MULTI-LABEL CLASSIFICATION FOR AUTOMATIC HUMAN BLASTOCYST GRADING WITH SEVERELY IMBALANCED DATA

Lisette Lockhart[1], Parvaneh Saeedi[1], Jon Havelock[2], Jason Au[2]

[1] Simon Fraser University, Burnaby, Canada

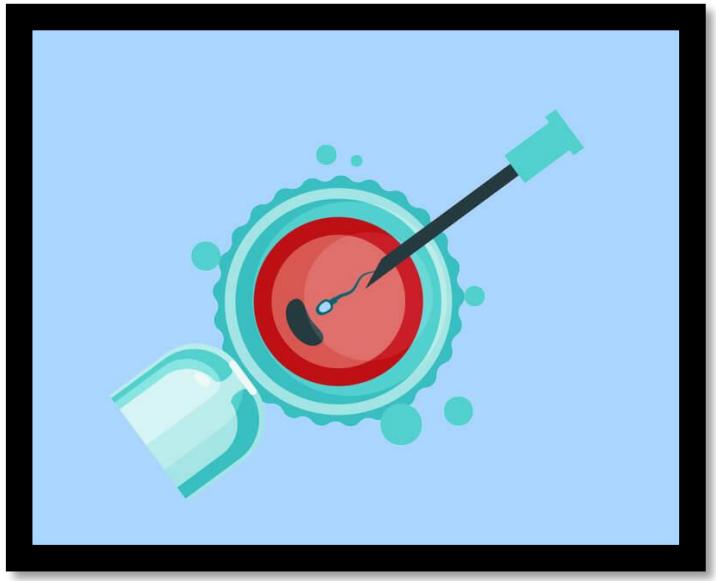
[2] Pacific Centre for Reproductive Medicine, Burnaby, Canada

SFU

Background

In Vitro Fertilization (IVF)

- Commonly performed medical procedure for couples suffering from infertility
- 30,000+ treatments performed annually in Canada; 230,000+ performed in U.S. [1]
- Steadily increasing with prevalence of reproductive issues and procreating later in life



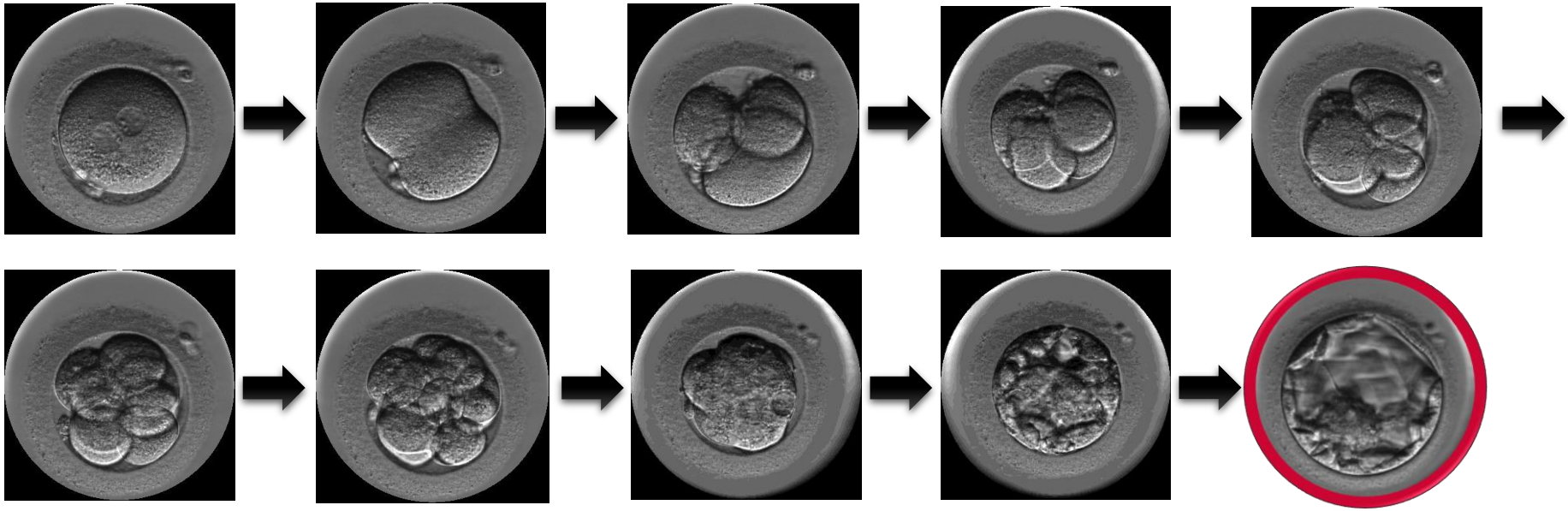
https://www.mcrmfertility.com/

Low success rate: only 29.4% of implantations led to live birth

Background

In Vitro Fertilization (IVF)

- Eggs and sperm are collected and combined in a lab
- Embryos developed for 5 days (until blastocyst stage)
- Highest quality blastocyst(s) transferred to patient



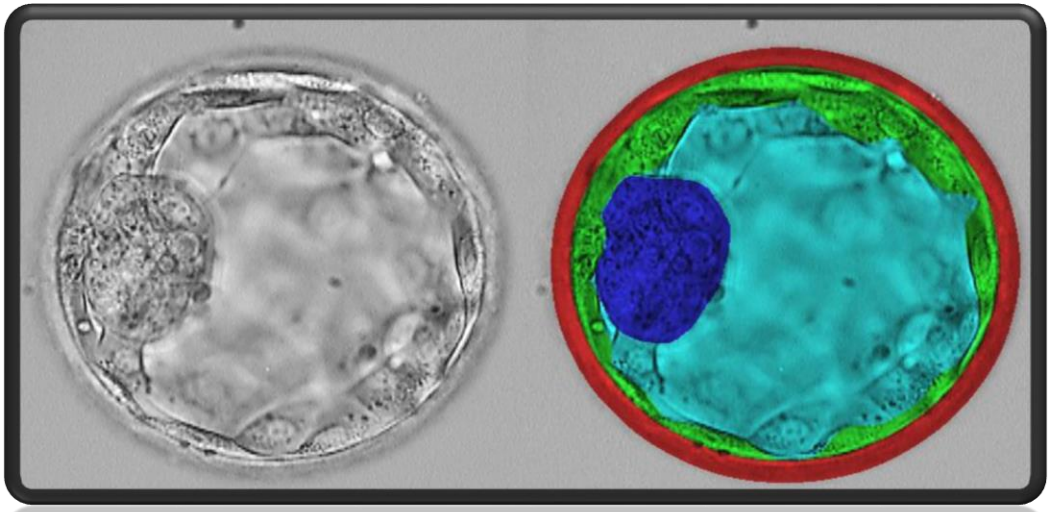
SFU

Problem Description

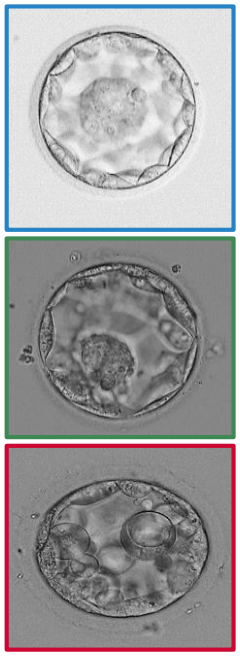
Automatically Assess Blastocyst Quality

- Quality assessed by visually inspecting blastocysts
  - Contrast microscopy imaging – non-invasive
- Blastocysts scored according to Gardner grading system
- Grading based on morphological components

- **Zone pellucida (ZP)**
- **Trophectoderm (TE)**
- **Inner Cell Mass (ICM)**
- **Blastocoel**

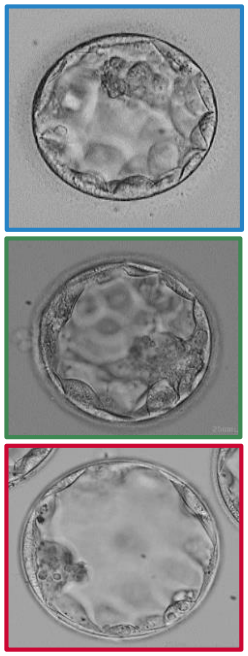


Problem Description



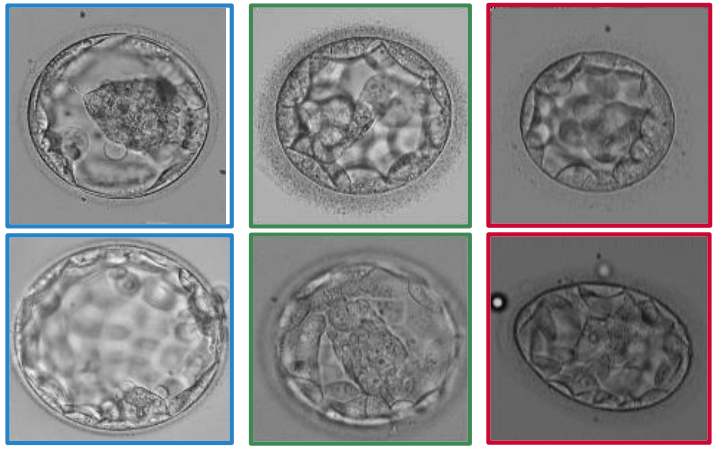
ICM score
- A (blue)
- B (green)
- C (red)

TE score



	ICM Grade
A	Numerous and tightly packed cells
B	Several and loosely packed cells
C	Few cells
	TE Grade
A	Many cells organized in epithelium
B	Several cells organized in loose epithelium
C	Few cells
	ZP (Expansion) Grade
4	Blastocoel volume larger than early embryo, ZP is thin
3	Blastocoel fills the blastocyst, ZP is thick
2	Blastocoel fills greater than half of the blastocyst

ZP (Expansion) score:
- 4 (blue)
- 3 (green)
- 2 (red)



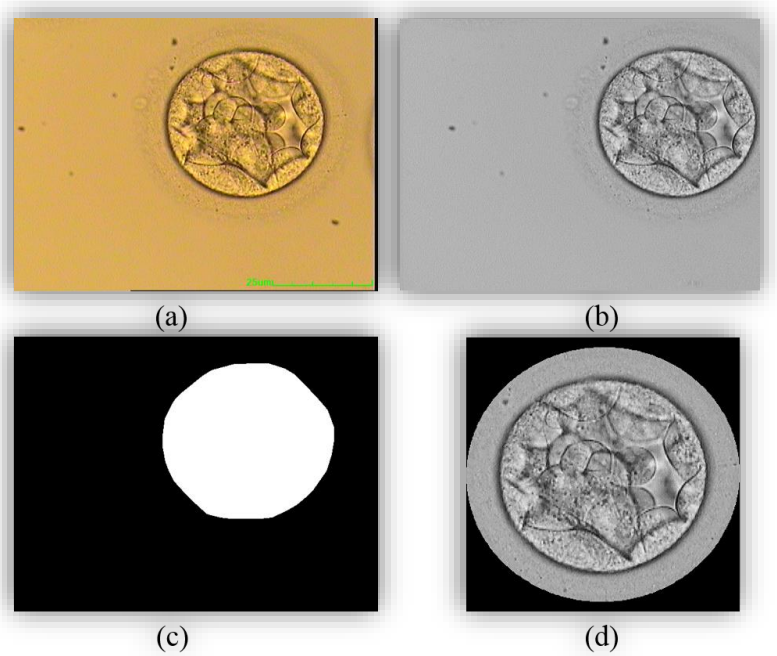
Related Work

- ✓ Blastocyst segmentation using image processing and deep learning
- ✓ Medical image classification using deep learning
  - ✓ Classification on more closely related images using small datasets
  - ! Typically binary or multi-class classification; **balanced datasets**
- ✓ Automatically classify blastocyst images into 'good' or 'poor' quality
  - ! Many blastocysts in a single batch are 'good' or 'poor'
  - ! How can they be ranked according to most important factors??
- ✗ No method for automatic blastocyst quality assessment with granular labels
  - ! More comprehensive assessment includes multiple grades (ICM, TE, ZP) each with multiple scores (A/B/C, A/B/C, 4/3/2)
  - ! Severely imbalanced training data (<5% samples in minority class)

Methodology – Pre-processing

Small dataset with significant variation within classes:

- Location/shape of blastocyst
- Fragmentation in background
- Neighbouring blastocysts



Reduce variation in data with pre-processing

a) Original image

b) Remove image borders and scale bar

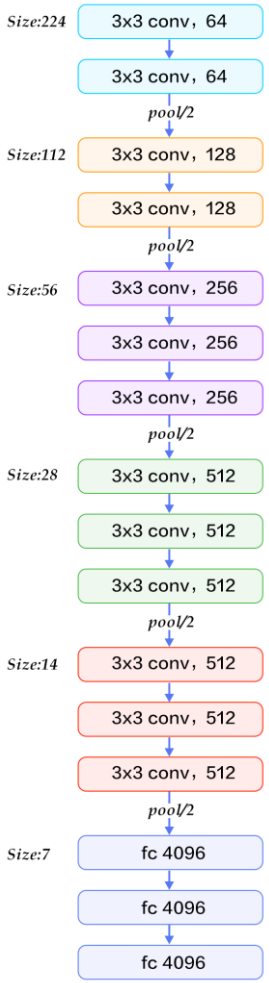
c) Sobel edge filtering and morphological operations

d) Best-fitting ellipse to blastocyst mask; crop to center

Methodology – Baseline Model

Transfer learning

1. Learn convolution kernels by training on large image classification dataset
   - VGG16, ResNet50, InceptionV3 models
   - Network effectively separates images into 1000 classes
2. Fine-tune on small blastocyst image dataset
   - Replace fully connected top layers with small dense network
   - Freeze most layers during second training period



VGG16 Architecture

Output Formulation

Single grade classification

- ICM or TE or ZP scoring

Multi-class classification:

- 3×1 output vector

X Requires 3 separate networks!

$$\begin{Bmatrix} A\ or\ 4 \leftrightarrow 0 \leftrightarrow [1 & 0 & 0] \\ B\ or\ 3 \leftrightarrow 1 \leftrightarrow [0 & 1 & 0] \\ C\ or\ 2 \leftrightarrow 2 \leftrightarrow [0 & 0 & 1] \end{Bmatrix}$$

Multiple grade classification

- ICM and TE and ZP scoring

Multi-class classification:

- 27×1 output vector

X Heavily imbalanced sample data!

$$\begin{Bmatrix} AA4 \leftrightarrow 0 \leftrightarrow [1 & 0 & 0 & \cdots & 0] \\ AA3 \leftrightarrow 1 \leftrightarrow [0 & 1 & 0 & \cdots & 0] \\ AA2 \leftrightarrow 2 \leftrightarrow [0 & 0 & 1 & \cdots & 0] \\ \vdots \quad \vdots & & \vdots & & \\ CC2 \leftrightarrow 26 \leftrightarrow [0 & 0 & 0 & \cdots & 1] \end{Bmatrix}$$

Multi-label multi-class classification:

- 3×3×1 output vector

✓ **Best balance of class distributions**

Concat

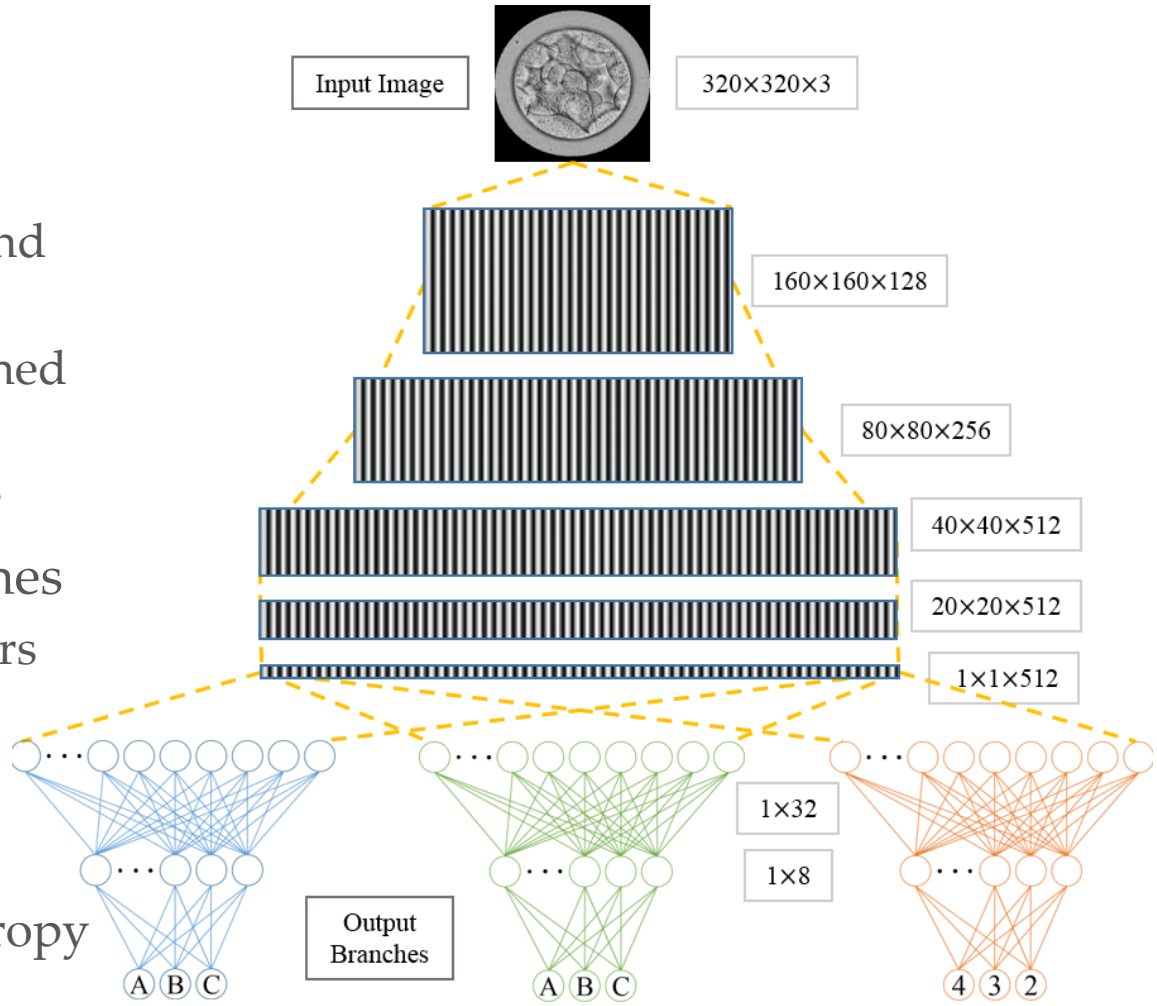
$$\left(\begin{Bmatrix} A \leftrightarrow 0 \leftrightarrow [1 & 0 & 0] \\ B \leftrightarrow 1 \leftrightarrow [0 & 1 & 0] \\ C \leftrightarrow 2 \leftrightarrow [0 & 0 & 1] \end{Bmatrix}, \begin{Bmatrix} A \leftrightarrow 0 \leftrightarrow [1 & 0 & 0] \\ B \leftrightarrow 1 \leftrightarrow [0 & 1 & 0] \\ C \leftrightarrow 2 \leftrightarrow [0 & 0 & 1] \end{Bmatrix}, \begin{Bmatrix} 4 \leftrightarrow 0 \leftrightarrow [1 & 0 & 0] \\ 3 \leftrightarrow 1 \leftrightarrow [0 & 1 & 0] \\ 2 \leftrightarrow 2 \leftrightarrow [0 & 0 & 1] \end{Bmatrix} \right)$$

SFU

Methodology – Proposed Model

Network diagram

1. VGG16 backbone
   - Shared convolution and pooling layers
   - Filter kernels pre-trained on ImageNet
   - Last 3 layers trainable
2. Multiple output branches
   - Individual dense layers
   - Trained in one pass
3. Cost function applied to each output branch
   - Categorical cross-entropy



Methodology – Stratified Sampling

Percentages of class samples are heavily skewed

- Image data collected at Pacific Centre for Reproductive Medicine in Burnaby, Canada from 2012-2018
- Data available only for embryos that were implanted
  - 1-2 highest quality embryos from an entire batch
- Greater percentage of A/A/3-grade blastocysts for ICM/TE/ZP
  - 1-2 highest quality embryos from an entire batch

ICM		TE		ZP (Expansion)	
A	**507 (72.0%)**	A	**382 (54.3%)**	4	248 (35.2%)
B	190 (27.0%)	B	268 (38.1%)	3	**300 (42.6%)**
C	7 (1.0%)	C	54 (7.6%)	2	156 (22.2%)

Methodology – Stratified Sampling

Why is data imbalance a problem?

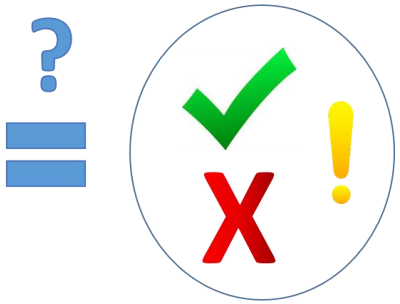
- Imbalanced data leads to different class sampling rates
  - X Majority classes are oversampled during training
  - X Minority classes are undersampled during training
  - X Choice of dataset splits can vastly influence performance

How to ensure network is trained and tested fairly?

- Require a minimum amount of samples be present in each dataset split
  - ✓ Partition samples according to severely imbalanced ICM label
    - Fixed number of samples in minority class
    - Fixed percentage of samples in majority/middle classes
  - ✓ Ensures all classes are represented

Experimental Setup

- Images resized to 320×320 pixels
- Converted to RGB format
- Normalization sample-wise
  - Mean subtraction
  - Division by standard deviation
- Random data augmentations:
  - Rotation
  - Horizontal/vertical shifts
  - Zooming
  - Shearing
- Trained with RMSprop optimizer
- Initial learning rate: 10^{-6}
  - Learning rate reduced by 0.3 on plateau
- 3-fold cross-validation
  - 70% samples in training set
  - 15% in validation set
  - 15% in test set



Experimental Results

Granular Results - Confusion Matrices

- Blue cells represents correct predictions
- **Bold** text represents highest number of correct predictions for that class across networks

Model		Actual Classes										
		Label	*ICM*			*TE*			*ZP (Expansion)*			
		Class	A	B	C	A	B	C	Class	4	3	2
Predicted Classes	ResNet50 [18] - 3×Single-Label	A	116	50	2	121	82	20	4	18	58	20
		B	83	**24**	3	40	28	9	3	11	21	9
		C	29	10	**1**	12	6	0	2	33	61	0
	InceptionV3 [19] - 3×Single-Label	A	203	79	6	89	93	10	4	6	7	11
		B	18	2	0	60	53	13	3	25	60	54
		C	7	3	0	0	0	0	2	36	76	43
	VGG16 [17] - 3×Single-Label	A	**228**	84	6	103	46	2	4	29	11	0
		B	0	0	0	66	**78**	23	3	45	103	18
		C	0	0	0	0	0	0	2	0	24	88
	VGG16 - Multi-Label (Proposed)	A	214	63	4	**151**	57	0	4	**57**	9	0
		B	14	21	2	21	59	19	3	17	**113**	11
		C	0	0	0	0	7	**4**	2	0	21	**90**

Experimental Results

General Results

Global accuracy

- Ratio of correct predictions to total number of samples

Macro-weighted precision and recall

- Precision and recall calculated individually for each class, then averaged across 3 classes

Model	*ICM*		
	Accuracy	Precision	Recall
ResNet50 [18] - 3×Single-Label	44.3	30.3	32.0
InceptionV3 [19] - 3×Single-Label	64.5	26.8	30.5
VGG16 [17] - 3×Single-Label	71.7	23.8	33.3
VGG16 - Multi-Label (Proposed)	**73.9**	**44.3**	**39.6**

Model	*TE*		
	Accuracy	Precision	Recall
ResNet50 [18] - 3×Single-Label	46.9	30.2	31.4
InceptionV3 [19] - 3×Single-Label	44.6	29.5	32.0
VGG16 [17] - 3×Single-Label	56.9	38.3	41.3
VGG16 - Multi-Label (Proposed)	**67.3**	**56.1**	**51.0**

Model	*ZP (Expansion)*		
	Accuracy	Precision	Recall
ResNet50 [18] - 3×Single-Label	23.9	24.9	25.3
InceptionV3 [19] - 3×Single-Label	34.3	32.0	30.2
VGG16 [17] - 3×Single-Label	69.2	71.0	65.6
VGG16 - Multi-Label (Proposed)	**81.8**	**82.5**	**81.7**

Discussion

1. Large networks (ResNet50 and InceptionV3) could not fine-tune on a small dataset
   - Overfitting during training led to poor test results
   - VGG16 had similar training and test performance
2. Deep learning models learn best using balanced data
   - Best classification performance on ZP grade label
   - Lower accuracy and precision/recall on ICM and TE labels
3. Bias towards majority class reduced in multi-label network
   - Less difference between accuracy and precision/recall
   - Baseline models precision < recall on ICM and TE labels
   - Proposed model precision > recall on ICM and TE labels

Conclusions & Future Work

- ✓ Multi-label multi-class blastocyst image classification with deep neural network
  - ✓ Performance improved by jointly learning to classify images in multiple grade labels
  - ✓ Parameter sharing led to more robust predictions
  - ✓ Stratified sampling enabled fair training and testing
  - ! Still exhibited weak performance for minority classes
- ➢ Train on all samples in IVF cycle – not just those implanted
- ➢ Utilize segmentation maps to associate grade information with different blastocyst regions
- ➢ Use images and grades to predict blastocyst implantation potential

References

1) National Center for Chronic Disease Prevention and Health Promotion: Division of Reproductive Health, “2015 Assisted Reproductive Technology National Summary Report,” 2015.