Automating Earthquake Field Data Parsing with Machine Learning: From Free-Text to Structured Observations

Neeraja Vasa; Harini Pootheri; Edric Pauk; Tran T. Huynh; Luke Blair; Kate Thomas; Timothy Dawson

University of Southern California, Statewide California Earthquake Center, University of Utah

Abstract

Spatial data collected from the field after earthquakes is heterogeneous and requires extensive manual post-processing before publication. The field observation dataset from the 2014 Napa earthquake took five years to publish due to paper-based collection, while data from the 2019 Ridgecrest earthquake took one year using form-based mobile apps. However, significant amounts of data were still received in non-standardized formats, creating opportunities for automated parsing to further reduce post processing times.

Parsing and standardizing observation data involves manually interpreting various terminologies, unit conversions, and free-text field notes across multiple input types. Scientists at the USGS and CGS undertook this manual process for the fault rupture observation datasets from Napa and Ridgecrest earthquakes. Using these datasets, we trained a machine learning (ML) model to parse and extract data from free-text fields, and classify it into structured fields.

Our model achieved an average accuracy of 88% in extracting structured data from free-text notes for the Napa and Ridgecrest datasets combined. This approach demonstrates potential to reduce earthquake field data processing from years to months. Future work will expand beyond fault rupture free-text parsing to handle other hazard types and additional data formats.

Introduction

After a significant earthquake takes place, vast amounts of field observation data are generated to document ground deformation. While these data aids in providing insights for future earthquake responses, its dissemination is delayed due to manual post-processing procedures.

This project addressed a key bottleneck in observation data processing: the manual parsing of free-text field notes. Investigators employ several methods to record observations, including spreadsheets, PDFs, KMZ files, and ESRI shapefiles, with information organized using varied approaches. By targeting free-text fields, where investigators document observations in their own terminology (e.g. vertical slip vs. uplift vs. vertical displacement), we addressed the most challenging aspect of data standardization, as descriptive language, units, use of acronyms, and technical terms vary among investigators due to the short timeframe available to record perishable data. Using machine learning techniques trained on the Napa and Ridgecrest observation datasets, we developed a semi-automated system to parse these free-text fields and extract structured information about fault ruptures.

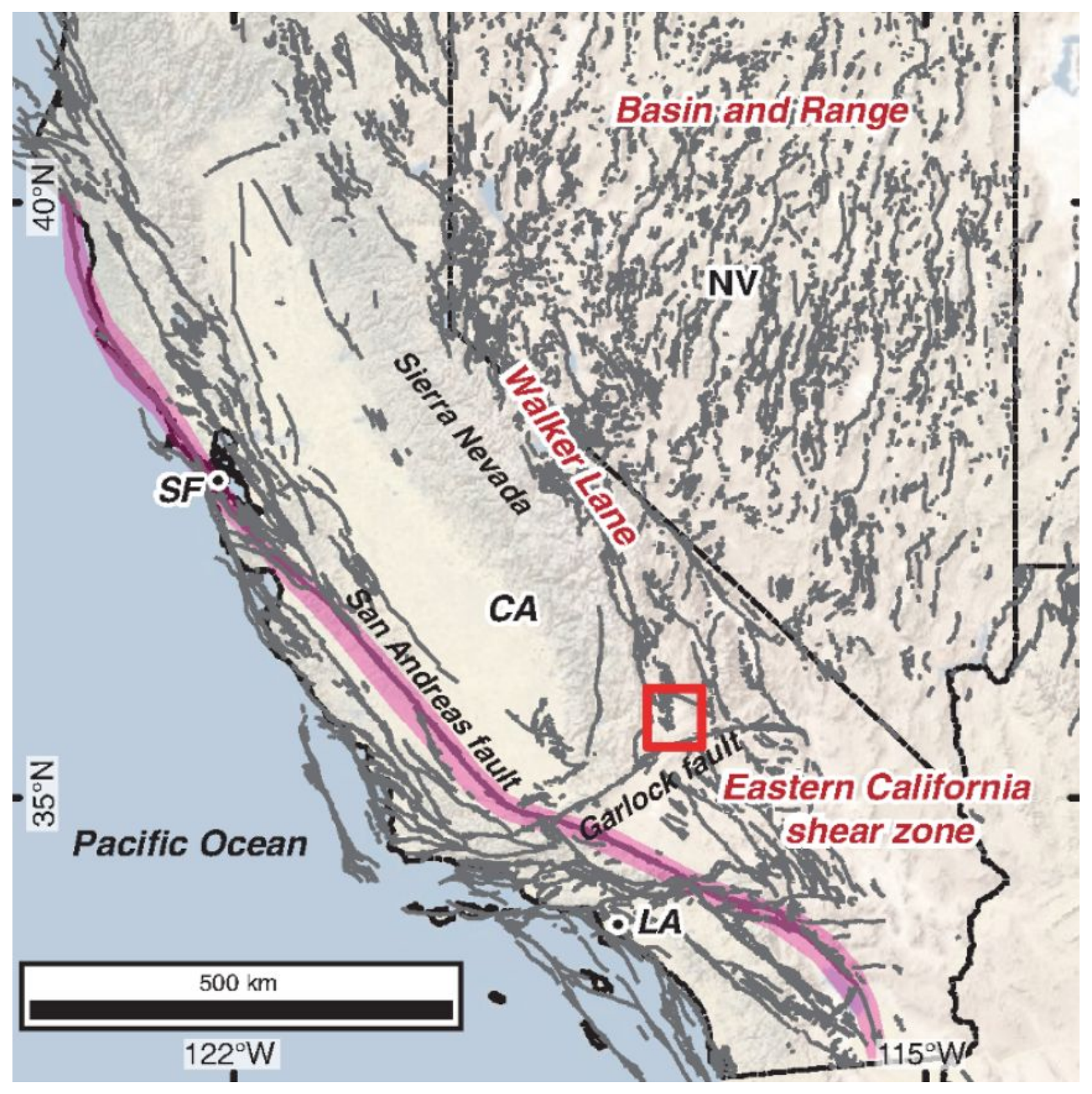
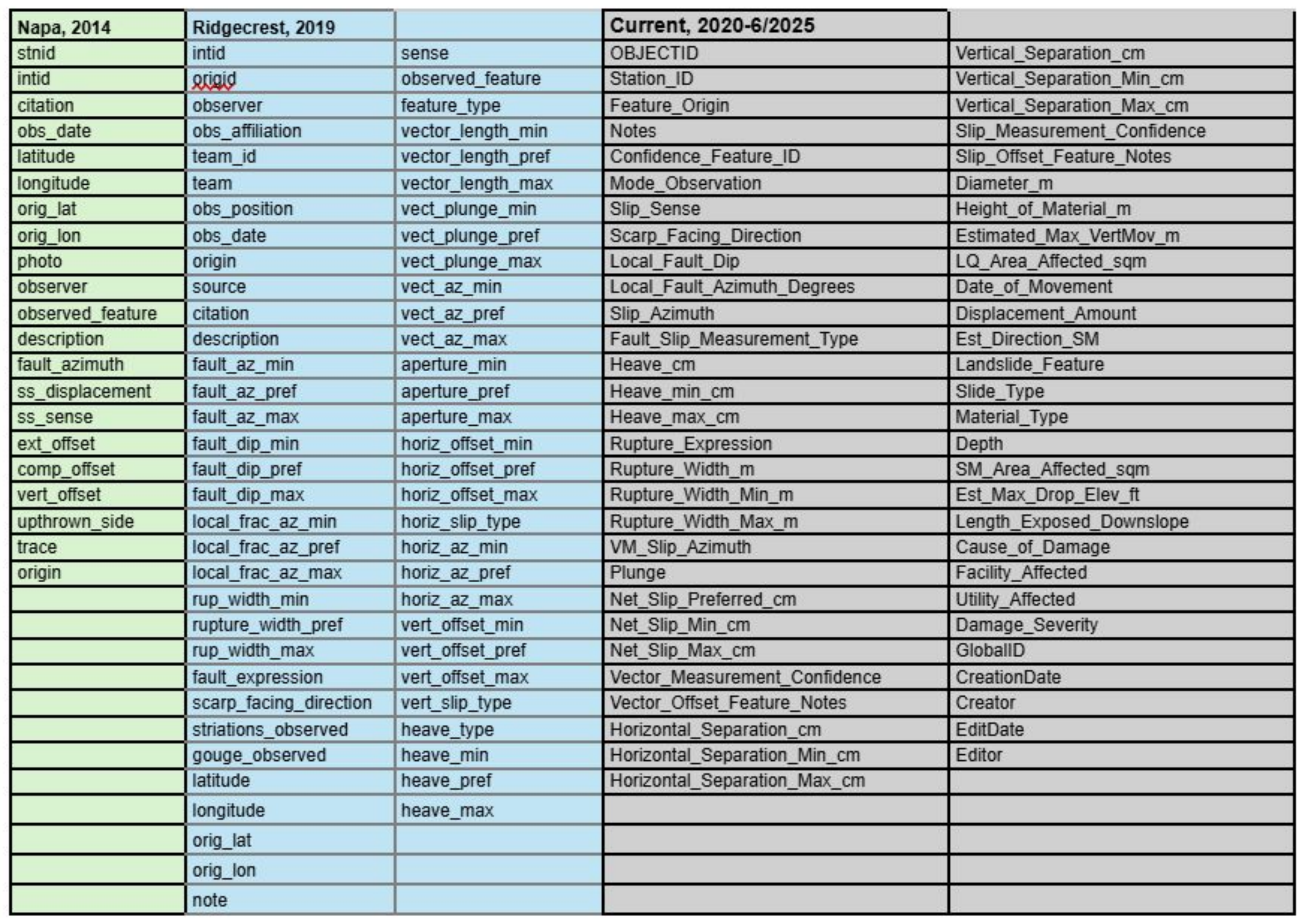


Figure 1: Tectonic setting of the Ridgecrest earthquake sequence in southern California. Red box shows earthquake locations; gray lines show Quaternary active faults from QFFD. (from Jobe et al., 2020)

Methodology

We began by consolidating post-earthquake fault rupture observation datasets from the 2014 Napa and 2019 Ridgecrest earthquakes into a unified dataset. The schema of this unified dataset is the most current schema that grew from incremental changes to what kinds of information was valuable enough to track. By comparing the fields between these datasets and the current schema, we systematically mapped corresponding fields with identical meanings. To preserve the original information from the Napa and Ridgecrest datasets that didn't fit in the current schema, we extended the current schema with new columns marked by underscore prefixes. This mapping strategy ensured comprehensive data retention while creating a larger, more robust training dataset. Our field mappings were the result of a collaborative effort involving experts Luke Blair (USGS), Kate Thomas (CGS), and Tim Dawson (CGS).



Napa, 2014	Ridgecrest, 2019		Current, 2020-6/2025	
stnid	intid	sense	OBJECTID	Vertical_Separation_cm
intid	origid	observed_feature	Station_ID	Vertical_Separation_Min_cm
citation	observer	feature_type	Feature_Origin	Vertical_Separation_Max_cm
obs_date	obs_affiliation	vector_length_min	Notes	Slip_Measurement_Confidence
latitude	team_id	vector_length_pref	Confidence_Feature_ID	Slip_Offset_Feature_Notes
longitude	team	vector_length_max	Mode_Observation	Diameter_m
orig_lat	obs_position	vect_plunge_min	Slip_Sense	Height_of_Material_m
orig_lon	obs_date	vect_plunge_pref	Scarp_Facing_Direction	Estimated_Max_VertMov_m
photo	origin	vect_plunge_max	Local_Fault_Dip	LQ_Area_Affected_sqm
observer	source	vect_az_min	Local_Fault_Azimuth_Degrees	Date_of_Movement
observed_feature	citation	vect_az_pref	Slip_Azimuth	Displacement_Amount
description	description	vect_az_max	Fault_Slip_Measurement_Type	Est_Direction_SM
fault_azimuth	fault_az_min	aperture_min	Heave_cm	Landslide_Feature
ss_displacement	fault_az_pref	aperture_pref	Heave_min_cm	Slide_Type
ss_sense	fault_az_max	aperture_max	Heave_max_cm	Material_Type
ext_offset	fault_dip_min	horiz_offset_min	Rupture_Expression	Depth
comp_offset	fault_dip_pref	horiz_offset_pref	Rupture_Width_m	SM_Area_Affected_sqm
vert_offset	fault_dip_max	horiz_offset_max	Rupture_Width_Min_m	Est_Max_Drop_Elev_ft
upthrown_side	local_frac_az_min	horiz_slip_type	Rupture_Width_Max_m	Length_Exposed_Downslope
trace	local_frac_az_pref	horiz_az_min	VM_Slip_Azimuth	Cause_of_Damage
origin	local_frac_az_max	horiz_az_pref	Plunge	Facility_Affected
	rup_width_min	horiz_az_max	Net_Slip_Preferred_cm	Utility_Affected
	rupture_width_pref	vert_offset_min	Net_Slip_Min_cm	Damage_Severity
	rup_width_max	vert_offset_pref	Net_Slip_Max_cm	GlobalID
	fault_expression	vert_offset_max	Vector_Measurement_Confidence	CreationDate
	scarp_facing_direction	vert_slip_type	Vector_Offset_Feature_Notes	Creator
	striations_observed	heave_type	Horizontal_Separation_cm	EditDate
	gouge_observed	heave_min	Horizontal_Separation_Min_cm	Editor
	latitude	heave_pref	Horizontal_Separation_Max_cm	
	longitude	heave_max		
	orig_lat			
	orig_lon			
	note			

Figure 2: The columns of the Napa and Ridgecrest field observation datasets as well as the columns of the current schema. The columns of these datasets were mapped to the columns of the current schema with the same meaning.

With validated field mappings in hand, we developed a Python script to transform the datasets into our unified dataset. The pipeline loaded the Napa and Ridgecrest datasets into Pandas DataFrames - chosen for their ability to handle mixed data types and seamless integration with machine learning libraries. Using our field mappings as a guide, the Python script populated corresponding schema fields, producing a consolidated dataset of 2,352 observations. This automated approach eliminated manual data transformation errors while ensuring consistent formatting across all observations.

Finally, we developed supervised learning models to extract geological characteristics from free-text field descriptions. The system converted unstructured field notes into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, then trained classification models to predict specific geological attributes like fault orientation, slip sense, and rupture expression. To facilitate rapid model development, we created artificial categorical fields by binning continuous measurements (e.g., converting precise slip values in centimeters to "Small/Medium/Large" categories). However, this approach sacrificed measurement precision for implementation speed.

Results

We used standard machine learning validation techniques to assess how accurately the system could automate the conversion of descriptive field text into structured data fields, replicating the structured data entry process that investigators typically perform manually.

We evaluated model performance using an 80/20 random train-test split of the consolidated earthquake field observations dataset. Our machine learning approach achieved an overall average accuracy of 88% across all predicted geological fields.

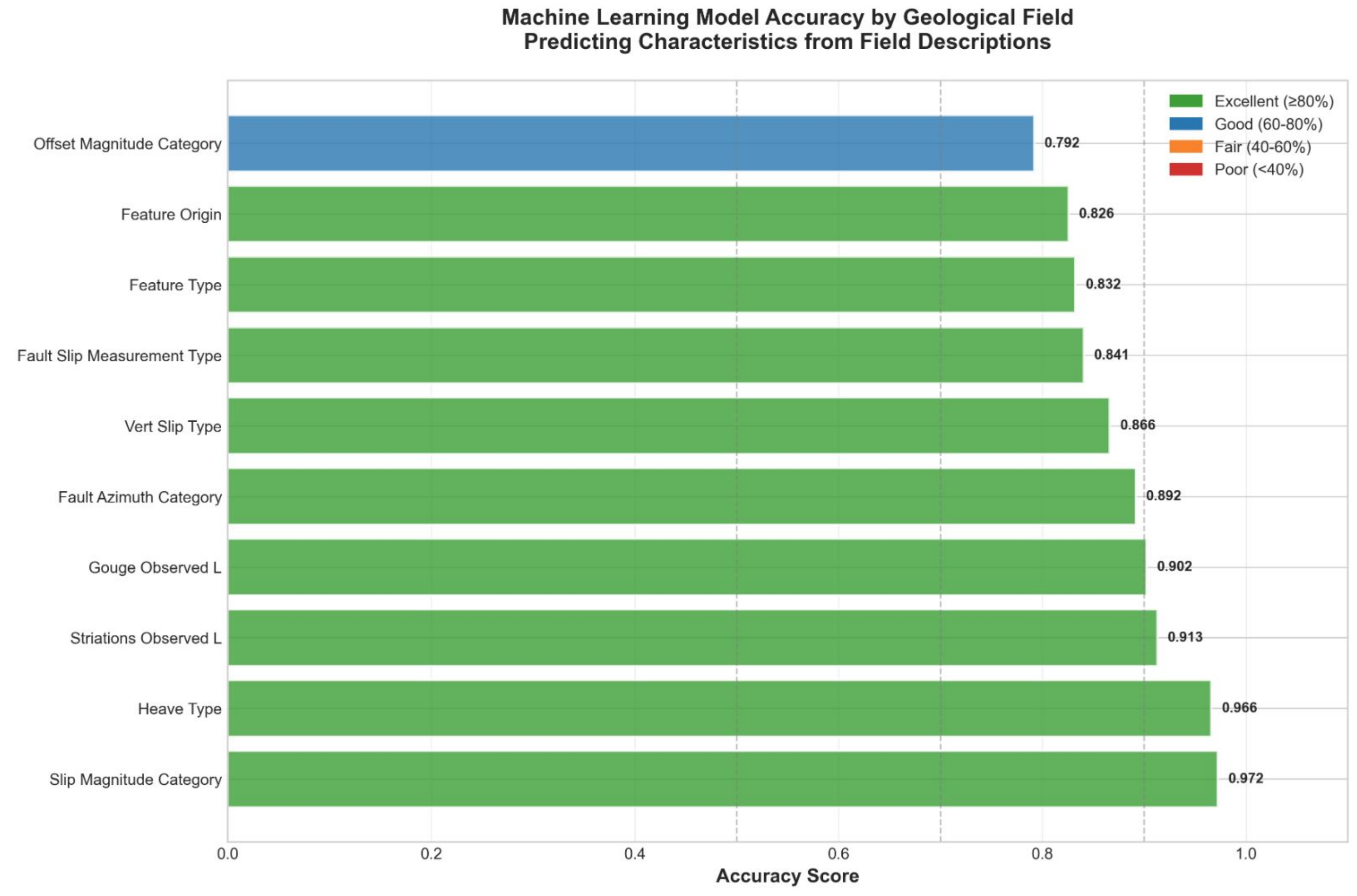


Figure 3: Accuracy of ML model predictions for each category

The model exhibited varying performance across different geological parameters. Binary classification tasks showed strong performance when accounting for class imbalance. For 'Gouge Observed' and 'Striations Observed', the model successfully identified the rare 'Yes' and 'No' cases despite the large number of 'Unknown' values in the training data.

One challenge was the prevalence of 'Unknown' classifications across most target fields. For instance, the 'Slip Magnitude' field, despite achieving the highest accuracy (97.2%), was dominated by 'Unknown' values.

These results demonstrate the model's effectiveness for automated processing of earthquake field data, particularly.. The high accuracy rates suggest potential for real-time field data standardization and quality control applications.

Conclusion

This project demonstrates that machine learning can significantly reduce the time it takes to pre-process data, thus reducing the time it takes to make the data available for use. The ML model in this project can be trained on other datasets to ensure that it yields more accurate predictions. The techniques used to create this ML model can be modified and used to predict data from liquefaction observations, not just fault rupture observations. Future work will focus on implementing direct numeric prediction for continuous measurements (e.g., precise slip values in centimeters) to preserve full measurement precision rather than using categorical bins.

References

1. Ponti, Daniel J., et al. "Documentation of surface fault rupture and ground-deformation features produced by the 4 and 5 July 2019 Mw 6.4 and Mw 7.1 Ridgecrest earthquake sequence." Seismological Research Letters 91.5 (2020): 2942-2959.
2. Ponti, D.J., Rosa, C.M., and Blair, J.L., 2019, Digital datasets documenting fault rupture and ground deformation features produced by the Mw 6.0 South Napa Earthquake of August 24, 2014: U.S. Geological Survey data release, https://doi.org/10.5066/F7P26W84
3. Ponti, D.J., Blair, J.L., Rosa, C.M., Thomas, K., Pickering, A.J., Dawson, T., Akciz, S., Angster, S., Bacon, S., Barth, N., Bennett, S., Blake, K., Bork, S., Bormann, J., Brooks, B., Bullard, T., Burgess, P., DeFrisco, M., Delano, J., Dolan, J., DuRoss, C., Ericksen, T., Frost, E., Funning, G., Gold, R., Graehl, N., Gutierrez, C., Haddon, E., Hatem, A., Hernandez, J., Hitchcock, C., Holland, P., Hudnut, K., Kendrick, K., Koehler, R., Kozaci, O., Ladinsky, T., Madugo, C., Mareschal, M., McPhillips, D., Milliner, C., Morelan, A., Nevitt, J., Olson, B., O'Neal, M., Padilla, S., Patton, J., Philibosian, B., Pierce, I., Sandwell, D., Scharer, K., Seitz, G., Singleton, D., Spangler, E., Swanson, B., Jobe, J.T. , Treiman, J., Valencia, F., Williams, A., Zachariasen, J., and Zinke, R., 2020, Field observations with quantitative displacement measurements obtained From surfaces faulting and ground deformation features produced by the Ridgecrest M6.4 and M7.1 earthquake sequence of July 4 and 5, 2019: Provisional release 1, in Ponti, D.J., Blair, J.L., Rosa, C.M., Thomas, K., Pickering, A.J., Dawson, T. E., compilers, 2020, Digital datasets documenting surface fault rupture and ground deformation features produced by the Ridgecrest M6.4 and M7.1 earthquake sequence of July 4 and 5, 2019: U.S. Geological Survey data release, https://doi.org/10.5066/P9BZ5IJ9.

Take a picture to see the code repository

U.S. National Science Foundation

Acknowledgements

This research was supported by the Statewide California Earthquake Center (Contribution No. 14487). SCEC is funded by NSF Cooperative Agreement EAR-2225216 and USGS Cooperative Agreement G24AC00072-00.