CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) Sandhini Agarwal	**Project Number** **S0801**

Project Title

Tracking Parkinsonian Tremors: A Wearable Device Application Utilizing Smart Sensing for Real-Time Monitoring & Analysis

Abstract

Objectives

Parkinson's Disease is a neurological disorder caused by neuron deterioration in the substantia nigra part of the brain. Such tremor-based diseases impacts tens of millions of people around the world and patients suffer from shaking, stiffness, and problems with balance and coordination. Unfortunately, there are no standard quantitative, efficient and reliable ways for patients and doctors to track tremor progression over time. The goal is to develop a wearable device application that focuses on assisting with diagnosis and monitoring of hand tremors. The app enables real-time quantitative measurement, analysis, data storage, and communication of results to healthcare provider. This allows the doctors to assess the progression of patients' symptoms and assists them in varying the treatment regimen for enhanced patient outcome.

Methods

The app utilizes the smartphone s built-in accelerometer sensor to monitor any movement of the device and is coded on the MIT App Inventor 2 platform. The Google Fusion Table is used to store the recorded data. Android Debug Bridge was used to transfer the app onto the wearable device (Android Tic Watch). All data was recorded through the smartphone, but can also be visually tracked using the wearable device.

Results

According to the data, the range of the sensor s coordinates for normal behavior is the deviation from the recorded original position: ± 0.2 m/sec2 for x-coordinate; ± 0.4 m/sec2 for y-coordinate; ± 0.2 m/sec2 for z-coordinate. Any value recorded outside the above mentioned range is considered abnormal. A normal hand behavior shows a steady graph line, whereas when the tremors are detected, the plotted graph shows peaks and valleys. The extent of peaks and valleys shows the severity of tremors. The tremor score on a standardized scale of 1 to 10 is being used to indicate the severity of tremors in numerical terms.

Conclusions

The developed mobile app for a wearable device can successfully record and plot the presence and severity of hand tremors associated with tremor-based diseases such as Parkinson's Disease. The plotted graph clearly shows the difference between normal and abnormal hand tremors to assist in the diagnosis and monitoring of Parkinson's Disease. The numerical tremor score and the graphical data can help the healthcare provider in keeping a historical record of the patient's condition before, during and after the treatment. The accurate and precise measurements of symptoms can allow doctors to assess disease progression, deliver optimal drug dosage, thus resulting in better prognosis.

Summary Statement

I developed a wearable device application that allows for real-time quantitative measurement, analysis, data storage, and communication of results to assist with the diagnosis and monitoring of Parkinsonian tremors.

Help Received

I designed and developed the app on my own. I trained myself on android programming using online tutorials.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)

Rehaan Ahmad; Brian Yang

Project Number

S0802

Project Title

Constructing Vegetation-Health-Map Image Forecasts Using a Novel Variable Length Attention ConvLSTM Network

Abstract

Objectives

Vegetation health forecasting is an important field of research that involves predicting annual crop growth patterns in order to provide information to humanitarian aid agencies, government crop monitors, and famine warning systems. We propose and develop a novel algorithm to construct accurate 32-day-ahead image forecasts of vegetation health, as measured by NDVI (Normalized Difference Vegetation Index), over a 100 km by 100 km crop region in Ethiopia.

Methods

The vegetation health data was obtained from NASAs MODIS satellites. To improve our models ability to learn the annual patterns of crop growth, we developed a modification to the standard method of spatiotemporal forecasting, the ConvLSTM. Our proposed model, the Annual Gate ConvLSTM, introduces attention-network-weighted data from previous years into the memory cell of each recurrent unit. Furthermore, our variable length attention network improves upon the standard attention network by ensuring a constant standard deviation of weights for sequences of different lengths. This is achieved by multiplying all unnormalized terms by a value k, determined by a regula falsi approximation, before feeding the terms into the softmax normalization function. We measure the accuracy of the standard ConvLSTM and the Annual Gate ConvLSTM using Root Mean Square Error (RMSE, 0 is perfect, 1.3 is worst) on a common testing dataset of 70 sequences of 10,000 sq km NDVI data.

Results

Our Annual Gate ConvLSTM achieves RMSEs of 0.0852, 0.0834, 0.0882, and 0.0911 for the 8, 16, 24, and 32 day ahead predictions of NDVI values over the 10,000 square km region. The standard ConvLSTM achieves higher, less accurate RMSEs of 0.1005, 0.1135, 0.1218, and 0.1412 for the four respective timesteps. Furthermore, our network outperforms baseline models and existing methods of NDVI forecasting.

Conclusions

The superior performance achieved by the Annual Gate ConvLSTM in comparison to the standard ConvLSTM suggests that the usage of a variable length attention network in a Recurrent Neural Network helps to discover and utilize periodic trends in sequential datasets. Furthermore, the best existing algorithm for large scale NDVI forecasts only provides 8 day, 10 km by 10 km forecasts of 1 square km resolution at an RMSE of 0.09 - performing fewer steps ahead, smaller scale, and less accurate forecasts.

Summary Statement

We develop a novel, period-aware modification to the ConvLSTM to construct accurate, long-term, large-scale image forecasts of vegetation health captured by NASAs MODIS satellites over a 10,000 square km region in Ethiopia.

Help Received

We designed, developed, and tested our algorithm entirely on our own. However, we obtained our data from the repository of NASAs MODIS satellite imagery. Furthermore, we would like to thank Dr. Wassila Thiaw & Dr. Pablo Bocca for their advice pertaining to the importance of NDVI forecasting.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) **Ayush Alag**	**Project Number** **S0803**

Project Title

Computational DNA Methylation Analysis of Food Allergy Yields Novel 13-gene Signature to Diagnose Clinical Reactivity

Abstract

Objectives

Current blood and skin tests are inaccurate (50-60% false positive rate) in distinguishing true food allergies (FA) from oral sensitivities. As a result, life-threatening Oral Food Challenges (OFC) are used, which has resulted in patient mortality and over-diagnosis of FA. I sought to create a highly-accurate diagnostic classifier of FA from blood sample (safer than an OFC) epigenomic data. I also sought to use a purely data-driven methodology, which would be rendered applicable to other diseases. Lastly, I sought to find biological associations with FA.

Methods

Working by myself on a dataset publicly available on Gene Expression Omnibus, I coded in Java (with Weka ML library) to develop a computational framework for feature selection and classification. My methodology was based on Sequential Forward Selection and ensemble classification methods. I later used the Illumina BaseSpace Correlation Engine to find gene and pathway associations for the diseases I found. I also used Gene Ontology Enrichment Analysis, Princeton University s Generic Gene Ontology Term Mapper, REVIGO, and NAVIGO, which are all publicly available, to find representational biological terms associated with the genes I found.

Results

An unbiased feature-selection pipeline was created that narrowed down 405,000+ potential CpG biomarkers to 18. Machine-learning models that utilized subsets of this 18-feature aggregate achieved perfect classification accuracy on completely hidden test cohorts. Ensemble classification was also shown to be effective for this High Dimension Low Sample Size (HDLSS) DNA methylation dataset. The 18-CpG signature mapped to 13 genes, on which biological insights were collected. Notably, many of the FA-discriminating genes found in this study were strongly associated with the immune system, and seven of the 13 genes were previously associated with FA.

Conclusions

I implemented an efficient feature-selection algorithm that found a condensed list of strong CpG biomarkers. I replicated the perfect classification found in previous works but with a much smaller CpG set (by Occam Learning, simpler models are preferable), and also with unbiased k-fold cross-validation accuracy measurement. Furthermore, the methodology I used was completely data-driven and generalizable to other diseases. I also found novel genes associated with FA. I am the sole author of this paper s publication in PLOS One, and it is currently in the minor edits stage.

Summary Statement

I created purely data-driven and highly-accurate machine learning models to perfectly classify true food allergies (as opposed to milder sensitivities), and in this process I found genes and biological pathways associated with the disease.

Help Received

Working independently at home, I consulted Dr. Joseph Hernandez from Stanford University for feedback on my paper, and he also advised me on collecting biological insights.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) Samuel Alber	**Project Number** S0804

Project Title

Developing and Applying a Novel Stochastic Model to Increase Production of Functional Cardiomyocytes via DNA Methylation

Abstract

Objectives

Heart disease is currently the leading cause of death throughout the world; however, an upcoming frontier in regenerative involves engineering patient-specific functional cardiomyocytes from induced pluripotent stem cells to treat many forms of heart disease. The goal of this project is to build a novel computational framework capable of describing large gene regulatory networks such as the developmental gene regulatory network within cardiomyocyte. This model is then used to provide novel insight into how DNA methylation perturbations could be used to increase the probability of obtaining a functional contractile cardioymyocyte which will make modern heart tissue engineering approaches more efficient. My hypothesis is that inhibiting global DNA methylation levels will increase the probability of producing a functional cardiomyocyte.

Methods

Using a MATLAB script, I converted the gene interactions within a known 29-gene cardiomyocyte gene regulatory network into a set of biochemical rules that were then stochastically simulated via BioNetGen. Additionally, I included a parameter responsible for the level of DNA methylation within my network model. I then incorporated the weighted-ensemble method to increase the efficiency of my model through using the WESTPA package in conjunction with BioNetGen. Finally, I ran my model simulations on the High Processing Computing Cluster at the University of California, Irvine, with varying levels of DNA methylation and visualized my large 29-dimensional data set using the t-SNE algorithm.

Results

The stochastic model coupled with the weighted ensemble method was able to successfully describe the cardiomyocyte gene regulatory network. The level of DNA methylation determined which phenotypes were expressed and at what probabilities. Inhibiting the global level of DNA methylation was shown to cause the network to favor the contractile cardiomyocyte phenotype.

Conclusions

My model is the first biologically relevant model to demonstrate that global DNA methylation perturbations can be used to increase the production of functional contractile cardiomyocytes, which is crucial in treating many forms of heart disease. Furthermore, my model is the most computationally complex model so far to successfully incorporate the weighted ensemble method. This novel framework can be extended to make other large-scale biochemical models significantly more efficient.

Summary Statement

I constructed a novel stochastic model with the weighted ensemble method that demonstrates how inhibiting global DNA methylation levels could increase the production of functional cardiomyocytes.

Help Received

I met with Professor Read of the University of California, Irvine, weekly to discuss the progress of my project and I met several times with graduate student Cameron Gallivan who helped me with debugging some of my code.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) **Suraj Anand**	**Project Number** **S0805**

Project Title

Machine Learning Ensemble Model for Improved Personalized Lung Cancer Risk Assessment and Malignant Nodule Detection

Abstract

Objectives

Current screening guidelines omit a large number of high-risk candidates that do not fit the traditional screening criteria. Furthermore, malignant lung nodule detection on CT scan is difficult as nodules are often miniscule and often benign/indeterminate. This causes radiologist screening of nodules to be expensive, low-throughput, and often inaccurate. This study develops an algorithm that utilizes machine learning and radiomics to build a complete lung cancer diagnostic pipeline that addresses these issues.

Methods

A large patient history dataset was obtained from Kaiser Electronic Medical Records and a separate large lung-CT scan dataset was compiled and hand-modified from various online sources. In order to assess a patient's risk of lung cancer, a 50-tree Gradient Boosted Machine (GBM) was constructed that employs personalized patient history variables including age, prescriptions, ethnicity, body mass index, blood pressure, and diagnoses to better assess true risk of patients. Once a CT scan is conducted to identify malignant lung nodules, an ensemble of 3D Convolutional Neural Networks (CNNs) of discriminator VGG-like and U-net architectures, trained with multitudinous augmentations and gradient clipping on a hand-engineered dataset, determines nodule morphology (calcification, spiculation, size), position, and malignancy. From these features, a linear classifier predicts lung cancer development in one year.

Results

The GBM significantly surpasses current guideline assessments, capturing omitted patient groups at high risk for lung cancer (sensitivity increased from 23% to 88%). Moreover, the CNN Ensemble obtained statistically comparable predictions to radiologist readings of scans. Further, the Ensemble substantially reduced the false positive rate of Computer Aided Diagnosis models (from on average 15.28 to on average 1.68 false positives per scan).

Conclusions

This model could serve as primary screen for lung cancer nodules to decrease radiologist involvement in screening. The combined automated lung cancer diagnostic system increases early-detection rates and reduces false positive rates of Computer Aided Diagnosis systems, thereby greatly improving the timeliness, accuracy, and affordability of lung cancer detection.

Summary Statement

I developed a machine learning algorithm that acts as a viable fully-automated lung cancer diagnostic pipeline.

Help Received

Dr. Drew Clausen aided in obtaining a patient history dataset and answered various questions regarding exploratory data analysis and model architecture. I accumulated a separate CT dataset from online sources. I further explored data relationships, modified datasets, and developed the algorithms on my own.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)

Dennis Chan

Project Number

S0806

Project Title

Implementation and Validation of a Machine Learning Model for the Early Detection of Alzheimer's Disease

Abstract

Objectives

Alzheimer s Disease (Alzheimer s or AD) is an irreversible progressive brain disorder that slowly impairs memory, degrades thinking capabilities. Eventually, it can destroy the ability to carry out daily tasks and lead to death.

While there is currently no cure for Alzheimer s, early diagnosis can be crucial for victims to seek proper medical treatment and support services in the early stages of the disease.

The objective of this study is to develop a machine learning model for the early detection of Alzheimer's Disease using various biomarkers.

Methods

To implement the machine learning model, I have developed an Artificial Intelligence-based Alzheimer s early detection system which is made up of three encapsulated but interconnected components: data loading module, training modules, and evaluation module. By accessing publicly available ADNI (Alzheimer s Disease Neuroimaging Initiative) data, my data module extracts, merges, and prepares the data for the AI training module. I have used 250 brain MRI scan subject datasets, which includes 3 categories of data, from control normal, to mild cognitive impairment (MCI), to AD patients, to repeatedly train and evaluate the model. Lastly, 150 brain scan subjects datasets are independently tested through the validation module.

Results

My result shows an accuracy of 86 percent to distinguish a healthy brain from those with MCI.

Conclusions

My solution effectively identifies MCI, which can happen 6 to 10 years before onset of Alzheimer s. Therefore, it helps people to plan ahead while they are still able to make important decisions on their care and on financial and legal matters. It also helps their families and caregivers relieve from enormous stress as they face new challenges and alleviate the financial burden for our society.

Summary Statement

I developed a machine learning model for the early detection of Alzheimer's Disease

Help Received

I got help from Dr. Tosun, a researcher at UCSF, on the ADNI database and which dataset can be potentially useful for my study.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) Cynthia Chen	**Project Number** S0807

Project Title

Decoding Neural Networks: Novel Computational Methods to Discover Anti-Tumor B Cell Receptor Binding Motifs

Abstract

Objectives

Cancer accounts for over 8 million deaths worldwide each year. Studying the binding interaction between the B cell receptor (BCR) and the tumor antigen has become a promising field for building a better understanding of how the human adaptive immune system attacks cancer cells. Currently, the recurring protein sequence patterns (termed motifs) for the BCR binding region remain largely undifferentiated between cancer types. Determining these cancer-specific BCR binding motifs is crucial, as they can inform more targeted immunotherapies for cancer patients.

Methods

I developed novel computational methods to uncover the BCR binding motifs encoded in a deep neural network. In previous research, the deep learning model was trained on 3 million BCR sequences from the TCGA database and achieved an 0.8 average AUC in predicting cancer-specific BCR binding affinities for 13 cancer types. To decode the key motif information that allowed the model to accurately predict cancer types, I implemented a computational pipeline, developed with 3500+ lines of Python and R code. My pipeline consists of several algorithms: generating random input sequences, running the model to rank sequences, visualizing top sequences to distinguish binding patterns, and clustering to identify motifs.

Results

Using my pipeline, I discovered 65 BCR binding motifs among all 13 cancer types and identified the 12 most significant motifs overall. The robustness of the motifs was validated through a synthetic data simulation and extensive correlation analyses. Last, I demonstrated the versatility of my computational pipeline by applying it to antigen-specific sequences and full-length CDR3 sequences.

Conclusions

My research is the first to reveal and validate anti-tumor B cell receptor binding motifs for specific cancer types. This discovery is a key step towards future synthesis of new motif-based antibody drugs and more powerful and precise cancer treatments. The approaches and methods that I developed are versatile and applicable to other types of cancer and disease. Furthermore, the novel computational pipeline that I propose in my research can be reused and employed to decode a wide range of deep learning models and ultimately lead to more transparent and understandable AI.

Summary Statement

I developed a novel computational pipeline to decode deep neural networks, and I employed my pipeline to discover anti-tumor B cell receptor binding motifs.

Help Received

This research was conducted at Harvard Medical School under the guidance of Dr. Sherlock Hu. My science teacher, Mr. Christopher Spenner, advised me on the presentation aspect of my project.

Name(s) Sauhaarda Chowdhuri	**Project Number** S0808

Project Title

PhonoNet: Deep Learning for Raga Identification in Indian Classical Music

Abstract

Objectives

Indian Classical music is an improvisational form of music based on ragas, melodic frameworks which are passed down through a fading oral tradition. I aimed to provide computational prediction of ragas so singers can receive live feedback when learning and so important features of the music can be preserved digitally in my machine learning system.

Methods

First, my system computed the short-term-fourier transform of the input audio data to form a chromagram representation of the notes being sung. This data was augmented using a novel transpositional data augmentation algorithm and split into 150 second chunks. The raga information was learned from these chunk samples using a deep convolutional neural network. The deep network was then modified with a recurrent layer to allow processing of multiple chunks in sequence, allowing full length songs to be learned.

Results

Experiments identified the optimal system which uses 150 second chunks and 12 chroma bins with data augmentation. This joint system achieved 78.9% accuracy on raga prediction from chunks and 98.9% accuracy on identification from chunks, a new state-of-the-art for raga detection.

Conclusions

The PhonoNet system documents the structure of Indian Music with deep networks and provides live feedback mechanisms for learning the art form. The proposed hierarchical system can extend to other tasks with long temporal sequences, and the proposed data augmentation algorithm may be applied to any form of music processing. These other applications will be pursued as future work in addition to extending the PhonoNet system to other forms of world music.

Summary Statement

I created a computational system capable of understanding and preserving the fading art of Indian classical music.

Help Received

The open-source Hindustani music dataset used was obtained freely online through the Universitat Pompeu Fabra's website.

Name(s)	Project Number
Andrew Chu	S0809

Project Title

A Novel Model to Optimize the Efficient Use of Lithium-Ion Batteries in Renewable Energy Storage Systems

Abstract

Objectives

Lithium-ion batteries are a central component of all renewable technology today, including hybrid/electric vehicles and grid energy storage systems. However, these expensive batteries are used inefficiently and replaced too soon as there exists no accurate way to measure their remaining capacity outside a laboratory setting. The objective of this project was to develop a practical method to estimate the remaining useful life of a lithium-ion battery by building a novel, prognostic model that predicts capacity loss with high accuracy. Such an algorithm could report in real-time how efficiently a battery is being used and extend battery life by years.

Methods

Initial analysis of capacity loss versus multiple battery cell operating conditions was done using published data from lithium-ion batteries studied under different experimental conditions in a laboratory setting where capacity is easily measured. MathWorks' MATLAB software was used for data processing and analysis to build the model. The model was validated on a second generated data set. In parallel, a neural network was trained to predict capacity.

Results

The stochastic prognostic model was able to predict capacity loss and therefore remaining useful battery life with high accuracy (average error <1.5%) using only readily measured battery characteristics. In addition, a cubic relationship between capacity and resistance was found. This is the first time a correlation between these two variables has been reported. It was also found that a battery cell s ratio also significantly affected the rate of capacity loss. Trends between the error of the model and various battery operating characteristics were established, allowing capacity to be predicted stochastically within a narrow confidence interval. Based on these findings, the remaining useful life of a battery can now be stochastically represented by a third-degree polynomial function of resistance for differing cell ratio. The neural network predicted capacity within 0.5% but is not stochastic.

Conclusions

This is the first time a prognostic model that stochastically predicts the capacity loss of a lithium-ion battery has been created. This can be used to maximize a battery s efficiency and extend its life by several years, as well as increase driver safety and grid reliability through accurate estimation of remaining useful battery life. Since the model can incorporate multiple inputs (such intensity and frequency of use) it is also able to provide an output tailored to each individual user. This model has important implications at both consumer

Summary Statement

I built a prognostic stochastic capacity model that uses easily measurable battery characteristics to accurately predict in real-time the remaining useful life of lithium-ion batteries used in renewable energy storage systems.

Help Received

Drs. Simona Onori, Harikesh Arunachalam, Abdullah-al Mamun, and Anirudh Allam from Stanford introduced me to MATLAB and provided me with guidance.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)

Karen Chung

Project Number

S0810

Project Title

Integrating Mathematical Modeling with Machine Learning for Cancer Driver Gene Identification

Abstract

Objectives

The identification of cancer driver genes is a critical component of precision oncology. Given the large feature space of The Cancer Genome Atlas (TCGA), which catalogs millions of somatic mutations observed in human tumors, machine learning techniques are ideally suited to driver gene identification. In existing models, however, the objective assessment of such machine learners is complicated by unexplained errors in the mutational data used to train the algorithms and by the absence of a perfect drivers list. This study employs mathematical modeling in tandem with machine learning processes to construct an objective and accurate classifier that identifies cancer driver genes.

Methods

In-silico dataset generation and machine learning training use the programming languages Go and Matlab, respectively; external datasets (TCGA, Cancer Genome Consensus, KEGG Database) are used later for classifier evaluation. A set of in-silico mutational data is generated by the stochastic simulation of a differential equations model of feedback-controlled cancer population dynamics. The synthetic dataset, validated through the assessment of mutational pattern distributions, trains a selected machine learning algorithm, producing a driver gene classifier.

Results

The gene classifier is shown to prioritize high-impact driver genes in four cancer types with >85% accuracy. Additionally, the quality of the ranked list of putative driver genes is validated through enrichment analysis on a list of generally accepted driver genes and biological pathway analysis. Top colorectal cancer driver genes from the classifier hold key roles in the PI3K-AKT and Wnt pathways, which have well-documented implications in carcinogenesis.

Conclusions

The interdisciplinary methodology developed here produces a more efficient and unbiased cancer driver gene classifier that can be utilized to identify henceforth unknown driver genes, providing insight essential for targeted cancer screening and treatment.

Summary Statement

I used mathematical modeling and simulation to generate synthetic data, which trained a machine learning classifier that accurately classifies and prioritizes cancer driver genes.

Help Received

I worked in the lab of Dr. Qing Nie at UC Irvine. My primary mentor in the lab was Dr. Seth Figueroa, who advised me on mathematical modeling and data simulation, and since Oct. 2018, I have been mentored by Matt Karikomi. Synthetic data generation was performed on the UCI cluster.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) **Patrick Cui; Stephanie Zhang**	**Project Number** **S0811**

Project Title

PocketOnco: An App for Diagnosis-Prognosis of Colorectal, Breast, and Skin Cancer Using Convolutional Neural Networks

Abstract

Objectives

To (1) develop a more effective and accurate convolutional neural network (CNN) algorithm to validate, identify, and classify colorectal, breast and skin cancer through feature segmentation and (2) integrate a machine learning model into a mobile app platform with potential treatments and clinical trials.

Methods

H&E-stained pathologist pre-labeled breast histology slides were obtained from the online public 2018 Breast Cancer Histology Image (BACH) Grand Challenge. Colorectal histology slides were obtained from the 2015 Gland Segmentation Challenge Contest (GlaS). External dermoscopic images of skin cancer were obtained from the International Skin Imaging Collaboration (ISIC) Archive. Data augmentation techniques were then used to increase the datasets to a total of 5, 000 images. After testing both the CreateML framework in XCode and the Turi Create module in Jupyter Notebook, we compared the two network model accuracies and selected the model with the greater accuracy to be used within the app. We then developed the app in Swift with bridges to Objective-C, and potential treatments and clinical trials were encoded using JSON with reference to the American Society of Clinical Oncology (ASCO) and the U.S. National Library of Medicine.

Results

The final network exhibited an accuracy of 100% for validation for all cancers, diagnosis accuracy of 96%, 78% and 75% and prognosis accuracy of 76%, 97%, and 80% for skin, colon, and breast, respectively. Using a confusion matrix, we identified a sensitivity of 64.29%, 91.87% and 100% and specificity of 68.75%, 85.71% and 92.31% for colorectal, breast and skin cancer diagnosis, respectively.

Conclusions

PocketOnco is a novel, user-friendly iOS app that grades and stages colorectal, breast, and skin cancer through the integration of a custom deep convolutional neural network algorithm. The mobile app platform allows for the exploration of personalized medicine and healthcare, with the rise of electronic health records, novel treatments, and clinical trials.

Summary Statement

A multi-cancer diagnosis-prognosis app was developed to grade and stage colorectal, breast, and skin cancer through the integration of CNNs which analyze imported histology images or taken dermoscopic images.

Help Received

All data was acquired from public sources online. Neural network implementation, algorithm design, and app development were done independently without external guidance.

Name(s) **Adham Elarabawy**	**Project Number** **S0812**

Project Title

Predicting Optimal Farming Regions via Machine Learning Trained on Novel Vegetation Index

Abstract

Objectives

As the world population grows and the usable farming area shrinks, the demand for nutrition increases, leading to an escalating pressure on farmers to increase their yield. Due to this, many farmers resort to using fertilizer and utilizing biotechnology to make the individual products larger and more nutritious. However, the uncertainty of overall crop yield is yet to be fully addressed, which this project attempts to undertake. Precisely and reliably predicting plant growth and relative crop yield proves to be very helpful for developing countries as well as existing farms, as it allows farmers to invest an optimum amount of resources due to their knowledge of its respective potential, as well as helps optimize their farmland placement for yield and reliability.

Methods

Firstly, in order to quantify the potential for vegetation, a novel composite vegetation index computed through calculations on image bands of different wavelengths from multispectral images is derived. To compute the novel composite vegetation index (NCVI), multispectral images are obtained from the Landsat 6, 7, & 8 satellites, from which the infrared and red bands are retrieved. In order to prevent cloud reflectances from influencing the NCVI calculation and consequently impacting the prediction algorithm, machine learning is utilized to disregard pixels determined as clouds. Once the cloud pixels are filtered out of the satellite images, NCVI is computed and passed into the output dataset utilized by the prediction algorithm.

Results

The algorithm reliably predicts relative crop yield to within 10% for 3 subsequent seasons when trained on only 11 seasons of prior data. Furthermore, the predicted relative crop yield output from the algorithm is compared to the true crop yield of two industrial farms in Fresno, the results of which lead the refinement and further development.

Conclusions

As opposed to the current leading vegetation indices, not only does the novel inclusion of moisture into the vegetation index lead to more immediate observations in short-term environmental changes, but it also counteracts the typical saturation of vegetation indices at higher densities of vegetation through the increased granularity.

Summary Statement

A novel composite vegetation index computed from multispectral images is passed to a machine learning algorithm to identify and predict optimal farming regions.

Help Received

I developed and tested the software presented independently. I received some preliminary help and exposure to the topic from my previous software engineering internship with GroGuru Inc.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) **Raghav Ganesh**	**Project Number** **S0813**

Project Title

Developing a Novel, Accurate, and Rapid Machine Learning Based Skin Disease Diagnosis Algorithm and Mobile Application

Abstract

Objectives

To design and implement a novel, accurate, and rapid machine learning based classification algorithm, computer vision software, and a mobile application to classify dermoscopy images as benign or malignant in under 30 seconds, with an accuracy of at least 80%. This is the documented average accuracy of dermatologists with varying levels of experience.

Methods

I designed my own algorithms and developed my own software. I constructed the algorithms with Scikit Learn, Keras, Tensorflow, Python and OpenCV. I also developed a cross-platform (iOS and Android) mobile application with Cordova, as well as a cloud service backend with PHP. I trained and tested each of my algorithms with a dataset of 4000 skin lesion images (about two-thirds train, one-third test) that consisted of 2000 malignant and 2000 benign images. The images were taken from the public ISIC archive of dermatoscopic images.

Results

My system successfully achieved my engineering goals and design constraints, taking on average 22.47 seconds to classify an image when tested with 1320 images that were not used in training. At the 2.5% significance level, my algorithm s accuracy (93%), sensitivity (88%), and specificity (98%) outperformed the corresponding data published from dermatologists with all levels of experience.

Conclusions

This year in the US alone, 96,480 adults are estimated to be diagnosed with melanoma, and 7230 are expected to be fatal. Early detection of melanoma saves lives. It has been reported that up to 70% of melanomas were first discovered by patients and brought to the attention of dermatologists for evaluation. This project demonstrates that the diagnosing effectiveness for melanoma using novel computer vision and machine learning techniques is higher than that of data published from dermatologists with varying levels of experience. My frugal mobile innovation can assist patients with a decision on seeking further professional evaluation.

Next Steps: When possible, dermatologists also factor in family history and a comparison of changes to the lesion over time to make their decision to biopsy. To further improve the accuracy of computer assisted diagnosis, expanding the algorithms to factor in the temporal changes of a lesion can be explored.

Summary Statement

I developed and demonstrated a machine learning based frugal solution delivered as a mobile application that diagnoses melanoma with a higher accuracy, sensitivity, and specificity than dermatologists with varying levels of experience.

Help Received

I designed and programmed the algorithm and mobile application myself using my prior experience with computer vision and machine learning. I reviewed my project with a dermatologist and my mentor (high school science teacher).

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)

Anusha Ghosh

Project Number

S0814

Project Title

A Novel Program for the Detection and Translation of the ASL Alphabet through the Use of Deep Learning

Abstract

Objectives

American Sign Language (ASL) is a common way for people who are deaf, hard of hearing, or verbally impaired to communicate, with over one million people using it as a primary method of communication. While this language is prevalent in terms of usage, and has been cited as one of the mostly commonly-used languages used today, many people who don t rely on ASL to communicate have no knowledge or rudimentary grasp on the language. Because of this, there is currently an almost insurmountable language barrier between hearing people and the hard of hearing/deaf community. The goal of this project was thus to create a program that translates the ASL Alphabet in order to provide a means of communication that bridges the gap between these isolated communities.

Methods

I collected data from a variety of different sources, including both a self-generated dataset and data compiled from various publicly available services. This data was then sorted into classes by letter, and uploaded to an Amazon S3 server in preparation for training my model. I then trained my model with Pytorch, using a modified version of the standard Resnet18 architecture to accurately classify the data through deep learning. By using transfer learning to hasten the training process, I was able to achieve accurate results that could generalize well to other datasets that the model had not yet seen. Using this model, I was then able to create a program that could process webcam input from a user to translate the ASL alphabet in real time. I used OpenCV to pass camera input to the deployed model, which would then output the most probable letter as a translation. This end-to-end system created a reliable way to classify and translate the alphabet.

Results

By evaluating the model on a set of data that was different than that given to the model during training, I found that the model had an overall accuracy of 82%, which exceeded my goal for my model s accuracy. This accuracy data is also backed up by the accuracy data given while training, which showed an accuracy of over 90% with the training and validation data it had.

Conclusions

My program was also able to generalize well to real time usage, which shows that my model was successful at translation. My detection program also performed admirably and is able to accurately detect letters in under a second, meeting my second criteria for this project. This speed allows people fluent in ASL to sign naturally and means that my program adapts best to the real need of people in the ASL community.

Summary Statement

I created a program that can accurately translate the ASL alphabet in real time in order to provide a better means of communication for various disparate groups.

Help Received

I programmed the entirety of my project myself, using existing documentation provided by the makers of Pytorch. Jason Su answered questions I had.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)

Pranav Kakhandiki

Project Number

S0815

Project Title

Automated Diagnosis of Aortitis Using Machine Learning

Abstract

Objectives

Aortitis is a rare heart disease characterized by inflammation in the aorta. My goal is to create a reliable tool in diagnosing aortitis by developing an accurate machine learning algorithm to analyze CT scans of the heart. Specifically, the algorithm analyzes the aorta and determine whether it is inflamed. Aortitis frequently goes undiagnosed due to its rarity, so my program will assist doctors in diagnosing it early.

Methods

To store and compare the CT scans, the python program uses HOG (histogram of oriented gradients) descriptors. They store a histogram of gradients (consisting of x and y derivatives which have direction and magnitude), which is more efficient than storing the entire image because the useful data consists of abrupt changes in the derivatives. In the case of aortitis, likewise, the key difference between an inflamed aorta and a normal aorta would be an increased area with a lesser change in the derivative (due to the thickened aortic wall). For classification, Linear-SVC, an algorithm which establishes a hyperplane between clusters of data, was used. LinearSVC uses the parameters which the HOG descriptor provides to train the program and draw the hyperplane, effectively classifying each image as either having or not having an inflamed aorta. This unique combination of Linear-SVC and HOG descriptors is more accurate than generic deep learning models as it is customized and specific to aortitis.

Results

With an overall accuracy rate of 94% and a type II (false negative) error rate of only 1.4%, the algorithm proves to be effective. Using the program along with analyzing ESR and CRP levels (two tests in which the levels will be abnormally high in an aortitis patient) makes diagnosing aortitis reliable and less error-prone for doctors. Studies show ESR and CRP to be roughly 96% accurate, so just with my algorithm and these levels alone, Doctors can diagnose aortitis with a 99.8% accuracy.

Conclusions

I developed a unique and innovative machine learning algorithm which uses HOG descriptors to extract features and Linear-SVC to classify CT scans of the heart to determine whether the aorta is inflamed. In short, I constructed a computer vision program to accurately diagnose patients with aortitis, a rare heart disease. My project expands our horizons in the field of diagnosing rare diseases as well as developing unique machine learning models.

Summary Statement

I developed a unique and novel machine learning algorithm to diagnose aortitis, a rare heart disease.

Help Received

My AP statistics teacher helped me with the 'Data Analysis' section of my project.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)	Project Number
Sruthi Kalavacherla	S0816

Project Title

Controlling the Chikungunya Virus in Dengue Endemic Areas through the Development of a Peptide Vaccine

Abstract

Objectives

The Chikungunya and Dengue viruses are arboviruses transmitted by the mosquitoes Aedes aegypti and Aedes albopictus. Co-circulation of Dengue and the Chikungunya virus (CHIKV) make it difficult to distinguish between the two diseases. As both cause febrile symptoms in the initial stages of infection, CHIKV is often masked and misdiagnosed as Dengue in Dengue endemic areas. This misdiagnosis affects how the symptoms of each disease are treated. This study targets the non-structural protein 2, nsP2, of CHIKV which has peptidase and helicase functions. nsP2 plays an important role in viral replication, the cleavage of the viral non-structural polyprotein, and inhibiting the protective immune response in the host cell. Thus, epitopes in nsP2 are ideal for a peptide vaccine.

Methods

In this study, potential cytotoxic (CD8) T cell epitopes in nsP2 that will elicit an immune response are determined through an immuno-informatics approach, using tools such as the Immune Epitope Database and Analysis Resource (IEDB), Net CTL Version 1.2, and VaxiJen. To test the accuracy of the procedure, the structural polyprotein of the Ross River virus is used as a control.

Results

Out of the top 82 CD8 T cell epitopes identified, three novel epitopes, YTYNLELGL, SILERKYPF, and MNNQLNAAF, are selected. These epitopes are highly immunogenic, show maximum binding to MHC Class I alleles, and are conserved in 99 strains. For the first time, a population coverage analysis is done on these epitopes by analyzing their binding ability to 1,089 MHC Class I alleles.

Conclusions

The development of a vaccine for Dengue is difficult because of its multiple serotypes and antibody-dependent enhancement of infection. As far as is currently known, all strains of Chikungunya belong to a single serotype. Therefore, an effective vaccine against CHIKV can help clinical management in CHIKV and Dengue endemic areas.

Summary Statement

I computationally designed an effective peptide vaccine against the Chikungunya virus which would aid clinical management in Dengue endemic areas.

Help Received

My project was performed independently at my house. I devised my method myself after a literature search on techniques, and I had my mom look over my project

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)

Shreyas Kallingal

Project Number

S0817

Project Title

Waterborne Parasite Infection Risk Mitigation via Microscopy-based Assays and Parasiticide Proteases

Abstract

Objectives

Waterborne disease remains an economic and health burden to approximately 2 billion people, particularly in impoverished regions. Current control methods for waterborne parasites are highly inefficacious and fail to manage parasite populations prior to infection. The objective of this study is to 1) create an automated assay for parasites in water bodies and 2) computationally develop a class of parasiticide proteases to reduce parasite populations.

Methods

Adaptive noise inclusion and feature extraction were performed on an image and video dataset of 5 parasite classes (Cryptosporidium, Schistosoma, etc). Individual organisms were then segmented out via Teh-Chain Approximation contouring and compiled into an ordered dataset. To address phenotypic similarity amongst parasites, a Long Short-Term Memory (LSTM) model was employed in conjunction with a Convolutional Neural Network (CNN) for contextualized image classification of the segmented organisms. Next, proteins from WBPS12 and EuPathDB were evaluated via BLAST and a literature search to find target biomarker protease substrates (TBPS) that were expressed externally in each parasite. A script was written to read PeptideCutter and PROSPER tool results for each TBPS and generate the most probable cleavage sites for protease development. PepComposer results were generated for each TBPS and analyzed to finalize the parasiticide protease scaffolds.

Results

The preliminary segmentation algorithm had an error rate of 1.2% and gathered novel, real-time metrics that were unattainable through previous methodology. Moreover, the hybrid LSTM-CNN model was 97% accurate for multi-class contextualized image classification. Additionally, parasiticide protease scaffolds were evaluated through FoldX energy scores, and 5 were selected with the highest stability and specificity.

Conclusions

This proposed system consists of the assay (1) to determine which parasites are present, after which the appropriate parasiticide (2) is disseminated to reduce parasite presence in a water body. The automated assay demonstrates unparalleled accuracy and range in parasitological study, and the designed parasiticides target parasites precisely. Furthermore, this system is extensible to other organisms due to the retraining potential of the developed pipeline. Thus, water bodies can be efficiently monitored and treated, thereby mitigating exposure and infection.

Summary Statement

I developed an inexpensive, accessible system using computational modeling to detect and treat parasites in water bodies of impoverished regions.

Help Received

Wherever appropriate, I cited freely-available published studies and web-based applications that I used. All model development, bioinformatics analysis, and experimentation was done independently at home by me.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) **Anthony Kim; Valmik Ranparia; Sky Shia**	**Project Number** **S0818**

Project Title

Sound Localization and Noise Cancellation to Assist the Hearing Impaired

Abstract

Objectives

Hearing loss can be caused by multiple factors, such as genetics, aging, or exposure to noise or infections, rendering it a difficult medical disorder to remedy with a cure-all solution. Deficits in hearing can contribute to onset of depression and anxiety and were identified in 2005 as an area of emerging research, particularly in its application to children. However, expensive rehabilitative services remain a high barrier to entry for many in the developing world.

Our purpose is to identify a cost-effective solution. The goal is to create a smartphone-based device that visualizes real-time locations of environmental sounds on a 3D-grid interface. The aim of this study is to determine accuracy and viability of our sound localization method.

Methods

An audio sample was played from a fixed position and recorded using 4 microphones. Recordings were compared by cross-correlation and yielded the calculated position of the sound source. This calculation was compared to known origin of sound. Noise cancellation was performed by removing wall-reflected and background noise. FFT-IFFT technique was used to further remove high-frequency noise.

Results

Analysis of calculations revealed that most origins were found within a 40-degree angle of error without performing noise cancellation. Analysis of noise-cancelled data indicated that error decreased by half.

Conclusions

This proof-of-concept study yielded promising rates of accuracy of the sound source position detection with pure signal analysis only. The result can provide a cause to begin prototyping a visualization device.

Summary Statement

By recording the sound with 4 different mikes and analyzing it with matlab, we could accurately find the exact sound source position.

Help Received

Dr. Young Kim in Northrop Grumman reviewed the noise cancelling equation derivation

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)	Project Number
Benjamin Lipman	S0819

Project Title

Accurate Identification of Cardiac Anomalies through Deep Learning

Abstract

Objectives

The objective of this research was to develop a neural network to automatically identify arrhythmia in the electrocardiogram (ECG) at diagnostically accurate levels.

Methods

108,240 heartbeats were extracted from the MIT-BIH Arrhythmia Database, comprising 11 beat types (normal beat + 10 types of arrhythmia) using the WFDB software package. Heartbeats were constructed encompassing the PQRST complex, from 0.25s before to 0.45s after the R peak, for a total of 256 data points per beat at 360 Hz. The heartbeats were converted into 256x256 grayscale images and labeled by heartbeat type. The images were randomly shuffled and divided 80/10/10 into training, validation, and test sets. A convolutional neural network comprising 4 convolutional layers and 2 fully connected layers was developed in Python using the Keras and TensorFlow machine learning packages. The network was trained over 500 epochs, achieving maximum training and validation accuracy without overfitting. Misclassified heartbeats were reviewed to identify potential sources of error during hyperparameter tuning. The test data set was evaluated by the network.

Results

The neural network achieved accuracy of 99.06% on the test data with an f1 score of 0.99. Individual beat level precision ranged from 0.846 to 1 and recall ranged from 0.733 to 0.999. Beat types of smaller sample size had precision and recall at the lower end of the range, although still at diagnostically useful levels.

Conclusions

This research demonstrates that deep convolutional neural networks can accurately classify cardiac arrhythmia from heartbeats that have been converted into images. This approach requires no feature engineering or noise reduction, producing high precision and recall results at diagnostically meaningful levels across all arrhythmia beat types. Although this project used a relatively small amount of data on a consumer-class GPU, the results suggest that this is a promising approach for further research at larger scale.

Summary Statement

I developed a convolutional neural network that automatically identifies cardiac arrhythmia in ECG data at diagnostically accurate levels.

Help Received

I designed and programmed the neural network myself.

Name(s)	Project Number
Jiaju Liu	S0820

Project Title

A Novel Approach for Understanding Early-Stage Epileptogenesis via Nonlinear Manifold Learning

Abstract

Objectives

High-frequency oscillations (HFOS) in EEG data are thought to be promising biomarkers of epileptogenesis. Instead of examining data from patients living with epilepsy, EEG data were analyzed from patients in the early stages of post-traumatic epileptogenesis. The goal of my study was to analyze large amounts of EEG data in a fully automatic, computationally efficient way and provide a meaningful clustering of the data in which points embedded to the same cluster have similar local geometries.

Methods

Five patients were analyzed, each with 12 hours of scalp EEG data obtained within 48 hours of initial brain trauma. Data was downloaded from the University of Southern California Laboratory of NeuroImaging. After downloading, the data were preprocessed with a surface Laplacian and IIR filter. Events of interest were identified with an energy-based approach. Finally, Unsupervised Diffusion Component Analysis was performed to cluster the data and detect relevant patterns.

Results

A total of 6,384 HFOs were detected and 79.94% were embedded near the origin of the graph. Upon visual inspection of the clusters formed, the cluster at the origin was composed of short spike artifact while clusters further away contained genuine HFOs.

Conclusions

Although no predictions may be made regarding whether the patients will develop epilepsy, the program outputs high-frequency waveforms and the embedded graph provides a guide to further visual inspection. Rather than examining thousands of events generated by existing HFO detectors in the literature, the nonlinear embedding allows epileptologists to only examine a few events in each cluster. In addition, the program was able to detect HFOs in scalp EEG which has been seldom used for HFO detection.

Summary Statement

I coded a fully automatic, computationally efficient program that detects and clusters high-frequency waveforms in noisy scalp EEG data which show potential as biomarkers of epileptogenesis.

Help Received

I was pointed towards manifold learning and given a paper on USC's overall plan for finding a biomarker of epileptogenesis. From there, I self studied Fourier Analysis, graph theory, learned MATLAB, proposed, and wrote my code by myself. My mentor registered me for an account to access EEG data and for cloud

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)

Natasha Maniar

Project Number

S0821

Project Title

MapAF: Deep Learning to Improve Therapy for Complex Human Heart Rhythm Abnormalities

Abstract

Objectives

Atrial fibrillation (AF) is a chaotic and irregular electrical disturbance of human heart rhythm that affects over 33 million individuals worldwide, causing serious fatal health effects, yet for which therapy is poor. Recent studies that have used voltage mapping videos reveal areas of rotational sources in the atria, for which ablation (burning of diseased tissues) at these particular regions terminates AF. Unfortunately, current methods to identify AF sources from these complex mapping videos are solely manual, therefore limited and subjective, with an average of 60-70% accuracy. I developed MapAF, the first computational approach to automatically recognize the location of these AF rotational sources from within chaotic electrical patterns.

Methods

Electrical mapping videos from 35 de-identified patients undergoing ablation with persistent AF were collected. Data was pre-processed in MATLAB so that each video was split into 5000 images. Each image was labeled either rotational or non-rotational by an expert. A multi-layered convolutional neural network (CNN) was then implemented and datasets were filtered through 25 different feature extraction and classification layers to classify the output. I used the AlexNet architecture as the base for developing the network. I then trained an unsupervised learning algorithm on the images which contained rotational source (s). Using the principal components of the images as inputs, I clustered the images into K=2, 3, and 4 groups using the k-means clustering algorithm.

Results

The sensitivity of the CNN is 97% and the specificity is 93%. Therefore, in blinded testing, the supervised network is 95.0% accurate for potential sources and detected all sites of AF termination, more accurately and efficiently than medical experts. Individual patients can be grouped into K=3 clusters based on their types of images in 3 clusters: Cluster A showed more ambiguous rotations of small domain sizes, Cluster B contained larger rotational patterns, and Cluster C showed unclear rotations.

Conclusions

This is the first tool to identify sources for AF and provide insight into its mechanisms. Promising unsupervised data showed that clusters could be linked to an individual patient s history and shed light on which patients may have more advanced AF for tailoring treatment. MapAF may standardize and streamline the treatment procedure by eliminating subjectivity, reducing the expert physician time to detect sources, and being accessible in areas lacking experts, making a potential global impact on AF therapy.

Summary Statement

MapAF is the first computational tool using machine learning to automatically pinpoint sources of AF better than medical experts, thus improving AF treatment.

Help Received

I conducted my research at Stanford University where the PI of my lab, Dr. Sanjiv Narayan, and two other graduate students assisted me with implementing the machine learning model.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)

Andrew Nazareth

Project Number

S0822

Project Title

Predicting the Presence of Pneumonia in Chest X-rays Using Deep Learning with Convolutional Neural Networks

Abstract

Objectives

Pneumonia is an infection that causes lung inflammation. In the US, 1 million people with pneumonia are hospitalized annually resulting in 50,000 deaths. A 2017 Stanford ChexNet study suggested that radiologists have a 95% accuracy in detecting pneumonia from chest X-rays. My goal is to train a Convolutional Neural Network (CNN) to meet or exceed this threshold.

Methods

- 8000 chest pre-classified (NORMAL, PNEUMONIA) X-rays from Kaggle.
- The resnet set (resnet34, resnet50) of CNN's from fastai pretrained on regular (non-medical) images,
- Linux hardware with a Nvidia GPU from Paperspace
- Software utilities: FastAI a framework for fast training CNN s, Python, Jupyter Notebook.

1. Pre-process the X--rays, randomly separating them into training (80%) and validation (20%) sets.
2. Select and train the resnet34 CNN to recognize X-rays that have pneumonia:
 - Measure the prediction accuracy of the pre-trained network,
 - Train the outer layers; re-measure the accuracy and loss rates
3. Improve the accuracy of the pre-trained model.
 - Identify a good learning rate.
 - Unfreeze the hidden layers and retrain the network.
4. Use input and test time data augmentation to improve the prediction accuracy
5. Repeat steps 2-4 to see if a deeper CNN s (resnet50) can provide better accuracy.
6. Validate results on random chest X-rays and correlate results with practicing radiologists.

Results

By training the outer layers only, I achieved a prediction accuracy of 95.6%. Using a graph of learning rate versus validation loss, I selected a learning rate of 0.05. With this learning rate, the prediction accuracy decreased marginally to 95.3%. With the addition of data augmentation and training the network for 3 epochs, the prediction accuracy increased to 97.1%. Furthermore, unfreezing the hidden layers and adding a differential learning rate yielded an accuracy of 98.1%.

Conclusions

CNN's can be used to predict the presence of pneumonia in a chest X-ray with > 98% accuracy. After tuning, false negatives were under 2% and false positives were 1%.

Summary Statement

I developed a Convolutional Neural Network that accurately predicts the pressence of pneumonia in chest X-rays.

Help Received

Erik Perkins is my project advisor at school. Dhar Rawal mentored me and provided me with the machine learning knowledge to help me to successfully undertake this project.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)

Nathan Oh

Project Number

S0823

Project Title

Modeling Intersections as an Asymmetric Non-Zero-Sum Game for Maximin Decision Theory and Traffic Flow Optimization

Abstract

Objectives

Develop a model for traffic flow that aligns with game theory rules. Use a decision rule to optimize and automate intersection lights.

Methods

Laptop with processing.py compiler and Python. Model tested using intersection data from Von Karman Ave crossing Campus, Martin, Dupont, and Michelson provided by the ITRAC (Irvine Traffic Research and Control Center).

Results

The model produced a 0-5% decrease in stopped time for the average car, while remaining autonomous and needing no infrastructure to implement.

Conclusions

My model slightly increased the efficiency of traffic flow by the measure of wait time, and requires no additional data-gathering infrastructure to implement. This is done through modeling intersections using game theory rules, and then applying Wald's Maximin decision theory and numerous pruning strategies to find the best possible light times. While the results show that game theory can be used to solve system-oriented problems, it cannot be used realistically yet due to the unpredictable variables in real traffic.

Summary Statement

I created a model of traffic with game theory and devised an algorithm to increase traffic efficiency.

Help Received

I learned game theory concepts independently using online resources. Information about how traffic is managed and real intersection data was provided by Mark Ha and Chris Lee from the ITRAC (Irvine Traffic Research and Control Center).

Name(s)	Project Number
Anjo Pagdanganan	**S0824**

Project Title

A Deep Learning Approach to E. coli Epidemic Prediction

Abstract

Objectives

E. coli ravaged the US in 2018, with the first outbreak tracked back to Yuma, Arizona infecting 210 and the second, tracked back to central California infecting 62. Outbreaks of this severity create paranoia around the produce at hand throughout the US. This leads to numerous agricultural industries coming under fire for an outbreak that they may not even have caused, such as the Salinas Valley. This damage is only dragged on considering that the CDC takes two to three weeks to recognize such an outbreak. Thus, this project applies a deep learning approach towards predicting trends in E. coli outbreaks to minimize the damage dealt to agricultural industries.

Methods

Using Python, I obtained search popularity data of phrases correlating with "e coli symptoms" from Google through the pytrends library. I also scraped approximately 13 years of weekly E. coli case data from the CDC with several web scraping libraries (namely BeautifulSoup and TQDM). I then used this data to train and evaluate an LSTM (Long-Short-Term-Memory) neural network with Keras. As a baseline metric, a naive forecast using the current week's data as predictions was used.

Results

RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) were used as accuracy metrics. The dataset obtained was split into 70% training data, 15% validation, and 15% testing, leaving approx. a year or two of testing data from each of the 9 CDC census regions. With an RMSE of 5.20 and MAE of 3.05 compared to the naive forecast's RMSE of 6.18 and MAE of 4.24, the model exceeded baseline performance, showing that it has predictive power. The model does have a slight flaw in that its predictions peaks around 15-25 cases, highlighting that it likely needs more data for a proper fit.

Conclusions

To my knowledge, this is the first time this technique has been used to predict food poisoning cases, although it was inspired by similar applications in dengue outbreak prediction. Despite being trained on a very small dataset, the LSTM neural network was able to recognize trends in E. coli epidemics. In the future, this model could potentially be developed into a dynamic E. coli epidemic prediction system, constantly tracking search queries and current E. coli data for training and future predictions.

Summary Statement

I applied forecasting techniques with deep learning towards predicting the number of E. coli cases in a specific region on a week-by-week basis.

Help Received

None. I researched all of the algorithms and scraped all the data by myself.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)	Project Number
Rishab Parthasarathy; William Zhao	S0825

Project Title

SkinSight: A Novel Implementation of a Convolutional Neural Network to Recognize Skin Diseases

Abstract

Objectives

Only 35% of skin disease biopsies are performed by dermatologists, and non dermatologists only correctly diagnose skin diseases with 23.9% accuracy. Many methods have tried to solve this issue, but most are computationally heavy, while algorithms that can segment lesions have not been used for diagnosis. This project aims to design a computationally light and effective convolutional neural network (CNN) to diagnose skin lesions while also evaluating two common deep learning based approaches, Mask R-CNN and FCN, for multiclass segmentation against results from the ISIC 2018 Challenge: Skin Lesion Analysis Towards Melanoma Detection.

Methods

Our physical materials included a laptop with Anaconda installed on it. The dataset was provided by DermnetNZ, but since there are only 4374 images, we used data augmentation to make it larger. To train Mask R-CNN and FCN, we produced a dataset of 520 images with pixelwise masks identifying the lesions. Then, we used code for Mask R-CNN and FCN from Matterport and Sagippel s Github, respectively, and for transfer learning, we used the Tensorflow Image Retraining Tutorial. For our CNN, we used transfer learning, where we took an existing CNN, Inception-v3, and added a classification layer. To improve accuracy, we grouped diseases by physical traits. For Mask R-CNN, we wrote a new supervision and dataset loading file and trained it for 60 epochs. For FCN, it trained for approximately 40 epochs.

Results

We found that as the size of each group for our transfer learning decreased, accuracy improved, but human error increased when choosing the correct group. As the number of training steps increased, the accuracy increased. For the segmentation techniques, the Mask R-CNN achieved 0.9053 mean average precision and a 0.7532 Jaccard Index while the FCN achieved a 0.7241 Jaccard Index.

Conclusions

Having 15 diseases per group and 16000 training steps offered the best cost/efficiency tradeoff for its 86.5% balanced accuracy and 0.85 average F1 score. FCN was less effective than Mask R-CNN because it introduced interference into its predictions, leading to a lower Jaccard Index. The transfer learning and segmentation methods we tested in this study achieved comparable results to studies in the ISIC 2018 Challenge which achieved 85.3% balanced accuracy and 0.813 Jaccard Index, proving the feasibility of both transfer learning and multiclass segmentation.

Summary Statement

We implemented an effective and efficient transfer learning approach to skin lesion diagnostics, and we also proved that multiclass segmentation of skin lesions is feasible, with the best method of doing so being Mask R-CNN.

Help Received

None. We worked on the project by ourselves.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) **Deepro Pasha**	**Project Number** **S0826**

Project Title

Developing an Artificial Intelligence-Based Assistive Robot: A Novel Approach to Prevent Falling in Elderly People

Abstract

Objectives

The main objective of this project was to design, build, and program an artificial intelligence based assistive robot to prevent falling in elderly people. A combination of appropriate hardware design and software interface (programming) was used. The main hypothesis was to test if an artificial intelligence (AI) based assistive robot was possible to be built and programmed that can detect and act to prevent falling.

Methods

Simulation, image processing, blob detection, sideways detection using Harris corner detection algorithm, were used to set three parameters such as Parameter 1: Is height longer than width? Parameter 2: Is width continually increasing? Parameter 3: Is height continually decreasing? to detect falling. Depending on if each of the parameters are met or not, outputs were generated. The outputs were analyzed in python and were turned into one large array input that was read by a pre-trained back propagation neural network to determine whether or not it should take action. The steps were a) using OpenSim simulation software, example videos of falling were developed. b) based on the fall videos, sideways perspective and forward perspective fall detection algorithms were developed. c) algorithms were then used to develop codes and were run for multiple test cases d) a prototype of fall catching assistive robot was designed. e) hardware and software installation was performed. f) the effectiveness of the prototype robot was tested with test cases using dolls. g) algorithm was refined and the testing was repeated for improvement. Hardware used: Raspberry Pi 3, Ribbon Cable for GPIO, 40 pin breakout board, Breadboard, Jumper Wires, Motor (12V DC), L298 Motor Driver, 2nd Generation Raspberry PI camera, 610 mm long flex cable,12V AC to DC adapter, wood, insulation and Software used: C++, Python 3, OpenCV, Raspberry Pi GPIO, OpenSim, Visual Studio.

Results

Final outcome of this project was a prototype assistive robot that was capable of detecting fall, capable to react to prevent fall within 2-3 seconds, capable of moving the robot arm at different angles.

Conclusions

My robot was tested on small dolls and was successfully detected and prevented falling. It was a prototype robot. For future improvement, the algorithm can be tested with large number of fall cases to further refine it. The response time of the robot can be improved from 2-3 seconds to real-time.

Summary Statement

My project is about developing an artificial intelligence based assistive robot to detect and prevent falling in elderly people.

Help Received

Discussed informally few times with mechanical engineering professor Dr. The Nguyen at California State University, Fresno and science teacher of my school Mr. Matthew Carter for suggestions and advice but conducted the project entirely by myself at home and no lab facility was used.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) **Govind Pimpale; Marek Pinto; Nitish Reuben**	**Project Number** **S0827**

Project Title

A Modular and Dynamic GPU-Based Maize Simulation Using L-Systems

Abstract

Objectives

Plant field testing is often a long and costly process that is crucial to the development of efficient agricultural techniques. We aimed to create a plant topology simulator that presents a significant improvement in speed over other plant simulators by utilizing the GPU in order to model plant growth based on abiotic factors accurately.

Methods

Each plant, composed of a binary tree structure, is grown iteratively. During each iteration, both the plant and the external environment are updated. Each node keeps track of values for various factors, which affect its growth in each iteration. Nodes are evaluated in parallel using the GPU, which provides significant performance gains and runs stably on most consumer grade computers. We chose to focus on modeling corn due to its status as a staple food in many parts of the world.

Results

Our final program was able to generate data that was within a 95% confidence interval for actual plant dimensions and resource consumption. Errors can be attributed to lurking and confounding variables unable to be distinguished by the given datasets.

Conclusions

Our program has applications in 3D modeling of plants, particularly for real-time renders/simulations and uses in plant placement and crop yield optimization. In the future, we hope to expand our program to include a full-fledged plugin system to accommodate other types of plants.

Summary Statement

We created a computer model that utilizes the GPU to simulate large scale plant growth based on several abiotic factors.

Help Received

We designed, programmed, and tested the algorithm independently. We received help from a mentor at our previous science fair and our statistics teacher when verifying our results. We also used data from federal databases, primarily data on plant height and leaf length from the USDA.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)	Project Number
Amol Singh	S0828

Project Title

STAC-STIC: Novel Computational Pipeline to Generate Digital Super-Resolution Static Representations of Pathology Slides

Abstract

Objectives

Accurate digital pathology image analysis depends on high quality images. As such it is imperative to obtain digital images with high resolution for downstream data analysis. While hematoxylin and eosin (H&E) stained tissue section slides from solid tumors contain 3-dimensional information; this data has been ignored in digital pathology. In addition, in cytology and bone marrow aspirate smears, the 3-dimensional nature of the specimen has precluded efficient analysis of such morphologic data. An individual image snapshot at a single focal distance or of a single scene is often not sufficient for accurate diagnoses and multiple whole slide images at different focal distances are necessary for diagnosis.

Methods

I describe a novel computational pipeline and processing program for obtaining a super-resolved image from multiple static images at different z-planes from a single microscopy video. This program STAC-STIC, in part, uses MULTI-Z, a program that constructs a final super-resolution image, as well as a novel image alignment and stitching program: V-STIC. This program performs image alignment, Gaussian smoothing, Laplacian filtering, homography calculation, perspective warp, and array manipulation to construct the final super-resolution static slide representation.

Results

I applied this algorithm and program to images of cytology and H&E stained sections and demonstrate significant improvements in both resolution and image quality by objective data analyses (24% increase in sharpness and focus).

Conclusions

With the use of our program, super-resolved Whole Slide Images (WSI) images of cytology and H&E stained tissue sections can be obtained to allow for optimal downstream computational analysis, hospital documentation, easy proliferation and collaboration, and can serve as the medium for the primary diagnosis. This method is applicable to whole slide scanned images.

Summary Statement

This software uses slow-motion videos to generate images of stained cytology slides that preserve vital 3D information, a promising advance over current 2D digital methods and a path to earlier, faster, and more accurate primary diagnosis.

Help Received

I received help with access to training data and troubleshooting any issues in the project.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)

Anish Singhani

Project Number

S0829

Project Title

Real-Time Freespace Segmentation Using Deep Learning on Autonomous Robots for Detection of Negative Obstacles

Abstract

Objectives

Many small unmanned ground robots are being developed to perform tasks such as delivery, surveillance, household tasks, and many other formerly-human tasks. It is essential to these robots' core functionality that they are able to navigate difficult terrain, requiring advanced perception capabilities. Although a significant amount of work has been done on the detection of standing obstacles (solid obstructions), almost no work has been done on the detection of negative obstacles such dropoffs, ledges, downward stairs. Detecting these negative obstacles using reliable, cost-effective sensors is crucial for the success of autonomous robots.

Methods

Small autonomous robot running Robot Operating System and an embedded GPU which was used to run the neural network, along with a desktop GPU used to train the network.

Results

This research developed a method of terrain safety segmentation using deep convolutional neural networks. The custom semantic segmentation architecture uses a single camera as input and creates a freespace map distinguishing safe terrain and obstacles. The network was trained using heavy data augmentation, enabling the network to generalize well, even when using very small hand-labeled datasets. The results showed that the system generalizes well, achieving around 94.9% mIOU accuracy on the validation dataset.

Conclusions

The neural network is deployed to an embedded GPU on an indoor robot. Because of its computationally-efficient design, the network is able to run at 55 fps and create a freespace map that can be used to create a costmap for navigation and obstacle avoidance. Experimentation with the neural network combined with pathfinding algorithms proved the robot's ability to reliably detect and navigate around both standing and negative obstacles in real-time, using only an RGB camera and the neural network developed in this research.

Summary Statement

I developed a novel method of terrain safety segmentation on autonomous robots using deep neural networks.

Help Received

None. I developed and trained my neural network, and tested it on the robot, completely by myself.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)	Project Number
Samyak Surti	S0830

Project Title

Cellular Automata-Based Mathematical Model for the Spread of Forest Fires

Abstract

Objectives

Taking into consideration the most important physical factors of fire spread, the objective is to develop a cellular automata-based mathematical model that can accurately predict the spread of forest fires in order to aid firefighters in fire containment and evacuation.

Methods

A cellular automata is developed modifying a simple Python-based cellular automata framework developed by Luis Antunes that uses matplotlib. His single layered cellular automata was modified to accommodate additional layers representing the physical factors considered, such as fuel load, wind patterns, and topological features. Functions to initialize the maps were added. To evolve the configuration of the cellular automata, a transition rule that defines the propagation of the fire is defined in the program, taking into account, separately, the effect of each of the physical factors.

Results

Under hypothetical ideal conditions, where the fuel is spread uniformly across the lattice and there are no wind patterns or topological features accounted for, the fire simulated propagated in a perfect circle. As physical factors were added in the program one by one, the shape of the fire's spread shared striking similarities to the empirical Rothermel model. Indigenously developed, the mathematics behind the propagation of the fire, in the context of the cellular automata, matched up with previously developed empirical models. However, new observations made by firefighters in the Carr and Camp fires in California were incorporated, that made the model potentially more accurate than existing models.

Conclusions

The performance of the cellular automata-based mathematical model demonstrates the effectiveness of modeling complex dynamic systems such as fires using a simple set of rules derived primarily from vector calculus. This approach provides, not only a more succinct way to define the spread of forest fires in comparison to previously defined empirical models, it also provides an avenue to model other similar complex dynamic systems. Computationally, cellular-automata based approaches can be made even more efficient using GPU-based parallel processing.

Summary Statement

A cellular automata-based mathematical model is developed from scratch to accurately simulate the spread of forest fires.

Help Received

Luis Antunes' GitHub page aided me in working with a simpler cellular automata framework without having to create one of my own, allowing me to focus on the mathematical aspects of the model. My father helped me understand how to use LaTeX.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) **Tirth Surti**	**Project Number** **S0831**

Project Title

Galaxy Morphological Classification through Convolutional Neural Networks

Abstract

Objectives

The creation of new telescopes in the future will lead to the generation of more images of new distant galaxies which will need efficient and accurate classification to better understand the processes that govern the evolution of universe. Thus, the objective of this project is to design, train, and test a neural network that can classify different types of galaxies based off of their morphological structures.

Methods

A dataset of over 100,000 images of hand-classified galaxies is obtained online and separated into sub-sets of elliptical and spiral galaxies for classification task one, specific spiral types (Sa and Sc) for task two, and specific elliptical types (E0-E3 and E4-E7) for task three. These datasets will be trained, processed, and tested on a convolutional neural network. A neural network is then created with multiple convolutional layer blocks for the processing of input images and for the generation of classification probabilities. This same neural network is used for all three classification tasks. For each task, the model weights are saved after training and used to predict on unseen datasets of galaxies to generate the official accuracy of the neural network.

Results

For the objective of classifying between elliptical and spiral galaxies, the training accuracy was 94% with a test accuracy of 100%. For the classification between spiral types, despite the training accuracy being 80%, there was 100% accuracy on the test dataset, and for the classification between elliptical types, there was a training accuracy of 95% with a test accuracy of 100%. For each classification task, the test accuracy was a result of the model predicting on 5 unclassified galaxy images and outputting a probability array of what the machine thinks is the classification.

Conclusions

The results demonstrate the capabilities of a convolutional neural network to classify different kinds of galaxies in place of simple and unreliable hand classification because it is not only fast, but accurate to a significant degree. As a result, this galaxy classifier can have large implications when new galaxies are imaged because the model can be used to rapidly and accurately classify those galaxies, giving a more accurate insight into the processes how the universe has evolved over time.

Summary Statement

I created a machine learning model that could accurately and efficiently classify images of different types of galaxies.

Help Received

Online resources helped me understand how to code the graphs to show the performance of my neural network, and everything else was done on my own.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s) **Yuansong Wang**	**Project Number** **S0832**

Project Title

Analysis of ADHD among Students

Abstract

Objectives

Coded for Logistic Regression and Artificial Neuron Network(package) in R language and tested the ADHD model accuracy with ROC.

Methods

Laptop computer with R language and packages. Analyze the data of 1752 school students from National Health Interview Survey with Logistic Regression model and Artificial Neuron Network and checked the model accuracy with ROC.

Results

With coefficients' exact numbers of different factors, I can use the logistic regression model to predict a person's probability of getting ADHD, and checked the results with ROC, getting 66 and 69 percent accuracy for logistic regression and Artificial Neuron Network.

Conclusions

I coded for logistic regression model and Artificial Neuron Network to test the factors of ADHD, and checked the model accuracy with positive distribution and Receiving Operation Characteristics. With my code I can analyze large amount of data to create a predicting model for the issues that have multiple factors.

Summary Statement

I coded to analyze data with logistic regression model and Artificial Neuron Network and checked the accuracy of the two models with ROC.

Help Received

I programmed the logistic regression myself after an internet search on techniques, and the artificial neuron network and ROC are the packages pf R language tools.

CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)	Project Number
Dustin Wu	S0833

Project Title

A Novel Fuzzy Clustering-Based Convolutional Neural Network Ensemble for Histopathologic Cancer Metastasis Detection

Abstract

Objectives

The objective of the project is to develop deep learning algorithms that can efficiently and accurately detect cancer cells in digitized lymph node slide images.

Methods

Used laptop that accessed a GPU through an online server, a public dataset consisting of digitized images of lymph node histopathology slides with and without cancer, Python 3, and the Keras and Scikit-learn machine learning libraries to design the architecture of and train a convolutional neural network (CNN) to determine the presence of cancer in images. An ensemble (group) CNNs that cooperate to make more accurate predictions was created by partitioning the dataset into clusters using image feature extraction and a clustering algorithm, and then training a CNN on each cluster. Compared the use of non-overlapping clusters versus overlapping (fuzzy) clusters, and the use of handcrafted features versus features automatically selected by an autoencoder.

Results

The first CNN created contained three convolutional layers and had an area under the ROC curve (AUC) (a performance evaluation metric ranged from 0 to 1) of only 0.7432. Through the use of data augmentation, deeper architecture, more training cycles, and other improvements, the CNN model, now with six convolutional layers, was able to achieve an AUC of 0.9796. The model performance was improved by using a non-overlapping clustering ensemble method, and the fuzzy clustering ensemble method (FCM) boosted the performance even further. Lastly, the use automatic feature extraction did not significantly impact model performance. The final FCM model improved to an AUC of 0.9862 and overall accuracy of 95% on the reserved test dataset.

Conclusions

In this project, I incrementally developed a novel ensemble framework to more accurately detect cancer cells in digitized lymph node histopathology images. This work will not only be useful in helping pathologists detect cancer cells in histopathology images but can also be applied to other machine learning tasks.

Summary Statement

I developed a novel clustering-based deep learning framework that can efficiently and accurately detect cancer cells in digitized lymph node histopathology images.

Help Received

I independently utilized machine learning libraries to develop and optimize various models and algorithms. I received help related to machine learning from Dr. Christopher Hart and help with understanding cancer histopathology from Dr. Jeff Engelhardt, both at Ionis Pharmaceuticals, Inc.