INVESTIGATE CANCER
A STEM INTEGRATED AP BIOLOGY UNIT

DAY 1 - 3
GENETIC-BASIS OF CANCER

Topics and Related Student Activities:
- Cell cycle function
  - Digital Activity
- Oncogenes and tumor suppressor genes
  - Card sort and classification
  - Identification on chromosomes
- Mutations and effects of mutations
  - Reading and response on factors that cause mutations

DAY 4
PHYSICAL DESCRIPTION OF TUMORS

Topics and Related Student Activities:
- How cell cycle function causes tumors
- Tumors and nodules (specifically lung)
  - Reading and response on physical characteristics of nodules
- Characteristics of benign and malignant nodules
  - Card sort of CT Images- "Benign or Malignant?"
  - Data analysis of sort results
- Imaging tools for identification of nodules

DAY 5
STEM CAREER CONNECTION

Topics and Related Student Activities:
- Using computer programming to evaluate CT Scans
- Computer Vision and CT Scans
- Machine Learning for nodule classification
- Data output and the meaning of those data
  - "How Well is the Model Working?" statistical analysis sheet
- Other connections to careers in computer vision and machine learning

Final Student Reflection
- Potential Field Trip to UCF

Integrating Computer Vision in AP Biology

@ Lake Nona High School

By: Valerie Ledford

Integration Theme: How Low-Dosage Computed Tomography (LDCT) Scans and Computer-Assisted Detection (CAD) systems aid in accurate identification of potentially life threatening lung cancer nodules.

Mission

As part of Research Experience for Educators (RET) the primary mission of the project is to expose Advanced Placement Biology high school students to the field of computer vision by providing them with a learning experience that both advances their knowledge in a specific biological topic while simultaneously integrating concepts and tools typically encountered in the field of computer vision.

ACCURACY OF LUNG CANCER DIAGNOSIS- HUMAN VISION VS. "COMPUTER VISION"

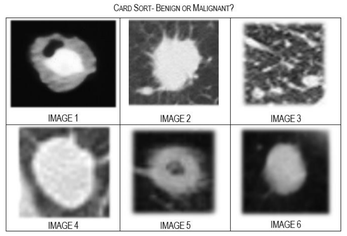
"THE HUMAN MODEL"

Students classify nodules using the table of characteristics and run statistical analyses on the results.

Table 2. Evaluating Probability of Malignancy using Characterization of Nodules

	Likely Benign	Possibly Malignant	Highly Likely Malignant
Diameter	< 1.5 cm	1.5 - 2.2 cm	≥ 2.3
Nodule Margins (Border)	Smooth	Scalloped	Corona radiate or spiculated
Cavitation (Hole/dark in the center)	Thin Wall		Thick Wall
Calcification (Calcium Deposits)	Concentric, Central, lamellar, popcorn or homogeneous patterns (Diffuse, Central, Laminated, Popcorn)		Non-calcified or eccentric calcification
Density	Solid	Subsolid	Nonsolid = "ground glass"
Shape	Round, Oval or Polygonal		Irregular

CARD SORT- BENIGN OR MALIGNANT?



1. Accuracy- $\frac{\sum TP + TN}{\sum P + N\,(TP + TN + FP + FN)}$
2. Sensitivity- $\frac{\sum TP}{\sum TP + FN}$
3. Specificity- $\frac{\sum TN}{\sum TN + FP}$
4. Precision- $\frac{\sum TP}{\sum TP + FP}$

TP = TRUE POSITIVE (Known = Predicted)

TN = TRUE NEGATIVE (Known = Predicted)

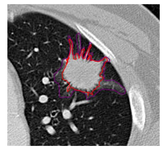
FP = FALSE POSITIVE (Known = Negative; Predicted = Positive)

FN= FALSE NEGATIVE (Known = Positive; Predicted = Negative)

Table 1: Confusion Matrix

True Positives *Total predicted as positive (malignant) and labeled as positive (malignant)*	False Positives *Total predicted as positive (malignant) and labeled as negative (benign)*
False Negatives *Total predicted as negative (benign) and labeled as positive (malignant)*	True Negatives *Total predicted as negative (benign) and labeled as negative (benign)*

VERSUS



Lesson Objective: Students will explain how computer vision and machine learning tools can improve accuracy and speed of lung cancer diagnoses.

"THE COMPUTER MODEL"

Students run statistical analyses on the results of the program and compare them with their results from the "human model".

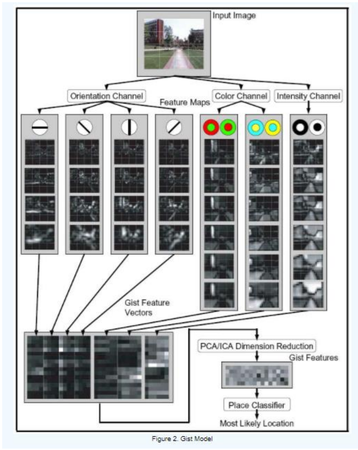
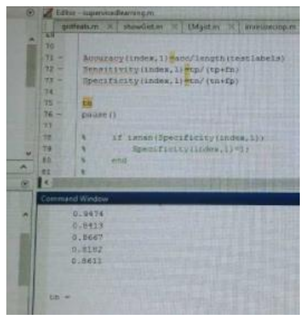


Figure 2. Gist Model



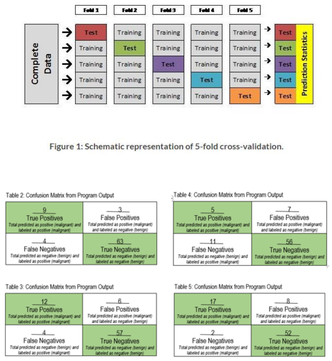


Figure 1: Schematic representation of 5-fold cross-validation.

Table 2: Confusion Matrix from Program Output

9 True Positives	3 False Positives
4 False Negatives	83 True Negatives

Table 4: Confusion Matrix from Program Output

5 True Positives	7 False Positives
11 False Negatives	56 True Negatives

Table 3: Confusion Matrix from Program Output

12 True Positives	8 False Positives
4 False Negatives	57 True Negatives

Table 5: Confusion Matrix from Program Output

17 True Positives	8 False Positives
2 False Negatives	52 True Negatives

EXTENSION GOALS

1. Help students understand that "science" is not done in a vacuum- there are connections to mathematics, technology, etc. AND there is a huge collaboration piece
2. Have interested students meet on a Saturday at the College of Computer Vision for tour and information
3. Provide a resource for students interested in computer vision focused science fair project
4. Make connections to computer vision technology in other areas of the curriculum (bacteria counting, population analysis, etc.)

Predicting Cancer Mutation Rate

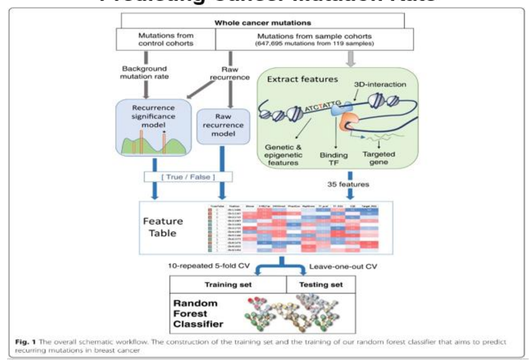


Fig. 1 The overall schematic workflow. The construction of the training set and the training of our random forest classifier that aims to predict recurring mutations in breast cancer

AP BIOLOGY LEARNING OBJECTIVES:

LO 3.6 . . .predict how a change in a specific DNA or RNA sequence can result in changes in gene expression

LO 3.7 . . .make predictions about natural phenomena occurring during the cell cycle.
LO 3.8 . . .describe the events that occur in the cell cycle.

LO 3.21 . . .use representations to describe how gene regulation influences cell products and function.

LO 3.22 . . .explain how signal pathways mediate gene expression, including how this process can affect protein production.
LO 3.23 . . .use representations to describe mechanisms of the regulation of gene expression.

AP BIOLOGY REQUIRED SCIENCE PRACTICES:

SP 2.1 . . .justify the selection of a mathematical routine to solve problems.
SP 2.2 . . .apply mathematical routines to quantities that describe natural phenomena.

SP 5.1 . . .analyze data to identify patterns or relationships.
SP 5.2 . . .refine observations and measurements based on data analysis.
SP 5.3 . . .evaluate the evidence provided by data sets in relation to a particular scientific question.

REFERENCES

American Cancer Society. (2014). *Cancer Facts & Figures 2014*. Atlanta, GA: American Cancer Society, Inc.

Cancer Commons (2013, April 10). Rates of Misdiagnosed Lung Cancer Vary Widely. Retrieved July 6, 2017 from https://www.cancercommons.org/news/rates-of-misdiagnosed-lung-cancer-vary-widely/

Geddes DM: The natural history of lung cancer: A review based on rates of tumor growth. *Br J Dis Chest* 73:1-17, 1979.

Golden Helix, Inc. (2015, April). Cross-Validation for Genomic Prediction in SVS. Retrieved July 6, 2017 from http://blog.goldenhelix.com/goldenadmin/cross-validation-for-genomic-prediction-in-svs/

Movsas, B., Brahmer, J., & Paller, C. Kernstine, K.H. (2001). Non-small cell lung cancer. *Cancer Network: Cancer Management*, 1-43.

Moyer, V.A. (2014). Screening for lung cancer: Recommendations from the U.S. Preventive Services Task Force. *Annals of Internal Medicine, 160*(5), 1-40. doi:10.7326/P14-9009

National Cancer Institute (NIH). (2014, September). National Lung Screening Trial. Retrieved July 6, 2017 from https://www.cancer.gov/types/lung/research/nlst

The National Lung Screening Trial Research Team. (2001). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine, 365*(5), 395-409. doi: 10.1056/NEJMoa1102873

Tat, M.J. Seeing the Random Forest from the Decision Trees: An Explanation of Random Forest. Towards Data Science, retrieved July 6, 2017 from https://medium.com/towards-data-science/seeing-the-random-forest-from-the-decision-trees-an-intuitive-explanation-of-random-forest-beaa2d6a0d80

Acknowledgements

This project is part of the SHAH RET Program, College of Engineering and Computer Science, University of Central Florida.

Special thanks to Dr. Niels Lobo, AliAsghar Mortazi, Dr. Mubarak Shah, Josue Urbina