Name(s)
Pranav D. Atreya

Project Number
S0801

Project Title
Portable Ultrasound Organ Tomography for Early Detection of Tumors and Blockages

Abstract

Objectives/Goals
Certain tests such as mammographies and colonoscopies are recommended for people every few years when they reach a certain age to detect medical conditions before symptoms arise. However the national percentages of people who commit to these tests are far below desired numbers due to inconvenience, cost, and the lack of local medical facilities. The objective of this project is to promote rates of early detection by creating a home-use ultrasound screening system for tumors and other types of deleterious conditions.

Methods/Materials
The system is built with hardware components such as a Raspberry Pi 3, an ultrasound transducer, and associated electrical equipment that together interact with custom developed Image Formation and Image Classification software. These software components were developed with the Python language to be compatible with external hardware. The Image Classification component was built using a Convolutional Neural Network developed with the TensorFlow library and trained on ultrasound images from scans of healthy body organs and organs with tumors, lymph node growth, or cysts present. The final scanning system was tested on models designed to simulate the medical conditions in observation.

Results
Testing was performed first on individual components of the system and ultimately on the final ultrasound diagnostic device. The ultrasound scanner hardware and the Image Formation software were tested on custom-built models that were designed to simulate the physical characteristics of tumors, cysts, and other medical conditions. The device was successful in its ability to perform scans and form a B-Mode ultrasound image from the data collected by the scan. This image was then passed to the ultrasound Image Classification software, which was able to determine the presence of a medical anomaly and classify what type of anomaly was present with 94% training accuracy and 77% test set accuracy.

Conclusions/Discussion
When tested on the models, the final diagnostic system was able to classify the ground truth with good accuracy. The system as a whole is portable and can be used for home-use. Ultimately this system will provide a person with the ability to perform small ultrasound diagnostic scans whenever they deem fit, and the results of the scan can prompt consultation of a doctor, thus improving rates of early detection.

Summary Statement
Built a home-use ultrasound diagnostic device and software to alert the user of suspicious growth for the purpose of improving rates of early detection.

Help Received
The work in this project was completed individually at home. I referenced a variety of sources for research material and used an online ultrasound image database for neural network training.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Srinivas Balagopal

Project Number
S0802

Project Title
The Effect of Nonlinearity on Recalibrating the AQI and Air Pollutant Forecasts for 3 Bay Area Urban Micro-climates

Abstract

Objectives/Goals
The Bay Area Air Quality Management District (BAAQMD) uses ozone (O(3)) and particulate matter (PM(2.5))readings to issue their daily Air Quality Index (AQI), eclipsing the impact of other criteria pollutants in urban microclimates. Current AQI forecasts also use linear and deterministic models that belie the impact of nonlinear pollutant interactions.Thus, my hypotheses are that (1) using nonlinear correlations and the Analytical Hierarchical Process (AHP) will result in an aggregated AQI accuracy of over 50% against the EPA's current AQI; (2) using nonlinear neural networks will produce accurate hourly pollutant forecasts.I obtained 3 years of hourly meteorological and pollutant data from BAAQMD for San Jose, San Francisco, and Oakland.

Methods/Materials
I applied Spearman's Rho to derive monotonic coefficients between the met factors and pollutants, which I used to develop an AHP that derived a weighted AQI scale for each pollutant per microclimate in Excel.This AQI was tested for accuracy of pollutant impact against the control EPA AQI. I applied the coefficients to enrich BAAQMD data to construct pollutant/microclimate-specific LSTM networks in Python to forecast hourly pollutants.The forecast results were tested for accuracy against the control BAAQMD forecasts and against actual pollutant data.

Results
The correlations showed that primary pollutants (CO, NO(2), & SO(2)) had greater impact on pollutant levels than met factors. My AHP-based AQI showed that 58.3% of EPA control AQIs reduce the impact of other pollutants, despite their having higher concentrations. My LSTM models increased the forecast accuracy by 57.3% for winter PM, 10% for summer O(3), and 8% for fall O(3), as compared to BAAQMD's control forecasts. Finally, annual forecasts were 96% accurate as tested against BAAQMD pollutant records.

Conclusions/Discussion
80% of global urban populations live in substandard air environments affected by anthropogenic primary emissions, compounded by topographical factors.My revised AQI provides an accurate pollutant representation for these microclimates that inform health impacts and empirically highlight the true sources of pollution.My nonlinear forecasts prove that dynamic urban environments are unsusceptible to linear and deterministic forecasting models and that using pollutant-centric nonlinear models provide accurate forecasts that allow individuals to plan their daily activities.

Summary Statement
Using nonlinear correlations, I proved that primary air pollutants have a higher impact on urban microclimates than meteorological factors, which accurately represents aggregate air quality and allows for robust pollutant forecasting.

Help Received
Mr. Daniel Alrick (BAAQMD) for answering questions and providing data and BAAQMD forecasts; Mr. A. Santhanam (Pinewood) for help on statistical analysis; and Mr. Haggai Mark (Pinewood) for his support.

Name(s)
Akhilesh V. Balasingam

Project Number
S0803

Project Title
Gossamer: A Monte Carlo Simulator for the Optimal Design of Nanowire Networks for Transparent Electrode Applications

Abstract

Objectives/Goals
I have developed a simulator to optimize the performance of a class of electrically conductive, transparent and flexible thin films. Such materials are needed to construct components such as touchscreens and solar panels that are non-rigid and shape-conformable.

I focus on films manufactured using "solution-processing" which creates random metal nanowire (MNW) networks. Large-scale experimental characterization of the resulting sample-to-sample variability of film properties is costly and time-consuming. Gossamer, the MC simulator I developed, responds to this challenge, allowing scientists to separate systematic effects of controllable manufacturing parameters, from statistical variability.

Methods/Materials
I wrote Gossamer in Java and it consists of three modules: (a) a geometry engine which generates a random MNW collection, computes their intersections and creates a list of MNW segments, (b) a network analyzer which identifies connected clusters of MNW segments using depth-first search, and (c) a circuit solver which computes the overall resistance of the network. I used Python for post-processing and visualization.

Results
I demonstrate three applications of Gossamer: (i) characterize current hotspots within the film which cannot be explored experimentally (ii) compute film resistance as a function of wire length (Le), diameter (Di) and areal wire mass density (phi_md), and (iii) optimize film conductivity under different tradeoff conditions (e.g., Le vs. Di) subject to constant phi_md constraints.

Key findings include: (a) when film resistance is plotted as a function of phi_md, the data falls on a nearly-universal dimension-independent L-shaped curve, which correlates well with experimental data, published recently (2015) by Lagrange and Langley (b) for each choice of MNW mass density, an optimal choice of wire dimensions exists for minimizing film resistance.

Conclusions/Discussion
I developed a physics-based simulator to model the resistance and internal state of a class of conductive and transparent nanowire-based films, validated it against recently published experimental data and used it to compute geometrical parameters that optimize overall film conductance, under constant mass constraints.

Summary Statement
Gossamer, my MC simulator, helps accelerate the ongoing search for transparent conductive materials which are needed for the construction of a wide range of emerging large-area and flexible electronic devices.

Help Received
I developed the code and performed the simulations and analysis on my own. I would like to thank my mentor for discussing the current experimental literature and pointing out outstanding challenges in the field. I would like to thank my school math teacher for helpful discussions and encouragement.

Name(s)

Prachi Bhagavatha; Jasmine Ngo

Project Number

S0804

Project Title

Deep Learning Real-Time Object Detection through Convolutional Neural Networks Using OpenCV for the Visually Impaired

Abstract

Objectives/Goals

The objective of our BerriVest device is to accurately do object detection by alerting the user of frontal obstacles and uneven surfaces and image processing, by specifically naming the obstacle(s) in an arbitrary environment.

Methods/Materials

Raspberry Pi 2 (Python 3.4.2), OpenCV 3.4, numPy software libraries, external power supply (5V battery), pi NoIR camera (8 megapixels), 2 HC-SR04 ultrasonic sensors, 6 M/F premium jumper wires, 2 1kohm resistors, 2 2kohm resistors, soldering gun, lead, headphones/earbuds, protective case for Raspberry Pi, LiDAR sensor, IR sensor, Adafruit Ultimate GPS Tracker, and Caffe model. Tested fifteen people in arbitrary environments, such as the living room, kitchen, garage, and outside in the neighborhood, where the BerriVest device detected objects to investigate the accuracy of the implementation of the convolutional neural network models we used: ImageNet and GoogleNet.

Results

With only the ImageNet model, we had each of the 15 test subjects conduct three trials for each of the nine obstacle detections for a bicycle, chair, car, person, dining table, sofa, TV monitor, stop sign, and fire hydrant. Then, after adding the GoogleNet model in parallel with the ImageNet model, we had our test subjects walk around arbitrary environments, such as their living room, kitchen, garage, and outside in the neighborhood. Results depicted that for 11 out of the 15 subjects, there was a high percentage of the actual object accuracy, which proves the consistency of the BerriVest in detecting various obstacles.

Conclusions/Discussion

We embedded artificial intelligence by experimenting with two pre-trained models to train the neural network into our program, so the BerriVest can efficiently do image processing and name exactly what the obstruction is in front of the user. We first successfully experimented with the MobileNet model and then added, in parallel, the GoogleNet model. With the GoogleNet model, the BerriVest can now detect objects under thousands of more detailed classifications due to its large database. We concluded that implementing supervised learning on Caffe yielded accurate and faster image processing -- reducing the lag time -- and object detection that enhanced our final product.

Summary Statement

We created a hands-free device and programmed it using Python to implement multiple optical flow algorithms using convolutional neural networks that detects frontal obstacles and uneven surfaces in live feed, aiding the visually-impaired.

Help Received

My partner and I designed the components of the BerriVest device and built it ourselves. We received help and supervision from Mr. Raghavendra Bhagavatha with the coding and understanding of the concepts of machine learning.

Name(s)
Cynthia Chen

Project Number
S0805

Project Title
Type 2 Diabetes Prediction Using Longitudinal Machine Learning Analyses and Integrative Personal Omics Profiling

Abstract

Objectives/Goals
Type 2 Diabetes (T2D) affects more than 200 million people worldwide. Steady-state plasma glucose (SSPG) values are crucial to determining T2D as they indicate a patient's insulin resistance. The goal of our project was to analyze large longitudinal omics datasets using machine learning and statistical analyses in order to accurately predict SSPG values for pre-diabetic patients.

Methods/Materials
Our dataset consists of SSPG values, BMI, and 3500 omics features for 23 patients at 4 different time-points. We preprocessed our data in three stages: 1) PCA dimensionality reduction for feature selection, 2) taking derivatives between consecutive time-points to preserve the longitudinal time sequence, and 3) data normalization and standardization.

For SSPG value prediction using machine learning, we developed 10 different classification algorithms and 5 regression algorithms in Python, and tested these algorithms for optimal performance. We also generated correlation matrices among the omics datasets and determined the most correlated feature pairs as well as the optimal microbiome taxonomy depth levels. Using these feature correlation analyses, we improved the prediction performance of the regression and classification models.

Results
AdaBoost classification achieved an accuracy rate of 87.5%, and LASSO regression performed the best with a root mean square error (RMSE) of 24.765. We improved these results to 90.0% accuracy and 22.455 RMSE by using the feature correlation analyses described above.

Conclusions/Discussion
We concluded that our computational model was successful in accurately predicting T2D for pre-diabetic patients. Our project has major implications in the medical field, as our novel longitudinal time sequencing and feature correlation methods can provide new, improved pathways for disease prediction. In the future, we would like to expand upon our project by analyzing more datasets, such as genomics and RNA sequencing.

Summary Statement
We created an accurate computational model using machine learning methods and statistical analyses to predict SSPG values for pre-diabetic patients.

Help Received
I worked at the Stanford Laboratory of Quantitative Imaging under the guidance of Dr. Imon Banerjee and Prof. Daniel Rubin. I developed the methods and algorithms independently and received help from my mentors.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Andrew C. Chiang

Project Number
S0806

Project Title
Automatic Basketball Shooting Trajectory Analysis

Abstract

Objectives/Goals
The objective of the project is to build a system to detect the basketball trajectory and provide feedback of the incident angle to the shooter.

Methods/Materials
I bought a stereo camera kit and built a custom frame to mount it. This allowed me to use a baseline of 39 cm which is much wider than most of the commercially available stereo cameras. I developed my programs on Intel i7 based PCs using C++ and OpenCV library.

I used the functions in the OpenCV library to calibrate my stereo camera. I evaluated many different object detecting techniques, and found that cascade detectors had the most promise in addressing my needs. I created an image library and annotated the sample images to train my own cascade models. I used cascade detectors to detect the backboard, rim, and basketball. I developed a custom algorithm to fine-tune the basketball location in the image. By using the reprojection matrix of the stereo camera, I could use the detected pixel coordinates from left and right images to find the basketball position in 3D space. I then wrote a program to collect all the sample basketball positions in the trajectory frame-by-frame, and fitted the samples to a parabola that minimized square errors. The parabola had the best fit to the projectile trajectory, and the incident angle can be derived from the parabola formula.

Results
I found that it was important to group foreground samples with different shot angles. The most critical factor was using the LBP (local binary patterns) feature type instead of the default HAAR feature type in training cascade models. After training the cascade models with LBP feature type, my program was able to detect the backboard, rim, and basketball. The RMS (root mean square) error of the trajectory detected by the stereo vision was about 7.4 cm at a distance of 7.9 m. I was able to fit the detected trajectory to a parabola and estimated the incident angle.

Conclusions/Discussion
Cascade models with LBP feature type were used to automatically detect the camera location relative to the backboard and rim, and detect the basketball trajectory. The trajectory was fitted to a parabola in order to estimate the incident angle for providing feedback to the shooter.

Summary Statement
I have developed a system that can automatically detect the camera location on the court, detect the basketball trajectory in order to estimate the incident angle, and provide feedback to the shooter.

Help Received
I wrote the programs and developed the algorithms myself. I used the OpenCV library extensively. I found answers through web searches for most of my programming questions.

Name(s)
Bryan H. Chiang

Project Number
S0807

Project Title
Illuminating Gene Dysregulation in Cancer: Deep Learning Identification of Disrupted Transcription Factor Binding Sites

Abstract

Objectives/Goals
Over 90% of mutations associated with cancer lie in the regulatory regions of the genome, driving tumor development by disrupting transcription factor binding - the "on and off switches" of key cell life, growth, and death mechanisms. The purpose of my project was to develop a comprehensive deep learning and statistical framework to pinpoint and characterize sites of irregular transcription factor binding in cancer.

Methods/Materials
Integrating over 50 million DNA sequences with corresponding chromatin accessibility and gene expression data from the ENCODE database, I first constructed high-capacity deep convolutional neural networks (CNNs) to accurately identify genome-wide regions of transcription factor binding. Next, I rigorously screened 1,500 regulatory breast cancer variants regions from the GRASP and HoneyBadger databases for statistically significant regions of differential binding across the previously uncharacterized healthy MCF-10A and cancerous T47-D breast epithelial cell lines, using data provided by the CCLE and GEO. To highlight putative misregulated genes and processes, I performed regulatory gene set enrichment analyses with GREAT. Lastly, I explored downstream roles of the putative dysregulated genes through Ingenuity Pathway Analysis (IPA).

Results
My networks had an average auROC curve score of 98.7%, high sensitivities and specificities (> 90%), and low false positive and false negative rates (< 10%) when evaluated in unseen celltypes. My networks outperformed current state-of-the-art methods by over 15% (auROC). Known binding changes for MYC, SP1, and BRCA1 were confirmed, and more than 300 unique disrupted binding sites across 8 cancer-associated transcription factors were identified ($p<0.05$). I found over 240 putative dysregulated genes and dozens of protein interactions, canonical pathways, and disease functions relevant to cancer progression.

Conclusions/Discussion
To the best of my knowledge, this is the first instance of leveraging deep learning to locate regions of genetic dysregulation in true cancer tissue. My results give us great insight into the key molecular components and mechanisms underlying cancer development that can be further validated through in vitro and in vivo experimentation. My framework also aids the development of clinical applications such as targeted drug therapies, prognostic biomarkers based on abnormal binding patterns, and base reversal technologies.

Summary Statement
I devised a novel high-capacity, integrative deep learning framework to discover over 300 disrupted transcription factor binding sites in cancer, characterizing hundreds of downstream genes and pathways possibly linked to tumor development.

Help Received
Mentored by Irene Kaplow. Questions on deep learning, statistical concepts, and software usage answered by Johnny Israeli, Anshul Kundaje, Daniel Kim, Jin Lee, Vincent Gardeux, Devon Ryan, and Kevin Blighe. Dongwon Lee gave me models to benchmark. Project sponsored by Ms. Nicole Della-Santina.

Name(s)
Lyron O. Co Ting Keh

Project Number
S0808

Project Title
A Novel Hierarchical Machine Learning Model for Non-Invasive Cancer of Unknown Primary Classification

Abstract

Objectives/Goals
Cancer of unknown primary (CUP) is the 4th leading cause of cancer-related deaths worldwide. Patient prognoses can be significantly improved with site-specific therapy. Thus, the objective of this project is to design and train a reliable and cost-effective model to carry out non-invasive tissue-of-origin classification.

Methods/Materials
The chimeric nature of cell-free DNA (cfDNA) in the bloodstream disrupts intraclass distributions and interclass independence assumptions and is subject to low signal-to-noise ratios. This model design is focused on alleviating these challenges with a hierarchical ensemble framework consisting of 3 Support Vector Classifiers (SVCs) and 4 Epsilon Support Vector Regression (SVR) predictors. Each model was trained with interpolated data from 3598 solid tumor profiles and 299 whole blood profiles at various tumor fractions. I conducted Recursive Feature Elimination (RFE) with a variety of random seeds and subsamples to produce a stable reduced feature set and a Grid Search with 5-Fold Cross Validation to tune hyperparameters. These models were integrated into a pipeline that employs the SVRs to infer the tumor fraction of a sample and then feeds the data into the SVCs to restore previously violated assumptions and produce a tissue-of-origin prediction.

Results
This multi-level model predicted tissue-of-origin with 96% accuracy on a withheld test set (n=525), corresponding to a 20% improvement from current non-invasive methods. Furthermore, this pipeline is able to achieve 82% accuracy in the context of early-detection and classification between 6 primary sites, whereas current screening methods only target single cancer types. This performance is maintained when the model is restricted to only 0.8% of the raw feature set, drastically reducing the computational and monetary costs of the test.

Conclusions/Discussion
The results demonstrate the capabilities of this hierarchical approach over standard single-level models when dealing with highly convoluted data with underlying structures, such as cfDNA profiles. This has important implications in treatment decision-making in CUP and in the development of early detection assays. In addition, the novel design and training technique presented in this study can be applied to other problems involving cfDNA such as monitoring cancer progression and observing treatment response to enhance the non-invasive examination of tumors.

Summary Statement
I designed a hierarchical machine learning model that utilizes a simple blood sample to affordably classify cancer of unknown primary with higher efficacy than published methods.

Help Received
Joanne Soo, Dr. David Kurtz, and Dr. Ash Alizadeh from Stanford provided me with guidance on the biological aspects of the project and best-practices when validating my methodology.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Ilan E. Cosman

Project Number
S0809

Project Title
Computer Vision for Detecting Errors in 3D Printing

Abstract

Objectives/Goals
This project aimed to create a computer vision system which can detect errors in real-time during 3D printing, and can pause the printer and notify the user when an error is detected. The errors that can be detected are filament running out, plastic drips, object motion, and extruder jams.

Methods/Materials
The materials and equipment used are a 3D printer (Prusa i3 MK2), PLA plastic filament for printing, a webcam, and a laptop or Raspberry Pi single-board computer for processing. The methods used are various image processing algorithms which were written for this project. Filament running out is detected by counting the number of pixels brighter or darker than a threshold near the top of the printer where the filament crosses contrasting light and dark bars. Plastic drips and object motion are detected by using the 3D object model to create a silhouette of how the object is supposed to look from the viewpoint of the webcam. During printing, the growing object shape is compared against the silhouette to count bad pixels (object pixels that are located where the silhouette says they should not be). Extruder jams are detected by seeing if the object is failing to grow.

Results
The errors that were detected were having the filament run out, an object shifting position during a print, a plastic drip such as that arising from the design having an excessive cantilever, and extruder jams. For each algorithm, when a misprint was detected, the print was paused successfully and an email was sent to the user.

Conclusions/Discussion
The algorithms work correctly and can be useful for preventing wastage of materials during long prints. The methods can be improved by using two cameras to observe different viewpoints, and by finding ways to detect errors in surface texture.

Summary Statement
I created a program that uses a webcam to detect various errors that can occur in real time during a 3D print.

Help Received
None, I coded the entire thing myself.

Name(s)
Rishi M. Desai

Project Number
S0810

Project Title
New Cartographic Network Visualization Technique for Analyzing Alignments of Protein-Protein Interaction Networks

Abstract

Objectives/Goals
In cartographic visualization, nodes are represented as horizontal lines and edges as vertical lines. This method provides advantages over node-link diagrams because nodes and edges cannot overlap. However, there is no software currently available that utilizes cartographic visualization of network alignments. I developed a network visualization tool based on cartographic visualization to analyze the alignments of protein-protein interaction (PPI) networks. This tool was added as a new feature to BioFabric, network visualization software developed by the Institute for Systems Biology, Seattle.

Methods/Materials
I calculated topological measures such as Edge Coverage (EC) and Symmetric Substructure Score (S3) with the alignment of PPI networks rat to yeast. I generated several alignments between the PPI networks yeast2K to yeast5K to analyze objective functions in alignment algorithms. I devised novel measures Node Group Distance (NGD) and Link Group Distance (LGD) to automate topological analysis.

Results
Using the width of link groups, I calculated the EC = .54 and S3 = .41 for the alignment between rat and yeast. The topological similarity between the two PPI networks can be visualized with the relative sizes of link groups. Researchers can alter the alignment so certain nodes do or do not align to each other. Cartographic visualization helps compare topology between the yeast2K to yeast5K alignments generated by different objective functions. I found that objective functions that utilize a combination of measures produce alignments closer to the perfect alignment than those that utilize only one measure. Alignments generated with a combination of measures consistently produced lower NGD and LGD values than those that utilized only one measure such as S3 and Importance.

Conclusions/Discussion
I developed a novel method using cartographic visualization to analyze PPI network alignments. I am the first to use cartographic visualization in the context of network alignments, and any other software currently available uses node-link diagrams. My layout shows topological measures, network connectivity, and allows researchers to improve alignment algorithms. I created novel numerical measures for the automation of topological analysis.

Summary Statement
I developed a novel software tool based on cartographic visualization that allows researchers to analyze topology in the alignments of protein-protein interaction networks.

Help Received
Mr. Longabaugh at Institute for Systems Biology, Seattle, and Prof. Hayes at UC Irvine provided guidance and valuable comments.

Name(s)
Leonardo E. Glikbarg

Project Number
S0811

Project Title
Predicting Terrorist Attacks in Afghanistan Using Generalized Multivariate Regression and Time Series Analysis

Abstract

Objectives/Goals
The objective of this study is to help prevent terrorist attacks in one of the provinces of Afghanistan by discovering the factors that significantly affect the rate at which terrorist attacks occur, and then by using that data to predict terrorist attacks.

Methods/Materials
Laptop computer with RStudio installed. Used data from the United Nations and the Afghan government to create a generalized linear model, ran a time series analysis which generated predictions of future terrorist attacks in Afghanistan.

Results
From the generalized linear model I created, I determined that opium production had by far the strongest correlation to terrorist attacks in Afghanistan out of the nearly 40 covariates I analyzed. I was also able to predict through a time series analysis, that there will be 39, 36, and 42 terrorist attacks respectively in the next three years in the capital of Afghanistan.

Conclusions/Discussion
One way to reduce the number of terrorist attacks in Afghanistan could be to limit the production of opium. Resources, both preventative and retaliatory, can be allocated to regions based on the number of attacks predicted.

Summary Statement
I created a prediction of terrorist attacks in Afghanistan using generalized multivariate regression and time series analysis.

Help Received
I learned the statistical procedures necessary for this project through independent study as well as explanations from Mihnea Andrei, a graduate student at the UCSB Department of Statistics. I used data from the U.N. and the Afghan Govt.

Name(s)
Lan Jiang

Project Number
S0812

Project Title
Defining a New Diagnostic Paradigm in Primary Central Nervous System Hypersomnias through Statistical Machine Learning

Abstract

Objectives/Goals
The multiple sleep latency test (MSLT) is the current gold standard for diagnosing primary central nervous system hypersomnias. While existing thresholds for defining a positive MSLT are sufficient for diagnosing narcolepsy type 1, the arbitrary nature of the thresholds result in both the mischaracterization of over 28% of hypersomnias on initial testing and the poor differentiation of other hypersomnias, thereby negatively impacting treatment efficiency. The study objective was to determine whether better differentiation of primary central nervous system hypersomnias - narcolepsy type 1, narcolepsy type 2, idiopathic hypersomnia - is possible, by incorporating data from preceding polysomnograms and defining new thresholds for the MSLT.

Methods/Materials
Cases from the world's largest hypersomnia database at the Stanford Narcolepsy Center were combined with a control population derived from the Wisconsin Sleep Cohort. Five machine-learning models - stepwise multinomial logistic regression, decision trees, random forests, gradient boosting machine, and recursive partitioning and regression trees - were developed to address the unique multinomial categorization problem. Transparent, reproducible, and comparable methods were then created to adjust for confounders and extract information from the machine learning "black box" to elucidate the mechanisms of each algorithm and thus improve clinical interpretability.

Results
For classification accuracies in the validation set, stepwise multinomial logistic regression performed the best (0.95 vs 0.83-0.88 for other models) and was the only model that had consistently strong category-specific accuracies. In addition to expected MSLT features, new features of interest from the preceding polysomnogram (e.g. total sleep time, N2 percent) greatly improved the ability to differentiate hypersomnias.

Conclusions/Discussion
By incorporating existing clinical information at different thresholds, all models perform excellently at categorization (well above the 25% accuracy expected for chance, with 4 categories) and significantly above current MSLT accuracies. By integrating additional elements from the diagnostic work-up, these results provide doctors with ways to improve the diagnosis and treatment of their patients without needing to reverse their fundamental clinical practice, and deliver great value to researchers hoping to better identify these disorders for investigation.

Summary Statement
I significantly improved the diagnostic accuracy of sleeping disorder hypersomnias by incorporating existing clinical information at novel thresholds, thereby elucidating disease patterns not readily apparent in common clinical practice.

Help Received
Dr. Logan Schneider of Stanford University provided access to data and other laboratory resources and guidance on my model formulation and analyses interpretations. Dr. Emmanuel Mignot of Stanford University provided me with the facilities to conduct my research.

Name(s)
Cameron C. Jones

Project Number
S0813

Project Title
Automated Identification of Organic Molecular Structure and Relative Concentrations from Infrared Spectral Data

Abstract

Objectives/Goals
Most of what we know about the composition of the universe is due to spectroscopy, the measurement of light intensity at varying wavelengths. My project investigates the application of machine learning techniques (convolutional neural networks) to the problem of identifying complex polycyclic aromatic hydrocarbon (PAH) molecules in IR spectroscopic data. I automate the nonlinear mapping of IR spectra to identify chemical composition and relative concentrations of unknown mixtures of molecular compounds found in telescopic data. Such techniques could greatly accelerate the analysis of the data captured from astronomical instruments (such as the upcoming James Webb Space Telescope).

Methods/Materials
My project looked at three specific problems: a) building models to identify PAH molecules from empirical IR spectroscopic data when trained on the approximate theoretical counterpart from NASA's PAHdb v2 database, b) building models to perform the same task but with NASA's 5x larger PAHdb v3 database, and c), building models to identify random compositions of theoretical PAH molecules (up to ten) from composite spectroscopic data using the v3 database. I implemented my models with the open source library TensorFlow.

Results
My principal findings are: a) convolutional models can be trained on theoretical spectra to accurately identify empirical PAH molecules (with 73% accuracy), and b) when trained on data for 3,139 PAH molecules, my models can identify the molecular concentrations of random compositions with weight vector correlations of ~85%, and correctly identify the largest constituent ~67% of the time. In all cases, the models dramatically outperform standard linear (logistic) models. These network models could aid scientists in identifying astronomical objects for further study and research, thus greatly amplifying research efficiency.

Conclusions/Discussion
My project proves the utility of convolutional neural networks to analyze telescopic IR spectra. My best model (ResNet5 with ~200M parameters) exceeded the performance of linear (logistic regression) models in all tests. The versatility of these models illustrates their unique ability to recognize patterns in the features of the spectra and to generalize across diverse datasets. The models can significantly increase the efficiency of analyzing empirical IR spectra and understanding the composition of gas clouds, stars, and exoplanets in our universe.

Summary Statement
I demonstrated the effectiveness of convolutional neural networks in identifying the composition of polycyclic aromatic hydrocarbon mixtures from infrared spectra.

Help Received
I created the models independently while consulting with Dr. Partha Bera, a postdoctoral researcher at NASA Ames, who helped me understand the concepts of spectroscopy.

Name(s)
Saeyeon Ju

Project Number
S0814

Project Title
New Visualization and Analysis Approaches Using 3D Electron Microscopy and 3D Printing Technologies

Abstract

Objectives/Goals
Serial block face scanning electron microscopy (SBEM) is a highly advanced technology to create 3D EM image stacks from 2D EM. Challenges of 3D EM have now shifted from how to capture the difficult-to-measure to what to do with all this big data. While the ability to acquire big 3D EM data is progressing rapidly, more advanced analysis tools and visualization methods are needed to assist in measuring precise 3D morphologies of micro-organelles.

Methods/Materials
In this project, manual, semi-automated, and auto-segmented 3D reconstruction methods were tested in the analysis of variable contrast SBEM datasets. Semi-automatic segmentation was performed with the "Interpolator plugin" from IMOD. Automatic segmentation was performed using modified tools in IMOD and ImageJ. Lastly, 3D printing is performed with programs Autodesk Meshmixer and Ultimaker Cura.

Results
IMOD and ImageJ combined automatic segmentation remarkably reduced annotation time and addressed the alteration and degeneration of the axon in the large brain cancer 3D EM dataset. Revealing the structure of the retinal neuron microcircuit can be accelerated using the alternative semi-automatic segmentation tool. Due to the limitations of printer technology, the delicate morphologies of retinal neurons pose the main technological challenge to constructing these 3D printouts. Reconstructed 3D retinal neuron model files were exported to OBJ or STL 3D model files and successfully produced the retinal microcircuit 3D printing model.

Conclusions/Discussion
Automatic segmentation is a strong tool to diagnose brain cancer illnesses in 3D EM datasets, which are relatively high contrast datasets. This technique should be extended to segment myelinated axon boundaries in brain images, where the axons do not follow the same direction and the staining is not limited to myelin sheaths. The IMOD Interpolator segmentation in 3D EM data is another advanced tool to accelerate the reconstruction of low contrast datasets. Types of retinal neurons and their synaptic interaction are able to be addressed in less time. Thus, these tools fill a critical need by allowing for the quantitative analysis of volumetric EM datasets at the nanoscale. The combination of the optimized annotation technique with 3D EM datasets and 3D printing can obtain high-resolution morphological data for microcircuits.

Summary Statement
Three segmentation tools were tested in SBEM datasets to reveal the advantages and limitations of software programs, and 3D EM and 3D printing visualization were combined to display these synapse-level models and their connectivity.

Help Received
Mrs. Gillum, Drs. Guy Perkins, Keunyoung Kim and Scott Mcavoy supervised me and helped me to analyze the datasets. UCSD provided equipments.

Name(s)
Elleen Kim

Project Number
S0815

Project Title
Using Machine Learning Algorithms to Investigate Genetic Burden of Scoliosis from Candidate Genes & GWAS Studies

Abstract

Objectives/Goals
The engineering goal was to create programs that cluster characteristics of Scoliosis candidate genes and prioritize pairs of Scoliosis patient DNA sequences to investigate the causes of Scoliosis.

Methods/Materials
A database of candidate genes was compiled via literature review of published papers. Candidate genes were grouped by running their numerical characteristics through KNN and K-Means clustering algorithms. Scoliosis patient DNA sequences from public databases were run through customized local and global alignment programs and through BLAST.

Results
Results from the KNN program provided an upper bound and lower bound threshold to quantify the validation test scores of clustering, allowing patient data to be quantified by quality according to this threshold. KNN and K-Means are meaningful analyses of gene groups that can elucidate underlying similarities of Scoliosis candidate genes while producing consistent results within the margin of error. Results from alignment provided a prioritized list of Scoliosis DNA sequences.

Conclusions/Discussion
In conclusion, this project is a proof of concept experiment that shows that machine learning algorithms can be applied to genetic data. Prioritized DNA sequences can be used to further study the association to Scoliosis. With the falling cost and rising prevalence of GWAS studies, this prioritization can be used to quantify the influx of GWAS data for various diseases. Alignment programs also provided meaningful differences in best alignment scores between patient DNA sequences to use as a starting point for further research on what possible genetic variations or phenotypic characteristics determine best alignment. The machine learning clustering algorithms produced a range in validation test measures to quantify genetic datasets. Future identified candidate can be assessed using this metric. In line with our experimental goals, identified clusters or groups of candidate genes also serve as a foothold for further investigation on potential relationships or patterns between the genes and Scoliosis.

Summary Statement
This project customizes machine learning clustering algorithms and DNA alignment algorithms to study Scoliosis candidate genes and patient DNA, and creates a metric to quantify the quality of GWAS datasets in other disease applications.

Help Received
I designed and built the programs by myself. I got help in understanding the algorithms from Dr. Neil Sarkar (Department of Biomedical Informatics, Brown University) and Sophie Kim (Department of Humanities and Sciences, Stanford University).

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Attila B. Koksal

Project Number
S0816

Project Title
Accuracy of Integration Schemes in Planet Trajectory Simulations

Abstract

Objectives/Goals
The objective of this study is to find the most accurate integration scheme using planet trajectory simulations.

Methods/Materials
Computer, Python Programming Language (VPython Library for Visualization and NumPy for Vector Operations), NASA JPL Ephemeris Interface.

Picked a start time, Got position, velocity, mass for the planets from NASA, Chose a timestep, duration, and a integration scheme, Ran the simulation, Found the simulated planet positions, Got actual planet positions from NASA, Compared the simulated and actual planet positions to get the errors.

Results
The Runge-Kutta and Verlet methods were more accurate than the Euler methods. These results tied to my research online and in book sources that the Runge-Kutta and Verlet integration schemes are far more accurate than the Euler integration schemes.

Conclusions/Discussion
Throughout many different timesteps and durations, it's evident that the Runge-Kutta integration methods and the Verlet integration methods were more accurate than the Euler integration methods. It's concluded that the Runge-Kutta and Verlet methods are far more accurate than the Euler methods.

Summary Statement
I created a computer program written in Python to test the accuracy of different integration schemes in planet trajectory simulations.

Help Received
My big brother, a computer scientist helped me with writing and teaching the visualization part of my code.

Name(s)
Patrick Liu

Project Number
S0817

Project Title
iCordisX: SmartPhone-Based Personalized Cardiac Monitoring Using Computer Vision and Bluetooth Low Energy

Abstract

Objectives/Goals
iCordisX aims to provide a personalized and data-driven supplement for cardiac anomalies that acts as a dependable healthcare interface for wireless ECG monitors.
Targeted Features: User-based system, daily symptom tracker, BLE functionality, diagnosis/monitoring mode.

Methods/Materials
Drew out and implemented project system: sensors + hardware. Developed algorithm for "diagnosis" of ECG. - Machine learning for extraction of baseline features from MIT-BIH Database, and object detection. Designed app interface: screens, user input, data flow and models. Tested fully-functional app system, review data output, compare accuracy to MATLAB algorithm, receive feedback from cardiologists/entrepreneurs. Create hardware casing (acryllic) and 3D shell.
ECG Circuit - 1 x Arduino Pro Mini and Cable - 1 x AD8232 board 3 x TENS electrodes
Computer + Software: Arduino IDE, Processing 3 Software, Anaconda-Navigator (Jupiter-Notebook), MATLAB R2016B, node.js, Visual Studio Code, XCode, AWS EC2 Instance

Results
Smoothing/peak detection method in MATLAB resulted in detrended signal, color-coded blue and red to distinguish the original signal.The Python algorithm was able to successfully filter, calculate specific intervals, and calculate heart rate, as seen by its percent error of only 8.08% for averaged features when compared to the MATLAB analysis. Average Signal Quality should be at least approximately .94. iCordisX has a calculated net price of $131 dollars (excluding the mobile device), an impressive feat for all it's capabilities.

Conclusions/Discussion
CordisX provides a unique value proposition: a personalized, simplistic monitoring system that is appealing to the aging society. The feature extraction algorithm is comparable to the accuracy of a standalone MATLAB program, which also verifies the device's accuracy of data output. The app is flexibility with data management and real-time data streaming, whether it be via monitoring or diagnosis. Users will receive a data-driven supplement for their daily heart health, all while logging it in the database for their personal physicians to see. Survival rates from heart attack may be increased from early detection in irregular heart rhythm, where the emergency protocol may be activated. All of iCordisX's features are driven by the user's information, and allows the user to view trends over a large period of time.

Summary Statement
iCordisX aims to provide a personalized and data-driven supplement for cardiac anomalies that acts as a dependable healthcare interface for wireless ECG monitors.

Help Received
Received resources at beginning of project from Nitish Nag (PhD Student @UCI) to begin algorithm engineering.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Marcus X. Luebke

Project Number
S0818

Project Title
Running on Water: Developing Novel AI/Optimization Techniques to Accelerate Research on Real-time Hydrogen Production

Abstract

Objectives/Goals
This project is a continuation of a four-year effort to generate hydrogen in real time to power automobiles, including last year's addition of a computer model of my physical system and an artificial intelligence (AI) to optimize the design based on user input priorities. This year, my objective was to create a faster and more accurate program that returns designs that better meet the user's priorities, accounts for more possibilities and how the design will be used, and converges faster to an optimized solution.

My objective was to improve my hydrogen production model & AI for greater application, accuracy, efficiency and speed, to find the optimum solution based on input priorities.

Methods/Materials
Hydrogen Production simulation: I updated my model to more accurately characterize the electrocatalytic (cathode, anode, solution) properties.
System optimization: I added an operating cost term to the Cost function, to better evaluate the time-based cost of maintenance and operations. The Cost function was also updated to better represent how well the AI is meeting the user's expectations, for more accurate and intuitive assessment by the user.
Novel AI algorithms: I developed new AI techniques and incorporated them into my evolutionary algorithm from last year:
1. "Food" based incentivization, to efficiently search the Design Space by allocating more resources to "organisms" with the most potential in each generation
2. Third order gradient descent line search, to improve speed by taking intelligent next steps

Results
Updating the cost function to include operating cost encouraged more efficient designs which took into account time based factors as well. In addition, the results of the model, when compared to previous data, were more accurate. Finally, my novel AI techniques consistently produced better designs (lower relative Cost) and converged to the best design approximately 6 times faster than a standard evolutionary AI algorithm.

Conclusions/Discussion
In addition to technical specifications, it is critical to consider realistic factors such as the cost over time of maintenance and operations. The successful changes to the AI algorithm demonstrate the importance of having the right algorithm to generate accurate, relevant results quickly. The novel third order line search and food-based population system I developed could be valuable additions to the arsenal of existing AI

Summary Statement
I developed two novel AI techniques and updated my computer model of a hydrogen production system, resulting in faster convergence to better designs.

Help Received
Assistant Prof. Kochenderfer of the Aero-Astro / Computer Science Departments gave me early access to his textbook "Algorithms for Optimization", which provided a comprehensive overview of the most current optimization techniques. These techniques served as a foundation from which I built my own

Name(s)
Alice Martynova

Project Number
S0819

Project Title
Using Generative Adversarial Networks to Enhance an Affordable Micoscope for Epidemic Prevention in Developing Countries

Abstract

Objectives/Goals
The Foldscope is a microscope made of paper and a removable lense which costs 25 cents to make. In this project Foldscope is used to target Schistosomiasis, a parasitic disease second to Malaria in deaths and economic effect in developing countries. Doctors diagnose it by using microscopes to count the number of Schistosomiasis eggs in urine. If the number is greater than 50, a certain medication is given, and otherwise a different medication is administered. However, developing countries lack not only microscopes but also access to medical professionals.

Methods/Materials
This project uses an algorithm called Generative Adversarial Network (GAN), a Foldscope, and a Raspberry Pi with camera to replace the current expensive diagnostics. First, I used miniscule plastic beads and artificial urine to model Schistosomiasis eggs found in urine. I took 350 images of these samples by attaching a Foldscope to a Raspberry Pi camera, and used them to train the GAN. GAN consists of two components: the generator and the classifier. The generator learns to produce fake images resembling the real ones, while the classifier learns to recognize both the fake and the real images and tell them apart.

Results
Four tests were conducted. In the first the classifier placed images into two categories: with and without eggs. 50% of sample images contained objects other than eggs, for urine often has other visible components especially in areas with dirty water. The accuracy in this test was 95%. The next test used three classes: 0 eggs, < 50 eggs, >= 50 eggs. This test mirrors how the medication is prescribed for Schistosomiasis, and it was 94% accurate. Third test estimated actual egg count for each sample, and measured the average deviation from the true count. By the end of training, the accuracy was 0.25 eggs, less than one egg off. The fourth test was the same as the third, but the network had no generator. This test resulted in an accuracy of 11.5 eggs which demonstrates that GAN algorithm is essential.

Conclusions/Discussion
The trained network is downloaded onto Raspberry Pi, and does not require Internet to operate, making it suitable for remote areas. All together the device costs less than $25, compared to $400 for the cheapest microscope, and replaces the need for a trained diagnostics professional. Finally, while this device was trained to combat Schistosomiasis, it can be re-trained on any parasitic disease.

Summary Statement
I used a novel image processing technique to add a computing component to a paper microscope to build a cheap autonomous device for parasite detection in rural areas.

Help Received
I conducted most of my research by myself, and would sporadically email Kevin Carde, a mentor from one of my past summer camps, giving him updates on my project.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)

Erik Mora; Alex Munoz; Jacob Zavala

Project Number

S0820

Project Title

Data Collection and Analysis of Aerial Drone Photography

Abstract

Objectives/Goals

In recent years, drone technology has had large advancements. Now that drones have gone from expensive to less expensive and more available to the public, this technology can help not only the common folk but larger companies as well. In the project here, a drone was used to take aerial photos in order to observe specific terrains. 40 flowers were randomly placed on a strip of land. A drone was coded in order to take photos of the land, which were then examined to see if the flowers that were placed could be identified in the photographs.

Methods/Materials

Place Markers, Notes, Drone, Species Props, DJI Mavic Pro, Calculator, Droneblocks Coding App, Meter Roll, Cell Phone

Results

Once the tests to get the working code were successful, the code was then used with props ,used to represent a species, which then showed the hypothesis was proven correct. Even with the props being rather small and hard for any camera to distinguish from the height of 55ft. The use of the drone allowed to analyze the terrain with the specific species allowed for more in-depth analysis of the surveyed area.

Conclusions/Discussion

The results of the data and project as a whole show the ability to gather data from a new angle. The methods can be applied to new research and even private companies to gather desired information on area.

Summary Statement

Using new technology can we use a code to run a program that can consistently gather data with photos over desired areas.

Help Received

Science Teacher Supplied Drone

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Andrew B. Nazareth

Project Number
S0821

Project Title
Extracting Wildlife in Puma Project Photos Using Java

Abstract

Objectives/Goals

As part of the Summer Internship Program in Environmental Studies, at the University of California Santa Cruz last summer, I manually tagged a number of images which had wildlife in them. As this was time consuming, I was curious if I could automate this process.

My objective is to develop a Java program to extract changes in images from the Puma Project footage to help identify wildlife found in the Santa Cruz Mountains. This program compares the pixels between two successive images in the Puma footage stream to highlight the presence of animals.

Methods/Materials

The Puma Project footage was obtained using motion sensitive cameras monitoring wildlife.

The program to extract the animals (from two successive images) was written in Java, using the Eclipse platform. The user interface was created using JavaFX with Scene Builder 2.0.

Four steps were coded to extract changes in the images.

1. Converting image(s) to greyscale.
2. Creating a new image that was the 'difference' of two input (color/greyscale) images.
3. Creating an adjustable filter to mask out the identical portions of the images
4. Applying the mask to the original image to extract the wildlife.

Results

I was able to extract the animal from the background images using my program.

I ran the program on 20 sets of the images from the Puma Project footage.

The masked image provided the best results. The mask value that provided the best extracted image varied across images.

It was not easy to recognize animals in the 'difference' image. Operating on greyscale versions of the images did not change the results significantly.

Conclusions/Discussion

My program demonstrates that it is possible to automate the tagging process from two successive images. It met my objective for this project, where my program using Java can extract an animal/object from a background image.

Summary Statement

I coded a Java program to extract wildlife from Puma Project footage by comparing pixels from two successive images using the Eclipse platform and the user interface was created using JavaFX with Scene Builder 2.0

Help Received

Mr. Williams, Mr. Askins and my school project advisor, Mr. Johnson helped me get started with the Java programming, user interface design, and provided me with valuable advice. Veronica Yovovich, program mentor for my SIP Internship, gave me permission to use footage from the Puma Project.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Anjo B. Pagdanganan

Project Number
S0822

Project Title
Analyzing the Efficiency of Subsequent Convolutional Layers with Small-Scale Images

Abstract

Objectives/Goals
My project attempts to find the optimal number of convolutional layers (a conv. layer teaches filters to recognize details) to place next to each other in order to improve the training efficiency of a neural network.

Methods/Materials
Using Python, four convolutional neural networks were trained on the CIFAR-10 dataset. Each model n had n conv. layers placed subsequently (otherwise, their architectures were the same). Each model was trained 5 times, running 100 loops over the training data, then assessed on its accuracy. The libraries used in this project were Keras (with TensorFlow as its backend), SciPy, Pandas, and Matplotlib.

Results
There was no significant improvement between the model that used blocks of three subsequent convolutional layers and blocks of four conv. layers.

Conclusions/Discussion
Neural networks using blocks of three convolutional layers trained the most efficiently. These results could have applications in feature detection with low resolution images.

Summary Statement
I found the optimal number of convolutional layers (filters in a neural network that can be trained to detect features like edges) to place subsequently in order to improve training efficiency.

Help Received
None. I designed and conducted the experiment myself.

Name(s)
Junseo Park

Project Number
S0823

Project Title
Diabetic Retinopathy Symptoms Recognition Using Image Processing

Abstract

Objectives/Goals
Diabetic Retinopathy (DR) affects 347 million people in the world, of whom 10% will lose their sight. The goal was to develop a tool to be used in diagnosing DR.

Methods/Materials
The research idea was obtained from Kaggle.com. The images were obtained from ADCIS.net. The image's brightness was curve fitted to a quadratic surface in order to normalize the brightness across the field. Then the color components were used to segment the blood vessels, optic nerve disc and other features that were neither healthy tissue nor blood vessels, i.e., anomalies. Morphological components were used to determine the shape and the size of blood vessels and anomalies. Computing and then measuring the distribution of these morphological measurements, the presence and the severity of the retinopathy was determined.

Results
Hemorrhages and hard exudates were detected successfully on images that were given.

Conclusions/Discussion
The results are very promising because these correct detections of the symptoms will lead directly into correct diagnosis of DR.

Summary Statement
I automated the detection of hard exudates and hemorrhages on fundus images using image processing.

Help Received
Dr. James Choi taught me image processing.

Name(s)
Nitya Parthasarathy

Project Number
S0824

Project Title
BiasCheck: An Artificial Intelligence Based Tool to Evaluate Bias in Social Media

Abstract

Objectives/Goals
The objective of this study is to establish a scientific basis using Artificial Intelligence (AI) techniques to study various forms of bias and stereotypes in social media.

Methods/Materials
Computer for coding and running AI programs. Public domain datasets were obtained from the Internet.

Results
New statistical metrics, some of which were adapted from diverse areas such as Information Theory and Language Modelling were introduced to evaluate gender bias in social media. These models were then further substantiated with novel algorithms using AI for bias prediction. Using Bayesian models as well as numerous sophisticated Neural Networks, the effectiveness of AI algorithms in studying biased text is then demonstrated on large social datasets. Incorporating these ideas, a web-based BiasCheck software is developed to automatically assess a BiasScore for any blog, webpage or document. Though particular emphasis is placed on gender bias evaluation, results are shown to readily extend to other types of bias evaluation.

Conclusions/Discussion
Comprehensive results were provided to demonstrate the presence of male and female gender stereotypes in social media. Furthermore, statistical techniques identified positive social sentiment for gender associated with specific behavior. Female gender was generally identified with softer roles while male gender was identified with leadership roles. AI algorithms (using numerous classifiers) developed were able to pick up this bias and aptly identify the gender in a sentence from surrounding words. In particular, female stereotypes was picked up more accurately indicating the presence of more overt bias for the female gender. Further, AI models also yielded interesting insights into social behavioral perceptions whereby a providing man was identified as successful whereas a providing woman was tagged as delicate!

Existence of bias in movie review datasets as a function of movie genre was also evaluated and shown to be more prevalent in specific categories. A valuable social commentary is also provided by studying the evolution of bias over time. In summary, a new direction of applying statistical techniques and AI for social good has been established in this work uncovering a rich set of topics for future study.

Summary Statement
A comprehensive scientific basis is developed for evaluating bias in social media using novel statistical techniques and Artificial Intelligence algorithms.

Help Received
I developed and coded the AI algorithms myself. I discussed results with Prof. Sameer Singh in the department of computer science at UCI.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)

Deepro F. Pasha

Project Number

S0825

Project Title

Intelli-Drip: A Sensor Based Autonomous Feedback Control System for Commercial Irrigation

Abstract

Objectives/Goals

The project objective was to automate water management in commercial irrigation using feedback control systems to save water. The automation was dependent on moisture availability in the soil for plants. Weather related parameters were used to validate the moisture levels. A combination of the appropriate hardware design, analysis and programming was used to accomplish the project goal.

Methods/Materials

Raspberry Pi 3 Version B, Ribbon Cable for GPIO, 40 pin breakout board , Breadboard, Relay Module , HDMI Monitor, Mini Pump, Small Reservoir, Cauliflower Plants, Drip Irrigation System, Jumper Wires, Soil Moisture Sensors and software: Python 3, Raspbian were used in this project. Ten cauliflower plants as experimental group and ten other cauliflower plants of same age as control group were planted in pots for the experiment. The control plants had a manually operated drip irrigation system. The experimental plants had the newly designed autonomous feedback control system for irrigation. Soil moisture sensors were connected to the experimental and control plants and data was collected and used for determining level of moisture availability in the soil for plants. Using the dry and wet moisture pulses each sensor was calibrated to calculate soil moisture level in percent. Weather data were collected and used to validate the soil moisture readings from sensors. An algorithm was developed on the Raspberry Pi using Python programming language to analyze the collected soil moisture and weather data to find out optimum time and amount of water to irrigate. Based on the analysis, a signal was sent to control the pump automatically to water the experimental plants.

Results

The daily water savings per plant (20.4%) found from this experiment can be extended to estimate the water savings for commercial irrigation to an acre of cauliflower crop field. Considering average spacing of 18 inches, and 19,360 cauliflower plants, the total water savings per day is approximately 1,597 liters/day/acre and for a 70 days cauliflower season, the total water savings can be 111,790 liters/acre/season or 29,535 gallons/acre/season.

Conclusions/Discussion

A control system based on a closed loop feedback system using moisture pulse was designed and operated in this project for irrigation. The control system triggers the pump on and off automatically based on the criteria set in the developed software.

Summary Statement

In this project, a sensor based autonomous feedback control system was designed and operated to control irrigation and save water using combination of appropriate hardware design, analysis and programming.

Help Received

I designed, analyzed and created the system myself but my science teacher helped me to understand how to set up the experiment.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Albert Qin; Samyak Surti

Project Number
S0826

Project Title
A Machine Learning Based Approach to Decrease the Lung Cancer Malignancy Detection Threshold

Abstract

Objectives/Goals
When creating our Machine Learning model, our objective was to be able to correctly identify the malignancy of the Lung presented based on the CT scan images. This was achieved through three main criteria. These included maximizing the classification accuracy of the model, minimizing the model's training time, and minimizing the model's overall run-time. We also wanted to measure the accuracy of our model in comparison to Google's Inception model. Our hope is that our model can be utilized in doctor's offices to forego the need to do biopsies or any other stressful surgeries.

Methods/Materials
Equipment: PC equipped with Nvidia GeForce GTX 980 GPU, Laptop with Nvidia GeForce GTX 1050. Software Components: Tensorflow - Google's Machine Learning and Artificial Intelligence Library for Python, CT scan image database acquired from a cancer imaging database, Google's Inception Deep Learning model

Results
After testing our model, we achieved around a 96% accuracy when classifying the cancer the patient had as being Benign, Malignant, or Metastatic, based on the testing data of CT scan images. The data that we used was sectioned off into training and testing sets to insure that the images in the testing sets haven't been seen by the model during training. This would result in an unbiased accuracy output. When compared to Google's Inception Deep Learning model, we consistently achieved around a 20 to 30% higher accuracy. These results were especially surprising, as we were able to achieve very high accuracy from a simplistic model in contrast to Inception which is extremely complex.

Conclusions/Discussion
Based on the accuracy we achieved with our model, we want to refine it so that doctors, especially radiologists and pulmonologists, can utilize this software to make an accurate diagnosis of the patient's condition without having to perform any biopsies. This will get rid of any stress or anxiety that is often attached to such procedures. By getting an accurate diagnosis of malignancy of the cancer early on, respective action can be taken without any delay, giving the patient peace of mind.

Summary Statement
By creating a simple Convolutional Neural Network, we were able to achieve a surprisingly high accuracy when diagnosing the malignancy of the patient's Lung Cancer.

Help Received
I received help from my dad in understanding some of the fundamentals of Machine Learning.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Tejas N. Rao

Project Number
S0827

Project Title
Logistic Regression and Decision Tree ML Algorithms to Predict Type-2 Diabetes

Abstract

Objectives/Goals
Compare two Statistical Models to predict Type-2 Diabetes - Logistic Regression and Decision Trees.
Determine which patient attributes - Age, Body Mass Index, Glucose Concentration, Genetics, % of time pregnant are most significant for Diabetes

Determine the following for each model to aid comparison: Accuracy, Sensitivity, Specivity, ROC Area under Curve.

Build a simple web application to use the model in mobile phones. Application should accept key patient data and return probability of diabetes
Application should run on phone and browser.

Methods/Materials
UC Irvine Department of Machine Learning Pima Indians Diabetes DataSet. This dataset provides details on 782 Pima Indians for Age, BMI, Pregnancy etc. Scikit-learn: Machine learning in Python Logistic Regression and Decision Tree algorithm packages in Python. Pythonanywhere for Hosting and running Python Applications. Jupyter notebooks running on Azure Cloud.

Methods
Scikit-learn Machine Learning toolkit in Python was used for running Classification Models DataSet has 768 patient records which were divided into 75% (576 records) for Training data and remaining 25% (192 records) for Test data.
Both models Logistic Regression and Decision Trees, are Trained and Scored with training data and test data respectively

Prediction Accuracy is measured as (TP+TN) / (TP+TN+FP+FN)
Sensitivity is measured as TP / (TP+TN)
Specificity is measured as TN / (TN+FP)
HTML5 was used to build a simple webapp that accepts Patient Data in a Form and calls backend Python App.

Results
Logistic Regression Model has

Summary Statement
Prevent Diabetes using Machine Learning Algorithms- Logistic Regression and Decision Trees

Help Received
Mr Wilke (San Mateo High School), Ms Bharathi Udupi (Oracle)

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Shivam Singhal

Project Number
S0828

Project Title
iDetect: A Machine Learning Algorithm for Non-Invasive Cancer Diagnosis through Epigenetic Biomarker Identification

Abstract

Objectives/Goals
Cancer remains a leading cause of death in today's world, only treatable if caught early. Epigenetic alterations are newly-discovered biomarkers that can facilitate early diagnosis. They are particularly attractive as a diagnostic tool due to their stability, frequency, and non-invasive accessibility in bodily fluids, such as blood plasma, in the form of cell-free DNA (cfDNA). The objectives of this project were to locate epigenetic alterations in cfDNA sequences obtained from the blood plasma of cancer patients and to map these sites to CpG islands (CpGI) to identify biomarkers for accurate and non-invasive cancer diagnosis using a machine learning algorithm.

Methods/Materials
This project was conducted in four steps: 1. Determining methylation in cell-free DNA sequences relative to the reference genome using the Bismark tool in Perl script. 2. Mapping the methylated and unmethylated sites to the CpG islands, which are important genomic regions for cancer detection. 3. Selecting features that indicate cancer presence and training the machine learning algorithm with tissue data obtained from the TCGA database to look for these specific features when making predictions for different types of cancer. 4. Testing the accuracy of the algorithm through cfDNA samples obtained from another researcher.

Results
The alignment of the cell-free DNA sequences in Bismark showed that different methylation levels are present, which enabled successful CpGI mapping. A positive correlation between differential methylation levels of cell-free DNA and tissue DNA samples allowed for the use of tissue data in the machine learning process, as well as validating the effectiveness of alterations in cell-free DNA as biomarkers for cancer. The algorithm was able to differentiate between cancerous DNA and the non-cancerous (control) DNA with an 85 percent sensitivity and 67 percent specificity.

Conclusions/Discussion
The program was able to accurately differentiate between different types of cancers for which cfDNA data was available. This algorithm can also be applicable to a host of other conditions in which identifiable differences in methylation have been reported, including neuropsychiatric disorders and cardiovascular diseases. In addition, this non-invasive technique can be extended to more advanced prenatal tests for unborn fetuses.

Summary Statement
This project developed a machine learning algorithm which considers methylation patterns in cell-free DNA to predict cancer presence accurately, efficiently, and non-invasively.

Help Received
I would like to thank Dr. Jiang for providing the cell-free DNA data. I would also like to acknowledge my computer science teacher, Mr. Steinke, for providing me with constant encouragement and my parents for their support throughout the project.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)

Nikhil Sundrani; Sameer Sundrani

Project Number

S0829

Project Title

SmartRate: A Machine Learning Approach to Predicting Cardiac Arrest

Abstract

Objectives/Goals

The goal of this project is to prevent people afflicted by cardiac arrest from dying, individuals whose lives are significantly shortened every year simply because they did not arrive at the hospital in time. This is an issue that has been plaguing society since the advent of modern medicine, and yet is has not yet been solved- until now, with an inexpensive but functional wearable. The prototype features three distinct and effectual algorithms: one for the wristband itself- to calculate heart rate accurately regardless of movement, one for the iOS application- to recognize cardiac arrhythmias and autonomously notify EMS, and one for the remote server- developed with machine learning to predict cardiac events before they occur.

Methods/Materials

To develop the prevention apparatus, an Adafruit based Atmel microcontroller was utilized, coded with the C/C++ compiler in Arduino. The sensor features a photoplethysmographic module, converting reflected light into signal values to measure heart rate. The iOS application was written in Xcode using the Swift language, and the server was coded in Python and hosted remotely via Django. It was tested with a dynamic and local database containing 300 electrocardiogram samples from MIT, deployed on the server.

Results

This project has proved incredibly accurate- the wristband has 98.5% accuracy compared to a 12-lead electrocardiogram, the application recognizes the abnormal heart rate and notifies EMS 100% of the time, and the innovative machine learning predictive algorithm has an accuracy of 90-95%.

Conclusions/Discussion

The cumulative device, which is a combination of three distinct and effective algorithms integrated within an inexpensive wearable that communicates via Bluetooth with an iPhone, creates the possibility of reducing excess funding within hospitals and diminishing diagnosis time for arrhythmias from days to seconds, thereby drastically decreasing Emergency Medical Service time from hospital to victim.

Summary Statement

We developed a wristband, an iOS application, and a remote server- utilizing machine learning to save the life of the user by predicting cardiac abnormalities and therefore preempting cardiac arrest.

Help Received

In addition to utilizing the MIT arrhythmia database, Dr. Nan Wang at CSU Fresno provided guidance in understanding machine learning classification functions.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Gideon Tong; Jane Zhang

Project Number
S0830

Project Title
Meloread

Abstract

Objectives/Goals

In the US, 32 million adults are illiterate and although new educational techniques are being developed, the literacy rate is not increasing by much. According to Time Magazine, kids read less and less as they get older, with "45% of 17-year-olds saying they only read twice a year". These percentages have tripled by 2014 which is also reflected by poorer scoring on reading comprehension tests as reported by research from Common Sense Media.

Thus, the literacy rates in the US are stagnant, leaving a large group of people struggling in today's faster pace society. Evidently, those who fail to become literate have difficulty finding jobs, maintaining relationships and simply navigating the streets.

How can we combat the declining literacy and better educate the people in the US? Make reading more entertaining. Nowadays, teenagers are drawn to watch the latest tv shows for hours due to the immersive experience complete with colorful graphics and dramatic music instead of settling down for a plain book or article. Thus, we created Meloread to provide students with an immersive reading experience by adding music that matches the tone of their reading material, motivating more students to read and ultimately increasing literacy rates.

Methods/Materials

2 laptop computers with text editor as well as JavaScript testing suite. Was tested by submitting the app to the Chrome App Store and having 500 users download it and fill out a survey

App is written in HTML, CSS, JavaScript and JQuery and connects to the IBM Watson API in order to use neural networks to determine the tone of the active text, then plays music from YouTube with the relevant mood

Results

Over 400 of the 500 users ages 13-19 surveyed reported that they would be happy to use Meloread again in the future as a reading aid in order to make the reading experience more enjoyable and interesting.

Conclusions/Discussion

In the future, Meloread could make use of iteration in order to improve upon itself and take into account user feedback, including consistent crashing of the Chrome extension as well as the inability to support background playback. In this sample it was determined that it is indeed more enjoyable to use Meloread

Summary Statement

Meloread is a computer software that matches music to the tone of one's reading material to provide an enhanced reading experience.

Help Received

We designed and programmed the app ourselves using internet searches on API usage and received suggestions from mentors at a hackathon.