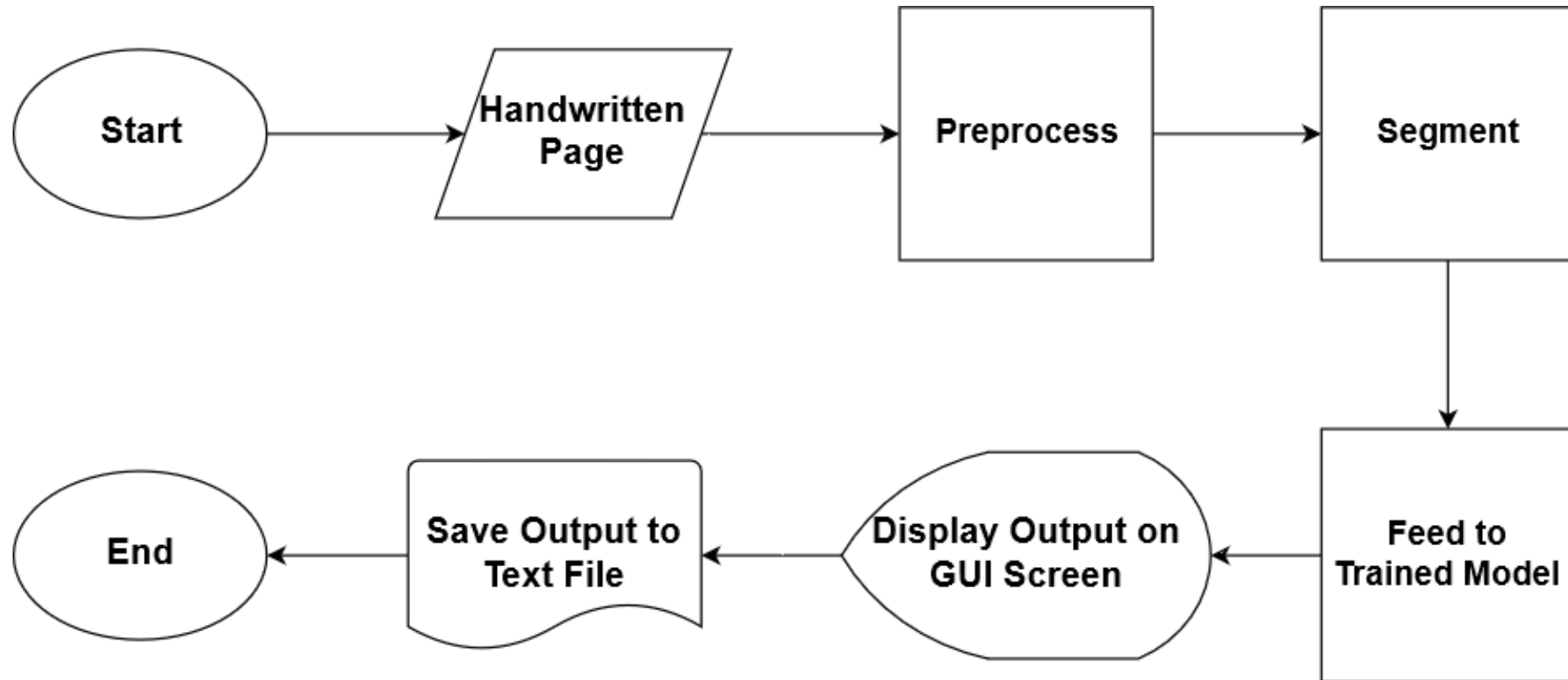
Block Diagram



Flow Chart - Training



Flow Chart - Deployment



Modern Tools Usage

- Programming Languages
 - Python
 - Java
- Libraries
 - Tensorflow
 - NumPy
 - OpenCV
- Computational Resources: Google Compute Engine for GPUs
- Graphical User Interface Frameworks
 - QT Creator
 - Tkinter

Industrial Linkage – Funding - Entrepreneur

- Applications
 - Digitization of land records (e.g. recent digitization by Punjab Information Technology Board)
 - Digitization of old handwritten records in universities, schools, libraries
 - Digitization of FIRs
- Funding
 - Center for Language Engineering (CLE), KICS, is funding the dataset part of the project
- May have the potential to be a commercial product
 - Useful for governments, universities, schools, libraries
 - Can help companies become paperless

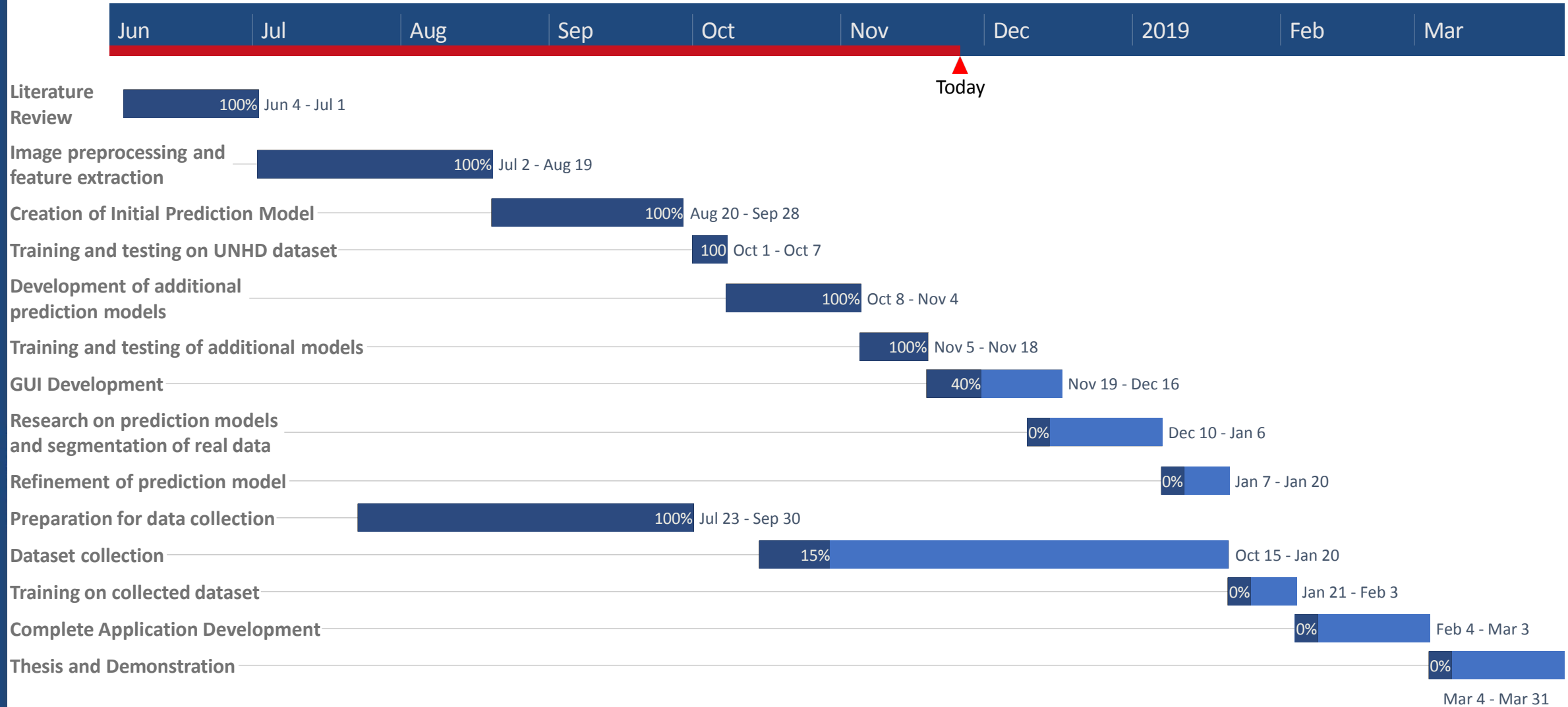
Future Deliverables

- Urdu Handwritten Text Dataset – 10,000 text lines with Ground Truths
- An executable file (and the source code) of the final GUI program
- Poster Competition –Industrial Open House
- Project Report
- Project Paper
- Project Video

Project Division

Members	Literature Review	Data Acquisition and Preprocessing	Segmentation	Model for Recognition	GUI Development
Shehryar Malik 2015-EE-167	<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
M. Naeem Maqsood 2015-EE-168	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		
Abdur Rehman Ali 2015-EE-188	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>

Gantt Chart



Thank You