Project ID: 231

Senior Division

Computer Science and Systems Software

Vasant Arul
Canyon Crest Academy
Gr. 12

Machine Learning Based Evaluation of Dehazing Algorithms

AWARDS:

CSEF Qualified

Autonomous driving vehicles are the new frontier in the development of transportation systems. Cameras and sensors are deployed to collect real world information which is then processed by many algorithms which perform tasks such as object detection and recognition. For these algorithms to function properly, they must be provided clear and high quality images, which introduces the importance of dehazing algorithms.

Dehazing algorithms act to remove "haze", such as dust, rain, smoke or fog, from a hazy image to turn it into one that is clear, significantly improving visibility for vision-based intelligence systems like self-driving or driving assistance. This important role has led to many dehazing algorithms being developed, but objective evaluation metrics have yet to be developed.

In this project, a mathematically-based objective dehazing quality index is proposed. An SVM Regression Model is used to predict the Dehaze Quality based on the features extracted from the dehazed image and its corresponding hazy image, and is compared to other machine learning models through cross validation. The SVM model has a good mean score of 0.95, and superior consistency, with standard deviation of 0.0039. The predicted DHQ Index for test images which are generated from the real hazy images and from other sources yields promising results. Additionally, we used consistency evaluation metrics, such as SRCC, PLCC, and RMSE, to compare the DHQI algorithm to existing methods of dehazing evaluation, with the results showing that DHQI has more consistent results and is less reliant on training data.

Project ID: 232

Senior Division

Computer Science and Systems Software

Alexander Backues
Canyon Crest Academy
Gr. 11

Improving PET Scan Resolution Using Machine Learning

AWARDS:

CSEF Qualified

The goal was to determine whether machine learning could be used to improve the accuracy of the reconstructed position of where gamma rays interacted with PET cells. Geant4 was used to simulate the generation of photons after a gamma ray produced from an electron-positron annihilation hit the anode in a PET cell. Data from Geant4 simulations were then fed into a machine learning model coded in Google Colaboratory, and the model was trained, tested, and graded based on accuracy. The model was repeatedly re-trained with varying amounts of data to find the optimal amount. It was found that sub-millimeter accuracy of the gamma ray position with detectors of size 1.5 cm x 1.5 cm x 0.2 cm was possible through the implementation of a machine learning model, and that positioning the anode 10 mm away from the photon detectors maximized the accuracy of the model. When the anode was 10 mm away from the photon detectors, the model's reconstructed position was, on average, within 0.6 mm of the actual location. The best result was achieved with around 1000 points of data. The contributions of this project are twofold. First, it showed that implementing machine learning into the process of determining gamma ray position was able to achieve sub-millimeter accuracy with detectors significantly bigger than 1 mm. Secondly, it also approximately determined the best anode position to maximize the accuracy of the machine learning model.

Project ID: 233

Senior Division

Computer Science and Systems Software

Arnav Dagar
Canyon Crest Academy
Gr. 10

ARFSNet - Deep Learning for Amblyopia Risk Factor Screening

AWARDS:

Grand Award - Senior Division Physical Sciences – ISEF FINALIST
Armed Forces Communications Electronics Association (AFCEA) Senior Division Winner
CSEF Qualified

Amblyopia or "lazy eye" is a childhood vision development disorder that affects 1-5% of the population. Amblyopia causes more vision loss in children than all the other causes combined. It can be treated by the age of 8-9 years when the visual pathways of the brain are still plastic. It is critical to screen for risk factors of amblyopia early for timely care.

Objective: Develop deep convolutional neural networks to predict some risk factors of amblyopia from images. This project focuses on refractive errors and strabismus prediction.

Method: GoCheck Kids provided 462 images of eyes taken using an iPhone based photoscreener. The associated cycloplegic evaluations of the patients (refraction errors, gaze type) served as ground truth. Only 150 images included strabismus prism diopters: I visualized and curated this data to remove duplicates, poor images, etc. I developed a mathematical model for strabismus and a python application to generate synthetic images from 25 real orthotropic images. I augmented the dataset through image flipping and random rotations. I designed, trained, and evaluated the CNN Amblyopia Risk Factor Screening Network (ARFSNet). ARFSNet contains two CNNs - Ocular Refractive Error Network (ORENet) and Ocular Gaze & Strabismus Network (OG&SNet). ORENet and OG&SNet provide quantitative results.

Results and Conclusions: I developed an algorithm for generating synthetic images for strabismus from mathematical eye models. Training with mostly synthetic images, OG&SNet can detect up to 5 PD of strabismus and ORENet can detect 1 diopter of refractive error in the real images. Hyperparameter tuning is continuing.

Project ID: 234

Senior Division

Computer Science and Systems Software

Nathan Dai

Canyon Crest Academy

Gr. 12

Performance Tradeoff in ML-Based Intrusion Detection Systems: Efficacy vs. Resource Usage

With the surge in IoT attacks and the increasing exploitation of IoT devices as entry points for data breaches, enhancing network security measures has become imperative. The resource constraints of IoT devices, however, pose significant challenges for executing computationally intensive ML tasks. To address these challenges, we assess the performance and resource usage of various ML models for network intrusion detection using the CICIDS-2017 dataset. Our research underscores the importance of balancing performance and resource efficiency when selecting ML models for IoT devices. We implement preprocessing steps such as standardization and Principal Component Analysis (PCA) for dimensionality reduction, facilitating the development of more resource-efficient ML models. To tackle the issue of an unbalanced dataset, we employ sample weighting. We measure the models' training time, inference time, and size as resource usage indicators. We conduct experiments using 30 cross-validation folds to ensure statistical significance. Our results reveal a trade-off between model performance and resource usage. Moreover, our findings suggest XGBoost and Decision Tree as suitable options due to their acceptable performance and moderate resource demands. In contrast, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) exhibit high resource consumption, while Naive Bayes delivers subpar performance. We also observe that rebalancing techniques can enhance recall at the cost of precision. The developed program fulfills the objectives by offering an efficient IDS solution for IoT devices with limited computational power. Unlike similar programs, this study concentrates on optimizing both performance and resource usage, delivering a comprehensive evaluation of ML-based IDS in IoT contexts.

Project ID: 235

Senior Division

Computer Science and Systems Software

Luke DeMott

La Jolla High School

Gr. 10

Investigating Weather Inputs for Neural Network Based Demand Side Forecasting in Home Energy Management Systems

The goal of my study is to determine if weather is a vital Neural Network (NN) parameter when optimizing energy efficiency in home energy management systems (HEMS). The increase in energy demand within homes has heightened need to help consumers reduce energy use and electricity cost using NNs within HEMSs. Procedurally I collected and reviewed NN data and findings to determine vitality of weather as a data set parameter for optimizing HEMS. Upon analysis, I found: 1. Consumption of energy gathered from solar panels can improve up to 8% and the amount of energy absorbed can be reduced by up to 25% when a NN is applied using weather information as an input [3]; 2. Accurate forecasts enhance reliability and reduce energy management costs by allowing efficient solar energy trading and secure distribution. Getting these accurate forecasts is not possible without weather as a parameter input; 3. NN based EMSs that did not include weather used less specific parameters, including historical energy consumption patterns and energy system loading, which led to less efficient energy demand prediction; and 4. Although reliable forecasting provides many benefits when managing energy distribution, NN algorithms must account for forecasting errors if network inputs require more than 24-hour advance notice. In comparing observed versus predicted 6-day forecast results from an Artificial NN focused on solar energy, a study found weather prediction error ranged between 5% at day one to almost 30% by day five [5]. In conclusion, information from weather forecasting plays a significant role in designing NNs to support HEMSs.

Project ID: 236

Senior Division

Computer Science and Systems Software

Ella Desai
Canyon Crest Academy
Gr. 11

Statistical Analysis for Deep Learning Model Selection

AWARDS:

CSEF Qualified

Problem: Deep learning enables anyone with an idea and basic coding skills to build applications that learn from data. Deep Neural Networks (DNNs) are universal function approximators and different types of models can be trained to fit the same data. It is of interest to investigate if the best performing model can be predicted beforehand via statistical analysis of the data. This would save on the time-consuming task of training and optimizing more than one DNN.

Hypothesis: A Convolutional Neural Network (CNN) will outperform other candidate models in anomaly detection from engine noise because of its ability to effectively detect patterns in sensor signals regardless of their position in the input sequence.

Procedure: Predict the best-performing of 3 DNN models {Long Short-Term Memory (LSTM), CNN, Transformer} from input data. Fine-tune each model's parameters and hyperparameters. Evaluate the binary classification accuracy of each model.

Results: A 1D CNN with batch normalization achieved the best training, validation, and test accuracy (~97%). The Transformer and LSTM architectures were less performant than the CNN, as correctly predicted from the dataset analysis.

Conclusion: The CNN effectively captured patterns from the motor sensor data by extracting its important features. The Power Spectral Density and Stationarity indicators correctly predicted that a CNN would achieve high accuracy. This is contrary to the conventional wisdom that an LSTM or Transformer is better suited for sequential data. Further work to extract additional meaningful patterns from other types of application data could lead to an automated DNN model selection recommendation procedure.

Project ID: 237

Senior Division

Computer Science and Systems Software

Aneesh Devulapalli
Canyon Crest Academy
Gr. 11

Immediate Diabetic Retinopathy Detection Using An Image-Based Machine Learning Model

This project aims to address the repercussions of late detection of diabetic retinopathy and the availability of such detection. By leveraging growing technological techniques of machine learning based image processing, those with diabetic retinopathy can be treated much earlier. According to CDC, more than 90% of vision loss caused by diabetes can be avoided with early detection and treatment. The objective was to develop a machine learning model capable of classifying 5 different levels of diabetic retinopathy while utilizing a light method.

The model used transfer learning to fine-tune a pre-trained model trained on the ImageNet dataset. In order to ensure accurate generalization of the model, it was trained on the Diabetic Retinopathy Detection dataset, including 35,000+ fundus images. Along with this, it employed filtering mechanisms, including gaussian blur, contrast, changing RGB values, and other transformations.

Previous works with ensemble techniques required a lot of computational power, rendering them inefficient for deployment on edge devices. With the use of transfer learning, filter mechanisms and hyperparameter searching, this model achieved a higher accuracy than previous models, while requiring less computational power. This enabled the operation of the model through a mobile app, making it easily accessible.

The model successfully achieved this goal, with a current test accuracy of 80.07%. This is significant as other CNN models with less computational power had a maximum test accuracy of 78.60%. With this improved accuracy, the project succeeded in making detection accurate and accessible with its low computational power.

Project ID: 238

Senior Division

Computer Science and Systems Software

Eric Feng

The Bishop's School

Gr. 9

Finding the Cause of Hallucinations Created by Generative Artificial Intelligence

This project uses prompts given to 2 different artificial intelligence engines, ChatGPT (OpenAI) and Gemini (Google), to determine the cause of hallucinations created by generative AI. The objective is to find what kinds of questions might influence an AI's response, and what factors might cause it to hallucinate. Different types of questions are given to attempt to provoke an inaccurate response from the AI to see what factors in the prompts given cause it to hallucinate.

For each part of the experiment, a set of questions are given separately to ChatGPT and Gemini to answer and their number of correct responses are recorded. The first and second part of the experiment consists of giving 10 trivia questions to each AI, while the third part of the experiment gives 2 critical thinking questions instead.

Overall, ChatGPT and Gemini had similar scores in each experiment. In the first two experiments, ChatGPT scored 9/10 both times and Gemini scored 10/10. The question they all missed was one that was stated vaguely and had multiple correct answers. One of the complicated riddles in experiment 3 was answered incorrectly.

The results of the experiment indicate that hallucinations found in generative AI are caused by the lack of clarity in the prompt given to the AI. Some wrong answers are also caused by the lack of deductive reasoning and logic in the AI's program. These results may help both AI users and programmers in the future to develop more effective technology with artificial intelligence.

Project ID: 239

Senior Division

Computer Science and Systems Software

Riya Khushu
Francis Parker School
Gr. 10

A Novel Gunshot Detection and Localization System for Public Safety

AWARDS:

Association for Women in Science - Winner
Society of American Military Engineers - San Diego Post - Honorable Mention
CSEF Qualified

Gun violence continues to cast a devastating shadow on communities worldwide. On January 17th, 2023, at my high school, panic struck when we were informed about a shooter on campus. This triggered pandemonium as students scattered in all directions, seeking what they believed to be safety. We didn't know if shots had been fired or the location of the shooter. Even though the incident turned out to be a false alarm, I wondered how chaotic and dangerous a real incident would be and if there is technology that can pave the way for better evacuation and managing the incident better.

This project explores the transformative potential of artificial intelligence (AI), particularly convolutional neural networks (CNNs), in revolutionizing gunshot detection and making it easily accessible to schools, public places and other locations that are vulnerable to gun violence. In addition, I explore localization methods that can be deployed in conjunction with gunshot detection to provide an approximate location of gunshots.

Using my computer hardware, microphones, open-source software, and python code, I built a CNN and trained it with custom data set from my Francis Parker High School. My method does not require any specialized hardware, sensors, access to cloud, or any third-party vendor. All the processing, including training with custom data, was done on my computer and without data ever leaving my computer.

After achieving an initial detection accuracy of ~90%, I used transfer learning to improve the accuracy to ~96%. In addition to detection, I measured the energy of sound at different locations and then used linear algebra to identify the location of gunshots with respect to a reference position of measurement.

Considering these unique features, including security and privacy, the solution can be deployed at schools, malls and other public places with minimal cost for deployment and maintenance. This AI-powered system holds immense promise by democratizing access to gunshot detection and localization across our communities.

Project ID: 240

Senior Division

Computer Science and Systems Software

Kaelyn Liu
Canyon Crest Academy
Gr. 10

Novel Machine Learning Approach to Prevent Misuse of Pharmaceutical Pills

AWARDS:

Association for Women in Science - Winner
Kaiser Permanente Blue Ribbon Award

Each year, millions of patients are negatively impacted as a result of preventable medication errors. Pill identification is error prone because of the thousands of FDA approved drugs and numerous look-alike pills. This task is especially challenging for individuals aged 65 and up who manage multiple prescriptions. Current pill identification and pill tracking applications lack the necessary robustness to effectively prevent the misuse of pills in a real-world setting. To address this issue, I developed a novel data augmentation pipeline to train a state-of-the-art YOLO deep learning model that could successfully identify different pills. I then utilized a Flask server to integrate this model into a mobile app I programmed with the Dart coding language. My final model achieved an mAP 50 score of 99.49% and an mAP 50-95 score of 95.42%, proving highly accurate even in real-world conditions. Additionally, my mobile app could successfully detect pills from user-captured images and help ensure correct pill intake. My mobile app's ability to display appropriate warnings based on the machine learning model's pill predictions has the potential to prevent pill misuse related hospitalizations. Moreover, my novel data augmentation pipeline could be used in other scenarios that lack a robust dataset. Future research will be conducted to enable my app to identify a broader range of medications.

Project ID: 241

Senior Division

Computer Science and Systems Software

Raunak Mondal
Del Norte High School
Gr. 11

Are LLMs Ready to Deliver Autism Inclusion?

This project focuses on reducing inherent biases resulting from a lack of context found in natural language processing models, particularly in the field of ableist hate speech detection. While previous literature has focused on improving performance of large language models (LLMs) in datasets related to hate speech, there hasn't been much work done focusing on autism, which is important because prior work has shown that speech detection has limitations for generalizations.

In order to gauge the performance of these models, a benchmark dataset is designed to serve as ground truth for anti-autistic hate speech detection. This project evaluates the performance of Gemini, Google's state of the art artificial intelligence model, for different methods of prompt engineering in perceiving and justifying their responses to ableist text. The statistical analysis focuses on a comparison of model accuracy and the reinforcement of anti-autistic beliefs, with a focus on computing probabilities in response to explicit and implicitly ableist prompts.

The project draws on references related to the development and labeling of benchmark datasets for hate speech identification, as well as literature on language model testing and dataset labeling.

The results of this project indicate an accuracy of 74.64% with the yes-or-no prompt and an accuracy of 32.57% with the prompt that required explanation. The results of this project can influence the understanding of biases in language models concerning autism and can especially influence the development of more inclusive and less ableist AI-powered chatbots.

Project ID: 242

Senior Division

Computer Science and Systems Software

Ronit Munshi

Canyon Crest Academy

Gr. 11

Making a Model That Can Predict Which Jobs Are Real and Which Jobs are Fake in the Most Efficient Method Possible

The main problem to be solved is to make a NLP Model that can detect the differences and correctly predict whether different job listings are real or fake. The procedure involved finding the dataset needed to train the model (the Real / Fake Job Posting Prediction dataset from Kaggle). From there, I preprocessed the data to make it fit for use on the NLP model I would create, primarily involving selecting which parts of the dataset I would use, cleaning up any unnecessary elements in the data, setting up a train test split with the part of the dataset I chose, and running the data through a count vectorizer to turn the text data into a numerical matrix of 1s and 0s. After the preprocessing, I tested out different NLP models to see which one would end up giving the best prediction, with the only one I got working in the end being a Logistic Regression Model. The accuracy of the model ended up being 99.48%, with the model correctly identified and predicted 2461 true jobs positions and 53 fake job positions, with there being only 4 incorrect true positions and 9 incorrect fake position predictions. And after running the data through Random Forest Hyperparameter Tuning, the model was left with a mean absolute error of 3.372721906660412% (low values are good!). Overall, the accuracy of the logistic regression model on its own was excellent in separating the fake job listings from the real ones, showing its capability in that regard. Though, possible potential ways to further improve the accuracy to a greater level would be through implementing a LSTM or BERT model with the same data, given how both can be more accurate than Logistic Regression. Though, both require their own amount of preprocessing of the data that is separate from the main preprocessing already completed.

Project ID: 243

Senior Division

Computer Science and Systems Software

Gavin Perez

The Cambridge School High School

Gr. 9

Program Complexifying Passwords in Ways Unique to a System

In this Project, I aimed to create a program that gets around password storage while still making complex and secure passwords easy to remember. With password storage systems, there are too many passwords stored in the same area, making them high reward targets for hackers. The idea of complexifying easy to remember password cues would get around saving passwords in one place while still letting them be secure and memorable. Despite this, there has been little study in that area, so I compared my program to the status quo: Password Storage Systems. My first sample run was nearly successful, waning in the security of capital letters. My second run was successful in the following: CIS Guidelines, different from other outputs, and different than the input.

Project ID: 244

Senior Division

Computer Science and Systems Software

Shounak Ray Chaudhuri
Francis Parker School
Gr. 10

Using Computer Vision for Object Detection/Classification on Roads of Developing Countries

AWARDS:

Armed Forces Communications Electronics Association (AFCEA) Senior Division Winner
Office of Naval Research - Senior Division Winner
CSEF Qualified

I developed an AI-powered computer vision model that can serve as a framework for autonomous vehicles (AVs) in developing countries, with an emphasis on Indian roads. The vast majority of AV developments have focused on the roads of developed countries, which face different challenges and situations, leaving the rest of the world behind. Roads in developing countries simply do not enjoy the same luxuries as those in developed ones, and it is clear that those technologies do not work for feasible autonomous vehicles in developing countries. The objective of this model was to be able to input dashcam video of a vehicle on Indian roads and get out annotated video featuring bounding boxes for more vehicle types, a prediction of where other vehicles will go, and the likelihood of the car being at an intersection or on a narrow road. The model I developed has been created from the ground up with datasets from developing countries with features in mind for the challenges that they face, including more types of vehicles, disorganized intersections, and the lack of lane driving. This project included data collection from dashcam recordings and formally annotated datasets (few and far between for developing countries) and training the models for each feature of the overall system (using Python, YOLOv8, OpenCV, and more), to finally bringing all the systems together so that dashcam footage fed in can come out fully annotated. A sample run of this model is shown below. The finished model can run with high accuracy at 55 FPS and is much better suited to Indian roads than previous developments.

Project ID: 245

Senior Division

Computer Science and Systems Software

Wonhee Ryu
Canyon Crest Academy
Gr. 10

Document Search Using Sentence Embedding

This project tested the effectiveness of sentence and word embeddings to find related text in documents given human-generated search queries. Two pre-trained models, SentenceTransformers, and Word2Vec, are used to test effectiveness of sentence embeddings for searching documents. The effectiveness of these are compared by using sentence block pairs. The final program is able to find similar sentences given queries and graded subjectively and using the tf-idf metric. It was hypothesized that the program would find a similar phrase within a document given a search query. The results indicate that the program is able to identify relatively similar sentences with common words and meaning with better results using word embeddings than sentence embeddings. This result is reflected in the tf-idf metric as the scores for the sentences found by the Word2Vec model are higher for every sentence than the SentenceTransformers model. When sentence block pairs within the text were compared, the two different methods indicated that even while comparing the same two sentences, most of the pair ratings were concentrated between 0.6 and 0.8 while for word embeddings they were concentrated between 0.8 and 0.9. This program was created as an improvement of the find tool and its ability to find parts of the document without exact wording would improve the convenience of document searching. The program showed better performance when the queries were similar size to the sentence blocks but struggled to find correlation between short queries and longer sentence blocks.

Project ID: 246

Senior Division

Computer Science and Systems Software

Parth Saxena
Scripps Ranch High School
Gr. 10

Streamlining Virtual Education Using Generative Artificial Intelligence and High-Performant Large Language Models

AWARDS:

CSEF Qualified

Generative Artificial Intelligence can be advantageous in virtual education in countless ways, however, there are setbacks to utilizing this technology, such as biases, hallucinations, and misinformation.

This program aimed to develop a Generative AI system using pre-existing large language models (LLMs) to develop educational content less susceptible to generating problematic content by providing users a pathway to upload their own material, which the LLM would learn from.

I integrated pai-001, an open-source LLM, with Flask for the server side of the platform, to develop an API that would interface with the front end, which was developed in HTML, CSS, and JavaScript. The system could now intake PDFs, TXTs, YouTube videos, and other forms of content, and collaborate with the model.

The system was tested on 3 datasets of various English, Biology, and History prompts. Source material was then provided to the model through the platform's content consumption features. The factor measured was the system's response accuracy, checking whether the generated information was factually correct. Each response was assessed on a scale of 0-10 on prompts spanning across 3 subjects and scored a 439/450 or a 97.6 percent accuracy.

This same procedure was used to test ChatGPT's 3.5 model, which scored 369/450, an accuracy of 82%. The key differentiating factor was that ChatGPT did not provide a robust pathway to providing source material. It went off the vast amounts of data it was trained on, inheriting the weaknesses of its training set.

Project ID: 247

Senior Division

Computer Science and Systems Software

Ethan Sun

The Bishop's School

Gr. 11

TrustworthyAI: Adversarial Attacks and Defensive Strategies in Self-Driving Systems Using Computer Vision and Artificial Intelligence

The project aims to improve the safety of autonomous vehicles against cyberattacks by developing a robust artificial intelligence system for precise stop sign detection. Its goal is to enhance the resilience of autonomous vehicles to cyber threats, contributing to the advancement of Trustworthy AI through the use of AI models.

The project involves employing and fine-tuning various iterations of the YOLO architecture, including YOLOv5mu, YOLOv8 from Roboflow, and YOLOv8s from Ultralytics. The YOLOv8 model from Roboflow emerges as the most effective. The development procedure includes calibrating the AI system to reliably detect stop signs, an essential aspect for the safe operation of autonomous vehicles.

A total of six different dataset versions were created, comprising over 46,535 images, featuring manually crafted simulations of cyberattacks. These simulations are designed to emulate real-world adversarial scenarios, allowing for a comprehensive assessment of the AI system's precision in detecting and classifying stop signs during such incidents.

Our model achieves up to 90% accuracy in classification confidence, which demonstrates the effectiveness of the YOLOv8 model and the comprehensive datasets in identifying and classifying stop signs, even under challenging conditions. The analysis underscores the critical role of computer vision techniques and extensive datasets in autonomous vehicle safety against cyberattacks.

This review emphasizes the success of the study in achieving its goals, showing how AI models can improve the security of autonomous vehicles against cyber threats. The innovative use of the YOLOv8 model and the extensive, targeted datasets underscores the project's contribution to Trustworthy AI and vehicle safety.

Project ID: 248

Senior Division

Computer Science and Systems Software

Olina Xia

Rancho Bernardo High School

Gr. 10

AI-Powered GIS-Based Mathematical Modeling Investigations on Climate Impacts of Transportation

Transportation is considered the largest greenhouse gas (GHG) contributor by economic sector. However, the extent of GHG emissions from urban freeway systems in our region remains subjects of debate. The predictive mathematical modeling might reveal that GHG produced by transportation is a function of various traffic parameters, such as daily roadway traffic volumes, speeds, and drivers.

The project commences with the comprehensive collection of urban freeway data in San Diego County, including transportation routes, count stations along the routes, and traffic volumes of each hour in an entire year at every count station. A unique aspect of this project is the application of Artificial Intelligent (AI) techniques to quickly predict the top 10 freeway segments in terms of the highest traffic and GHG emissions, based on the huge raw data sets. The prediction of AI-powered program was then being verified using the large-scale GIS-based mathematical models, creating a detailed spatial understanding of the Annual Average Daily Traffic (AADT) data and emissions of carbon dioxide, the main GHG, on San Diego's freeway systems.

The predictive mathematical modeling did reveal that GHG produced by transportation is closely correlated with daily roadway traffic volumes. However, when compared to the AADT freeway map, the CO2 emission map also showed that GHG emissions are not singly determined by traffic volumes. Another main factor could be traffic speed.

Employing AI-powered algorithms along with the large-scale regional transportation models provided a promising opportunity for transportation professionals to quickly identify inefficiencies and potential improvements in urban freeway systems.

Project ID: 249

Senior Division

Computer Science and Systems Software

Kayley Xu
The Bishop's School
Gr. 9

A Comprehensive Approach to Addressing Respiratory Disease

AWARDS:

Association for Women in Science - Winner
Kaiser Permanente Blue Ribbon Award
Society of Women Engineers - San Diego County Section - Senior Division 3rd Place
CSEF Qualified

This project stemmed from a personal experience with pneumonia and aims to address several challenges with auscultation interpretation that I both observed and experienced.

My project uses machine learning techniques to detect the presence of crackles and/or wheezes in digitally recorded respiratory audio samples. The audio samples were converted into spectrograms through Fourier transform, allowing image classification approaches to be applied. Techniques such as data augmentation, weighted random sampler, windowing, and frequency region cropping were applied, and a variety of machine learning networks were developed and trained. With an ensemble network consisting of seven best performing individual networks, 4-class (normal, crackle, wheeze, both) classification accuracy of 69.94% and sensitivity-specificity-average score of 68.69% are achieved. Separate machine learning networks were developed for respiratory disease diagnosis classification based on respiratory sounds and achieved 87.12% disease classification accuracy and 70.21% sensitivity-specificity-average.

I also designed and built a wireless digital stethoscope prototype that costs under $20. Lung sounds of multiple volunteers were collected using this prototype and were tested with the ML network, producing correct classification results.

This project brings an affordable and reliable solution to auscultation interpretation challenges such as inter-observer variability. Providing this solution to patients at home allows more breathing cycles to be factored into diagnosis and enables patients to conveniently perform self-examination.

In addition, a therapeutic breathing exercise web game was developed. With multiple features that encourage patients to practice breathing exercises, it provides a fun way to expedite recovery from respiratory disease.

Project ID: 250

Senior Division

Computer Science and Systems Software

Rex Tran

Scripps Ranch High School

Gr. 11

"If Humans Evolved, Machines Should Too": Novel Evolutionary-Merged Ensemble Multi-Layer Perceptron Model

Researchers have been working with neural networks since the 1950s; the starting architecture being a perceptron--a simple model bearing resemblance to a biological neuron. Fast forward, deep learning has advanced to CNNs and RNNs, vast strides from the past, but issues dating back to the perceptron still persist today. The brunt of the issues are overfitting and premature convergence. These two are at the heart of my research because when training NNs, A.I model waste considerable compute power trying to get out of local minima (premature convergence) or cannot be applied to testing data (overfitting) which is extremely taxing in real-life applications where unexpected situations run rampant. Several ensemble methods such as bagging and boosting have been created to reduce bias and variance but fail at fixing these two primary issues. I propose a novel evolutionary merged multi-layer perceptron (EMLP) that draws inspiration from Darwinian Evolution: natural selection, crossover, and mutation. By aggregating a population of models and evolving them over multiple generations, I can improve accuracy through selecting the fittest individuals (natural selection), I can incorporate model merging to lower variation (crossover), and I can streamline convergence by fostering diversity in the population. The EMLP model outperformed the baseline MLP model by 30.02% and 21.10% on Digit and Fashion MNIST datasets respectively. This novel ensembling configuration is imperative to the future as we continue to stride into an integrated A.I society where self-driving cars, disease diagnosis, and pharmaceutical drug discovery demand stronger neural networks.