Computer Science Clinic

Statement of Work for
Auburn / USDA

Time-series Modeling, Analysis, Interface, and Insight from Entomological Electropenetrography

April 29, 2025

Team Members

Mehrezat Abbas (Team Lead)
Milo Knell
Devanshi Guglani
Lillian Vernooy
Zachary Traul

Advisor

Prof. Gabriel Hope

Liaisons

Dr. Elaine Backus
Dr. Anastasia Cooper
Dr. Kathryn Reif

1 Introduction

Founded by President Lincoln, the United States Department of Agriculture (USDA) has also been called the "people's department" of the U.S. Government. The Agricultural Research Service (ARS) falls under the purview of the USDA's Research, Education, and Economics (REE) department. Its mission area is "to create a safe, sustainable, competitive, and equitable U.S. food and fiber system." [USDA] . The USDA, in partnership with the Department of Homeland Security created the National Bio and Agro-defense Facility (NBAF) to conduct researched aimed at protecting the nation's food supply and public health from zoonotic diseases (those afflicting both humans and animals) affecting livestock production [NBAF]. Pathogens transmitted by arthropods, such as Japanese encephalitis virus to livestock by mosquitoes, bean pod mottle virus to crops and anaplasmosis to livestock by ticks, are detrimental to agricultural production. In addition, many of the arthropod borne diseases affecting livestock are zoonotic and can affect public health as well.

The electropenetrograph (EPG) is a valuable tool for unraveling the transmission mechanisms of arthropod-borne pathogens. EPG involves passing a mild electrical current through a host, such as a plant or animal, and attaching a recording electrode component to an arthropod. Based on alterations in the electrical activity that occur throughout the arthropod's feeding patterns, we can glean information on their consumption behaviors. By visualizing and interpreting EPG signals, researchers gain vital insights into arthropod feeding behaviors, which are key to reducing the threats these pests pose to agricultural productivity.

Unfortunately, signal interpretation is incredibly time-consuming, taking up to months of labeling per recording session. This limits the utility of EPG for large-scale studies. Hence, the USDA and NBAF funded Dr. Reif's Parasitology lab at Auburn University (AU)'s College of Veterinary Medicine, a laboratory that specializes in EPG with blood-feeding arthropods, to form a specialized interdisciplinary team to develop an automated signal interpretation solution.

Thus, our goal is to reduce interpretation time on the part of entomologist researchers by automating the waveform labeling and identification process. This time reduction will increase the speed of EPG research and propel advancement towards pest mitigation and against agricultural and human losses due to arthropod-borne diseases.

2 Problem Statement

Manually labeling feeding behaviors on EPG waveforms is an unfortunately lengthy process for entomologists, including our liaisons at the AU and the USDA. As labeling waveforms takes away time that they could be spending on lab work, manuscript preparation, and other tasks, it also puts a bottleneck on research throughput. The importance of removing this bottleneck has risen in recent years due to changes in arthropod ranges from climate change that threaten agriculture in both the United States and worldwide. As EPG is a powerful tool for studying how parasitic arthropods interact with their hosts, improving the speed at which EPG data can be interpreted will be critical for understanding and combating threats to agriculture from existing and newly introduced pests.

Due to our funding specifications as laid out by the National Bio and Agro-Defense Facility and National Science Foundation, our project's focus will be on blood-sucking arthropods such as mosquitoes and ticks, which threaten both humans and livestock. Our aim is to use machine learning techniques to develop software that can largely automate the process of labeling EPG waveforms. As our sponsor's entomologists lack the technical expertise to do so themselves, they are seeking the assistance of Harvey Mudd College's Clinic team for development of this tool. In pursuit of this, our Clinic team aims to develop a machine learning model capable of labeling feeding phases and behaviors of bloodsucking arthropods in EPG waveforms and to create a software interface for its use by entomologists. If successful, our tool will improve EPG research output and help to modernize EPG for combating the current and future challenges faced by agriculture.

3 Project-Specific Background

The following sections describe EPG methodology, outputs, and the specific datasets that we will be working with in this project.

3.1 The Electropenetrograph

The electropenetrograph allows researchers to study arthropod feeding in cases where it is difficult to observe directly because the behaviors occur inside host tissues. To do this, the system uses the connection between the arthropod and their host to complete a circuit. A thin wire is attached to

the arthropod using conductive glue, and an electrode is placed in the soil (for plant-feeding arthropods) or on the skin of the host (for blood-feeding arthropods). This is placed in series with an amplifier, and a DC or AC voltage is applied to the circuit (Backus et al. (2019)).

As the arthropod feeds, the varying depth of the arthropod's mouthparts (e.g. stylets) in the host and flow of electrolytes through the mouthparts result in a time-varying resistance. In addition, small time-varying voltages are sometimes generated, such as when the stylets of a plant-feeding arthropod punctures a cell membrane. These are known in the community as the "R" and "emf" components of the signal, respectively. These signals result in a time-varying voltage at the input of the amplifier.

If an AC applied voltage is being used, the amplifier output is passed through a rectifier. While this is used to convert the AC signal to a DC signal before digital conversion, "rectifier foldover" can obscure voltage drops caused by varying biopotentials in the arthropod-plant system, and distort waveforms in arthropod-animal systems, if an appropriate DC offset is not included in the applied voltage. The signal is then converted into a digital signal and recorded using a computer. Currently, researchers mainly use the WinDAQ software to record, view, and annotate the generated signal.

The input impedance of the head stage amplifier can be varied between $10^6\,\Omega$ and $10^{13}\,\Omega$. As this is varied, the R and emf components of the signal are inversely represented in the final signal. In general, the best sensitivity tradeoff is obtained when the input impedance of the amplifier is similar to the impedance of the arthropod-host system.

Patterns in the output signal can be correlated with different phases of arthropod feeding. While a given species of arthropod has patterns in the signal which are characteristic of different phases of feeding, signals can be significantly different across different species, both in waveform shapes and in total recording duration. An example of the waveform generated by an EPG recording of a mosquito feed on a human hand is shown in Figure 1, and recording of a blue-green sharpshooter feeding on grape is shown in Figure 2.In agriculture, which is the primary focus of our sponsors, entomologists use these waveforms to study how agricultural pests interact with crops and livestock under different conditions.

3.2 Datasets

Our liaisons have provided us with two sets of data from two arthropod categories.

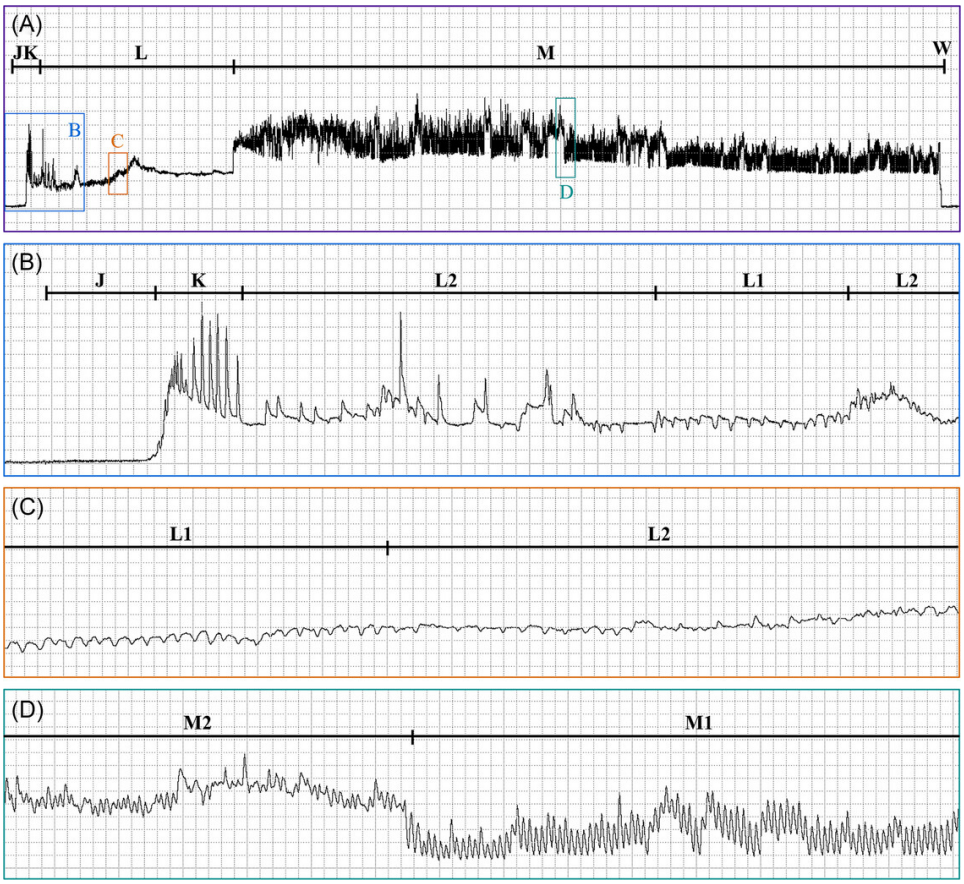

Figure 1 EPG waveform of mosquito feeding on human hand Cooper et al. (2024)

- Dr. Anastasia Cooper has provided WinDaq and text representations of waveforms collected from *Culex tarsalis* (mosquito) EPG recordings. Each file contains waveforms from one mosquito at a time, for one to five probes from each insect. The files represent 17.5 hours of recording time in total. The waveforms are labeled to the family level. We were also provided with a list of all possible behavior family transitions in mosquitoes.

  Advantages of this dataset include that it consists of data collected on the target species of our funder's and that it is composed of recordings from 62 individuals. Despite the number of recordings, the recordings themselves are rather short (on the order of minutes) and the

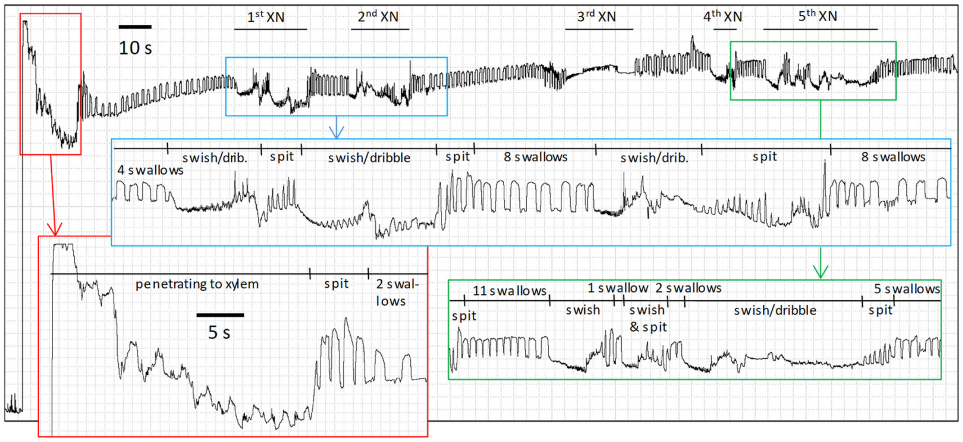

Figure 2 EPG waveform of blue-green sharpshooter feeding on grape Backus et al. (2021)

dataset is somewhat imbalanced, which is to say that there are less data points for certain types of waveforms like *J*, which only occurs for a few seconds in each file. Overcoming this imbalance when training models with be important for developing useful models.

- Dr. Elaine Backus has provided WinDaq, Excel, and CSV representations of waveforms collected from melon and cotton aphid EPG recordings. Because the duration of recorded waveforms for mosquitoes is limited, we will also work with an aphid dataset that has a greater depth of EPG recordings available. Working with the aphid dataset in parallel will allow us to fine tune aspects of programming the automated system that benefit from having more data. The aphid recordings are sectioned into five experiments and in total represent hundreds of hours of recording time. The files contain some labeling errors and varying waveform quality due to hardware, although these issues are insubstantial and we do not expect them to meaningfully detract from the usefulness of these data.

  Data volume is of course an advantage of this dataset. Additionally, aphid EPG waveforms are very well-characterized which makes them a useful reference point as they are the model species for EPG recordings. Although they are not the focus of our funders, we believe there is useful insight to be gained about automated labeling of EPG waveforms from this dataset. Aphids are also an agricultural pest so improvement in automation of their labeling would be a welcome

advancement in its own right.

4 Roadmap

The following sections detail our goals for the project and how we plan to achieve them.

4.1 Goals

The high-level goal of this project is to produce an automated system to label arthropod (with mosquitoes as the primary focus and ticks and biting midges as a secondary focus) EPG voltage data with waveform names. Because EPG has historically been used to study arthropods that feed on plants, there exist larger datasets for these arthropods. Therefore, towards continuing to refine aspects of the automated system that would benefit from having a deeper dataset that what is available for mosquitoes, we will also work in parallel on expanding the application of this automated system to include plant-feeding arthropods (with aphids as the primary focus and sharpshooters as a secondary focus). We will additionally create a user interface to make it possible for researchers without computer science experience to use this technology on their own data.

4.2 Objectives

- Define a metric for performance of our machine learning algorithm so we can compare its performance to the standards of our liaisons and the researchers who will use this tool.
- Create a machine-learning tool that automatically labels arthropod (in particular, mosquitoes and other bloodsucking arthropods) EPG voltage data according to the waveform.
- Stretch Goal: we hope to provide a secondary tool that can provide extra desired functionality:
  - Given labeled data for other arthropod species (e.g. aphid), re-train the machine learning algorithm to label novel data for the new arthropod species.
  - Given new data for the same arthropod species but on potentially different machinery or collected under different circumstances,

re-train the machine learning algorithm to label data under these different circumstances.

- This tool, including potential secondary tools for re-training under different circumstances, should be usable by an audience that does not have a computer science or machine learning background. This includes, but should not be limited to:
  - Labeling data with the model should take a reasonable amount of time on a desktop computer. In terms of reasonable, we mean that it should run for inference (making predictions) on the order of minutes, not hours or days. Re-training might potentially take longer but can be run in the background and is only a one-time cost for a new species or setup.
  - The interface and model should be compatible with standard desktop operating systems, in particular Mac and Windows.
  - The interface and model should be a stand-alone application. In particular, it should not rely on the software being produced concurrently with this project by our sponsor's Engineering Clinic team.
  - The interface and model should accept data in the form of comma separated value and WINDAQ files as these are the two most common formats in which waveform data has historically been recorded.
  - The interface should be easy to use by entomologists, particularly entomologists at Auburn/USDA. While evaluation of this objective is somewhat difficult, we will know when we have achieved it by feedback from our liaisons.
  - The interface and model should provide the user a metric of its confidence in each classification and flag uncertain areas for review which can then be classified by the user. This will require that our interface has the ability to display waveforms and labels and edit them.
  - The interface should have the ability to export labels and timestamps in a format compatible with existing tools used by our sponsors for further analysis.

4.2.1 Tasks

Time Series Classification Research First, we need to gain familiarity with machine learning algorithms for time series classification. We will begin our exploration with a two-pronged approach which will compare the performance of deep learning models which have yet to be applied to EPG data and existing EPG classification methods that have not yet been applied to our dataset.

On the deep learning side, we will begin by looking into pretrained deep learning models. This requires research into existing methods for classification of other waveform data such as EEG. There is a large body of existing work in time series classification using deep machine learning algorithms for EEG or other time series data, which primarily rely on convolutional network networks (CNNs) O'Shea and Nash (2015). They often include more complex models as well, including encoder/decoder with CNNs Ma et al. (2023) Kashiparekh et al. (2019) Craik et al. (2019). This is a promising starting point for our EPG data, since these datasets look similar (particularly the EEG data), even though the underlying data is slightly different. We could use one of these models and fine-tune it on our EPG data which will help us avoid overfitting with our limited data. Our lack of data is a major concern when training a deep model, so while transfer learning is one way to fix this, another is to simply use a less complex model.

For this, we reach towards more traditional machine learning approaches. Existing methods for classifying EPG waveforms largely work by using feature engineering (usually in the frequency domain) combined with some form of decision trees or neural networks to classify short intervals of the waveform as seen in Willett et al. (2016), Dinh et al. (2024), and Xing et al. (2023). While the performance of these models varied by the species of the arthropod being studied, these papers generally reported an accuracy greater than 85% in classifying ten second intervals of waveforms. While this accuracy may not necessarily be representative of the usefulness of these models, it is certainly an indication that they may be a good starting point. Additionally, as these models are less flexible than deep models, they may achieve better performance in a data limited environment. As labeling EPG waveforms is arduous, this could be an advantage of taking a less flexible approach. On the other hand, lack of flexibility may be the difference between 90% and 99% accuracy, which is why we also need to consider deep models.

Experimentation Next, we will produce experimental machine learning algorithms, based on the starting points described above or other methods we discover from our research to this point. We will measure the classification performance of these algorithms on the *Culex tarsalis* (mosquito) waveform library provided by Dr. Cooper. Further information on this library is available in Cooper et al. (2023). Besides evaluating classification performance using multiple metrics (discussion below), we will also evaluate how well they meet other objectives like ease of re-training, inference running time, and training data size requirements. The results of these experimental models will be the basis for our selection of a final model to move forward with.

Selecting a Performance Metric In previous work such as Willett et al. (2016), the usual performance metric is classification accuracy on ten second segments of the waveform. While this will certainly be a starting point for measuring performance, it does not necessarily mean that it will be the only metric we will evaluate. For example, perhaps shorter segments would be better at capturing short waveforms. Another factor to include in a performance metric could include the number of impossible transitions, where an impossible transition is one that is not biologically possible (like changing from feeding to non-feeding behavior without an intermediate state). Finally, it will likely be important to evaluate accuracy across states (easily visualized with a confusion matrix) as opposed to overall accuracy as this will allow us to ensure that the machine learning model is well-rounded, especially if the dataset is unbalanced label-wise. Of course, the comparison and choice of performance metrics will need to be a two-way conversation between the Clinic team and the liaisons to come to an agreement on what features of a performance metric are desirable.

User Interface Finally, we will create a graphical user interface (GUI) that allows users to apply the chosen machine learning algorithm to data in an easy and user-friendly way.

To create the interface we intend on utilizing the Qt framework. As Qt has bindings that allow us to create GUIs in Python, it will provide for programming language consistency within our project and smooth integration of the GUI and our machine learning models.

Some key metrics to determine the success of our GUI will include the following:

1. Task completion time: how long it takes a user to perform a particular

task: both in terms of how long it takes them to navigate and operate the system and for the interface to respond and process their tasks.

2. Perceived ease of use and user satisfaction: we can conduct A/B testing to determine the qualitative nature of how simple and intuitive our interface is.

3. Customer effort score: this measures how much effort our users need to interact with a product. This includes how easily they can resolve issues or utilize features. We should intend to make a guide or preliminary tutorial on how to operate our interface and run models.

   We intend to do user testing to determine our success in the above metrics when we present our interface to the entomologists.

4.3 Additional Details

This is a difficult task, with limited and highly noisy data. We can produce a predictive algorithm with some level of performance, but we cannot guarantee that its level of performance will be sufficient to label data automatically. It's possible its predictions are too noisy to use for practical research.

5 Schedule

5.1 Deliverables

Table 1 shows our current list of deliverables. We plan to create more deliverables, particularly for Spring semester (January to May), depending on what our research concludes at the end of Fall semester (September to December).

Deliverable	Date
Statement of Work	October 1, 2024
Decision of Machine Learning Approach	October 2024
Fall Presentation	October 22 or 29, 2024
Initial Prototype of Identification Model	December 2024
Initial Prototype of Interface	December 2024
Midyear Update	December 6, 2024
Second Prototype of Identification Model	February 2025
Feature Freeze	April 11, 2025
Final Identification Model	April 2025
Final Interface	April 2025
Code Freeze	April 25, 2025
Poster	April 29, 2025
Final Report	May 9, 2025
Project Handoff	May 9, 2025

Table 1 Schedule of Deliverables

Our deliverables are broken down in the following manner:

1. Decision of Machine Learning Approach: After a thorough data and strategy exploration phase, we will present a decision and rationale for our suggested approach to solving this machine learning problem according to our liaisons' requirements and boundaries. This will be our first major step in the technical development of the project.

2. Initial Prototype of Classification Model: We will present a prototype of our machine learning waveform identification model. This will include a locally usable model trained on a subset of the datasets given to us by our liaisons. The model is expected to have sub-par accuracy at this stage. With this prototype we will present a list of obstacles to success whose removal we deem possible to our liaisons.

3. Initial Prototype of Interface: We will present a prototype of the executable interface through which users will access our machine learning model. We will gather feedback on the options and visual aspect of the interface for refinement from our liaisons. The interface is a less intensive portion of the project but its correctness is extremely important to the success of our project's users. This interface will be used for testing during our site visit.

4. Second Prototype of Classification Model: Based on feedback from our liaisons and new solutions to accuracy shortcomings, we will present a second prototype of the identification model. We hope for this model to have an accuracy close to our defined measure of success, but continue to expect some problematic barriers and model failures.

5. Final Classification Model: We will complete the final iteration of the waveform identification model with a classification accuracy as high as we are capable of producing, with completed improvements determined after the second prototype presentation.

6. Final Interface: As the method of accessing our programs, this will be the final output of the project to our liaisons. The interface will be fit to researcher needs as determined during the initial prototype presentation and will function as a desktop application.

6 Conclusion

Auburn University and the USDA fund us in a movement not only to safeguard our massive agriculture industry against pest damages, but also to defend human lives against dangerous arthropod-borne pathogens. By successfully completing the development of an automated EPG waveform identification model, we will remove months of human work time from EPG-utilizing research. This will enhance the speed of advancement in pest management and help to prevent millions of dollars in agricultural production damages. With this project, we hope to contribute to the nutritional and financial security of communities across the country.

References

Backus, Elaine A, Felix A Cervantes, Raul Narciso C Guedes, Andrew Y Li, and Astri C Wayadande. 2019. AC–DC Electropenetrography for In-depth Studies of Feeding and Oviposition Behaviors. *Annals of the Entomological Society of America* 112(3):236–248. doi:10.1093/aesa/saz009. URL https://doi.org/10.1093/aesa/saz009. https://academic.oup.com/aesa/article-pdf/112/3/236/28565976/saz009.pdf.

Backus, Elaine A, Raul Narciso C Guedes, and Kathryn E Reif. 2021. Ac–dc electropenetrography: fundamentals, controversies, and perspectives for arthropod pest management. *Pest Management Science* 77(3):1132–1149. doi:https://doi.org/10.1002/ps.6087. URL https://scijournals.onlinelibrary.wiley.com/doi/abs/10.1002/ps.6087. https://scijournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ps.6087.

Cooper, Anastasia M. W., Samuel B. Jameson, Victoria Pickens, Cameron Osborne, Elaine A. Backus, Kristopher Silver, and Dana N. Mitzel. 2023. An electropenetrography waveform library for the probing and ingestion behaviors of culex tarsalis on human hands. *Insect Science* 31(4):1165–1186. doi:10.1111/1744-7917.13292. URL http://dx.doi.org/10.1111/1744-7917.13292.

———. 2024. An electropenetrography waveform library for the probing and ingestion behaviors of culex tarsalis on human hands. *Insect Science* 31(4):1165–1186. doi:https://doi.org/10.1111/1744-7917.13292. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1744-7917.13292. https://onlinelibrary.wiley.com/doi/pdf/10.1111/1744-7917.13292.

Craik, Alexander, Yongtian He, and Jose L Contreras-Vidal. 2019. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of Neural Engineering* 16(3):031,001. doi:10.1088/1741-2552/ab0ab5. URL https://dx.doi.org/10.1088/1741-2552/ab0ab5.

Dinh, Quang Dung, Daniel Kunk, Truong Son Hy, Nalam Vamsi, and Phuong D. Dao. 2024. Machine learning for characterizing plant-insect interactions through electrical penetration graphic signal doi:10.1101/2024.06.10.598170. URL http://dx.doi.org/10.1101/2024.06.10.598170.

Kashiparekh, Kathan, Jyoti Narwariya, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff. 2019. Convtimenet: A pre-trained deep convolutional

neural network for time series classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. doi:10.1109/IJCNN.2019.8852105.

Ma, Qianli, Zhen Liu, Zhenjing Zheng, Ziyang Huang, Siying Zhu, Zhongzhong Yu, and James T. Kwok. 2023. A survey on time-series pre-trained models URL https://arxiv.org/abs/2305.10716. 2305.10716.

Mirsky, Yisroel, Tomer Doitshman, Yuval Elovici, and Asaf Shabtai. 2018. Kitsune: An ensemble of autoencoders for online network intrusion detection. *CoRR* abs/1802.09089. URL http://arxiv.org/abs/1802.09089. 1802.09089.

O'Shea, Keiron, and Ryan Nash. 2015. An introduction to convolutional neural networks. URL https://arxiv.org/abs/1511.08458. 1511.08458.

Willett, Denis S., Justin George, Nora S. Willett, Lukasz L. Stelinski, and Stephen L. Lapointe. 2016. Machine learning for characterization of insect vector feeding. *PLOS Computational Biology* 12(11):e1005,158. doi:10.1371/journal.pcbi.1005158. URL http://dx.doi.org/10.1371/journal.pcbi.1005158.

Xing, Yuqing, Baofang Li, Lili Wu, and Fengming Yan. 2023. Waveforms eavesdropping prevention framework: The case of classification of epg waveforms of aphid utilizing wavelet kernel extreme learning machine. *Applied Artificial Intelligence* 37(1). doi:10.1080/08839514.2023.2214766. URL http://dx.doi.org/10.1080/08839514.2023.2214766.

Funding Acknowledgement

USDA (58-2034-3-445)
USDA (58-3022-4-034)
NSF (DBI - 2304787)

Collaborators

Harvey Mudd College
USDA-ARS San Joaquin Valley Agric. Sci. Ctr.
Auburn University College of Veterinary Medicine