SEMESTER PROJECT 6501

SAKET UPADHYAY

FUZZDISTILL

FuzzDistill: Intelligent Fuzzing Target Selection using Compile-Time Analysis and Machine Learning

Saket Upadhyay
Ph.D. Student
Dept. of Computer Science
University of Virginia
saket@virginia.edu

***Abstract*—Fuzz testing is a fundamental technique employed to identify vulnerabilities within software systems. However, the process can be protracted and resource-intensive, especially when confronted with extensive codebases. In this work, I present FuzzDistill, an approach that harnesses compile-time data and machine learning to refine fuzzing targets. By analyzing compile-time information, such as function call graphs' features, loop information, and memory operations, FuzzDistill identifies high-priority areas of the codebase that are more probable to contain vulnerabilities. I demonstrate the efficacy of my approach through experiments conducted on real-world software, demon-**

II. Background

Fuzz testing is a widely employed technique for identifying vulnerabilities in software by providing invalid, unexpected, or random data inputs to a program in an attempt to elicit errors or crashes. While fuzzing has demonstrated efficacy in uncovering bugs, the extensive volume of code in contemporary software systems and the inherent unpredictability of the outcomes render it challenging to apply fuzz testing in a time-efficient and comprehensive manner.

AGENDA

- **WHY** we need it?
- **WHAT** it does?
- **HOW** does it work?

WHY

TARGET DISTILLATION IN LARGE PROJECTS

- Chromium as approx. 120,911,775 lines of code.
- Fuzzing is time and resource intensive.
- We don’t have forever to run fuzzing campaigns.
- Target’s surface changes rapidly.
  - By the time we fully fuzz version 1, version 3 or more is released.
  - Part of APIv1 we fuzzed is now deprecated.

WHAT

Compile-time information ⚙️

+

Machine Learning 🔍

=

Educated Guess 📊

COMPILE-TIME INFORMATION

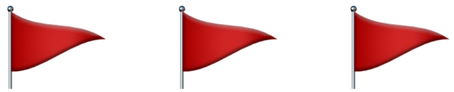
- Language Independent Features:
  - Control Flow analysis
  - Programming pattern analysis
  - Pointer analysis
  - Loop complexity analysis
  - Syscall sequence
- Language Dependent Features:
  - ABI Calls
  - Std library calls
  - Memory management
  - Pointer sanitization
  - Type information

FUNCTION COMPILE-TIME INFORMATION

1. Instructions
2. BBs
3. In-degree
4. Out-degree
5. Num Loops
6. Static Allocations
7. Dynamic Allocations
8. MemOps
9. CondBranches
10. UnCondBranches
11. DirectCalls
12. InDirectCalls

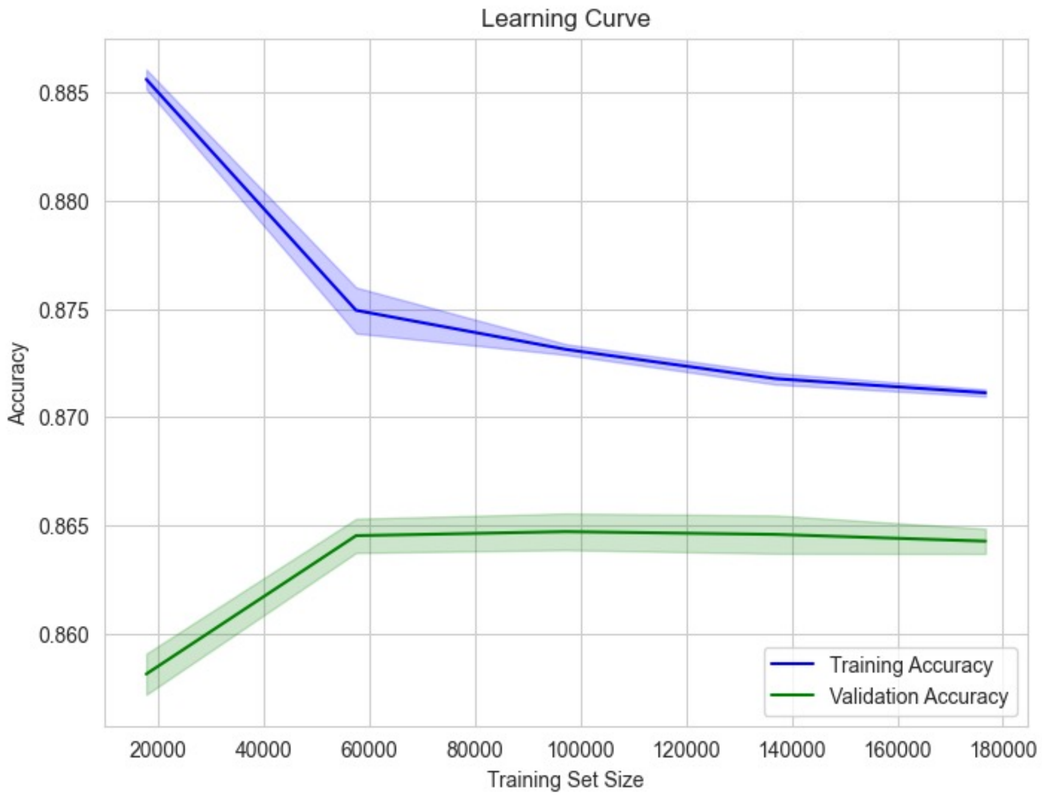
LEARNING CODE PATTERNS

```
complex loops + user input + manipulates *ptr + memcpy
```

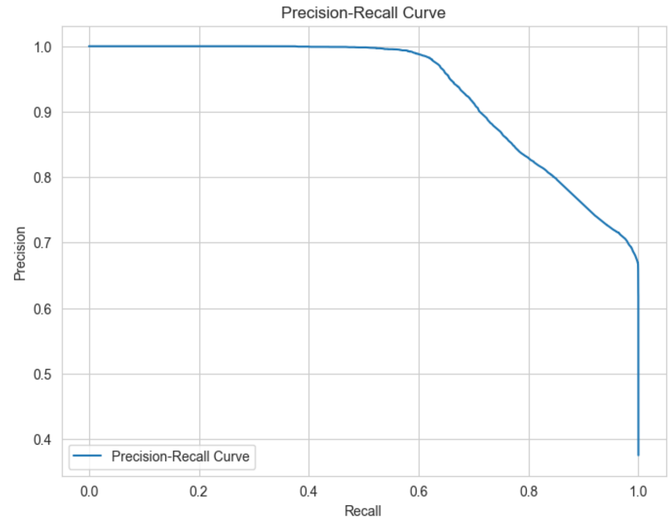


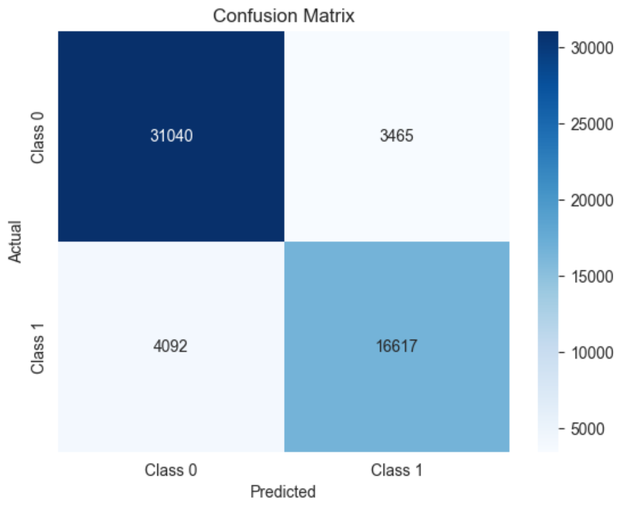
- Throw selected program features at Machine Learning algorithms.
- Generate ML models that can (reasonably) make sense of these features.
- Use this knowledge to get best candidates for fuzzing.
- Spend your resources and time on these selected candidates.

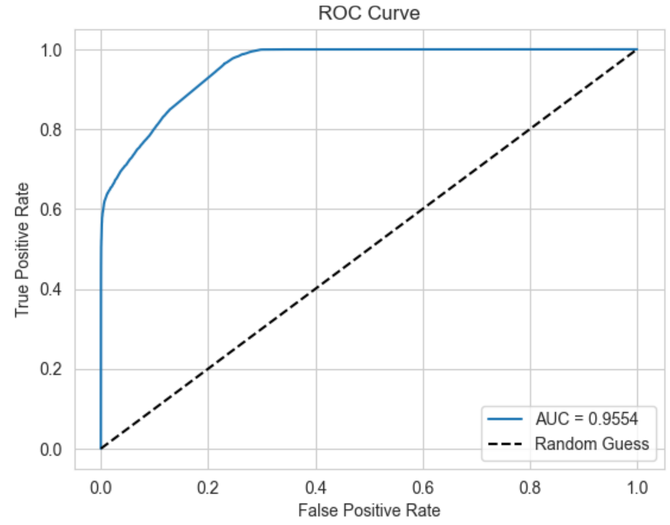
FUZZDISTILL-ML (XGBOOST) [86%]



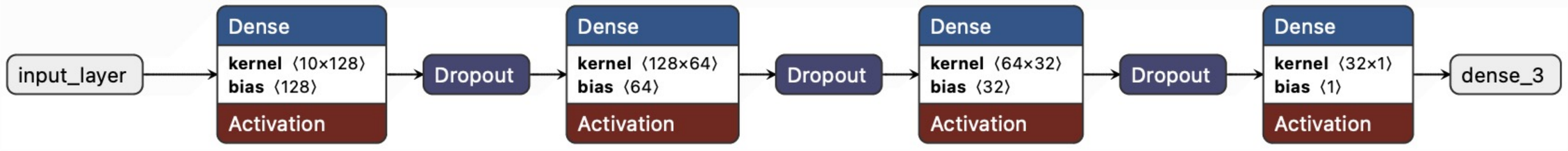

```
model = xgb.XGBClassifier(
  objective='binary:logistic',
  eval_metric='logloss',
  random_state=GLOBAL_RANDOM_STATE,
  colsample_bytree=0.8,
  learning_rate=0.05,
  max_depth=10,
  n_estimators=400,
  subsample=0.8
)
```

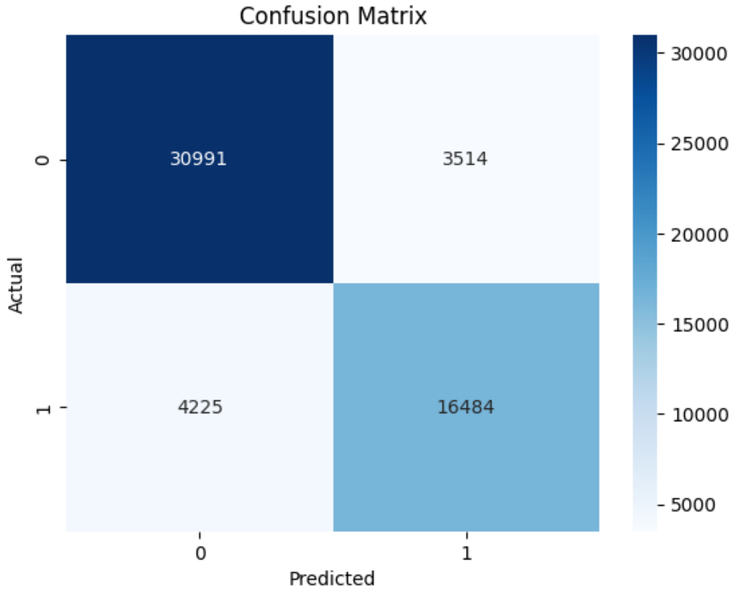


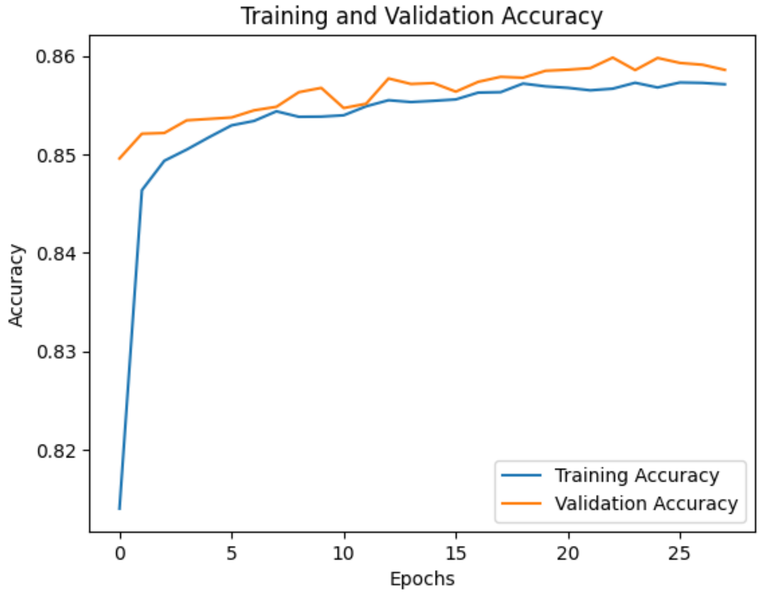


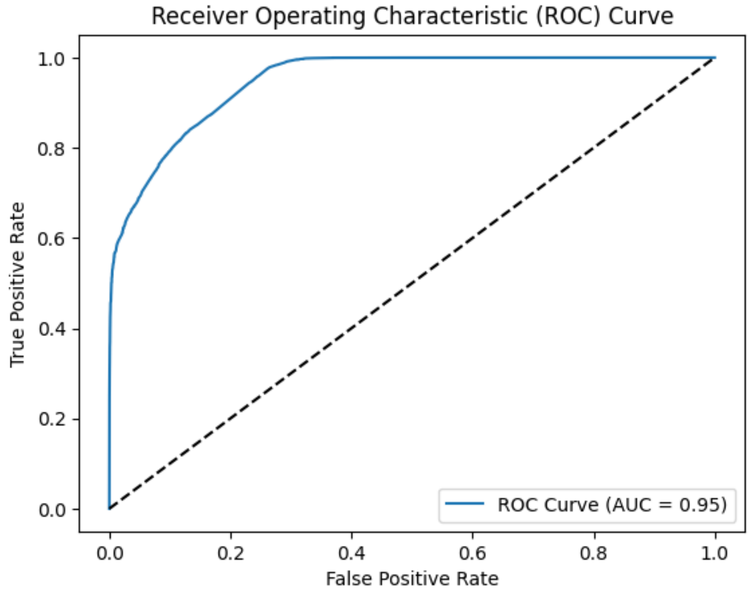


FUZZDISTILL-ML (DEEP NEURAL NETWORK) [86%]

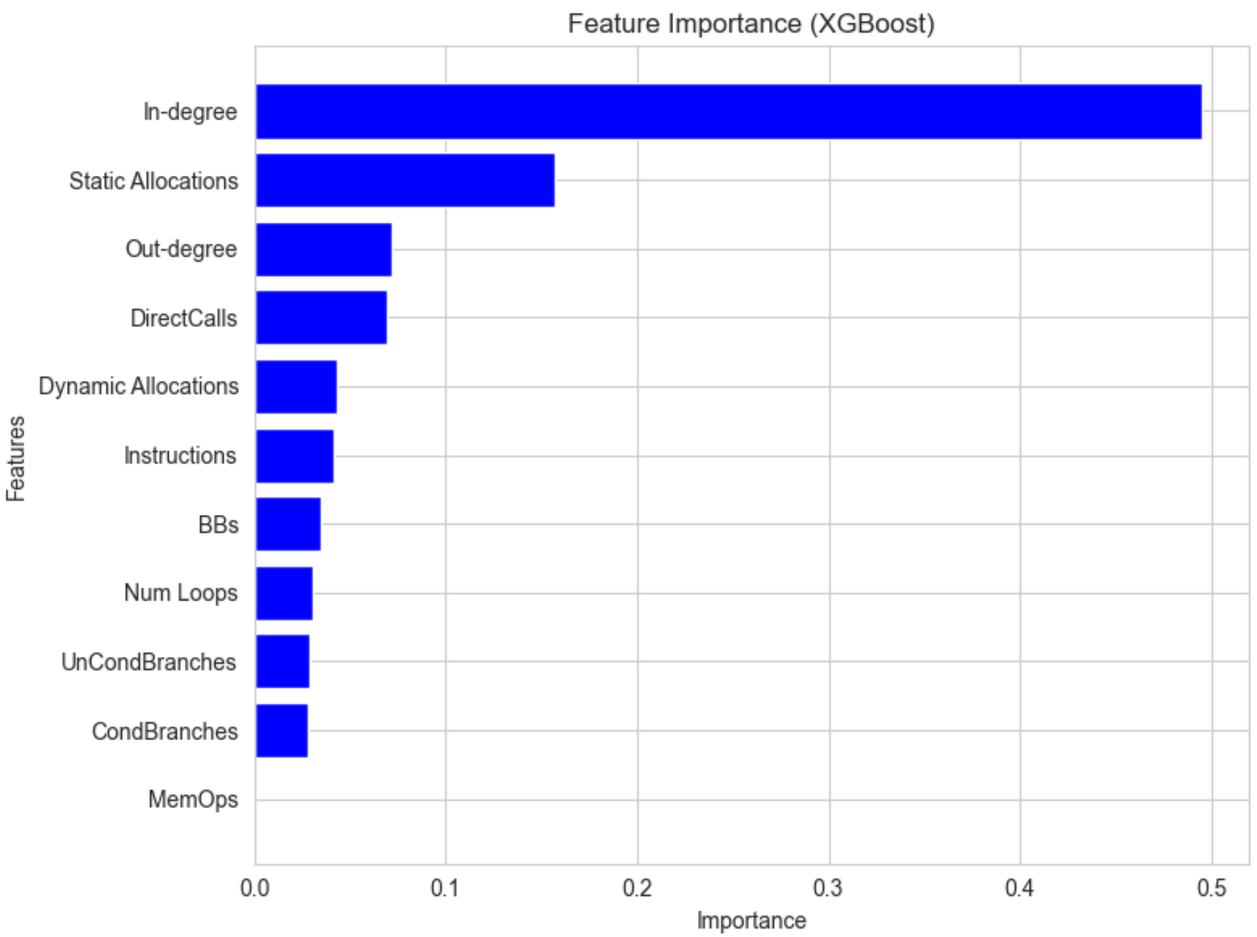


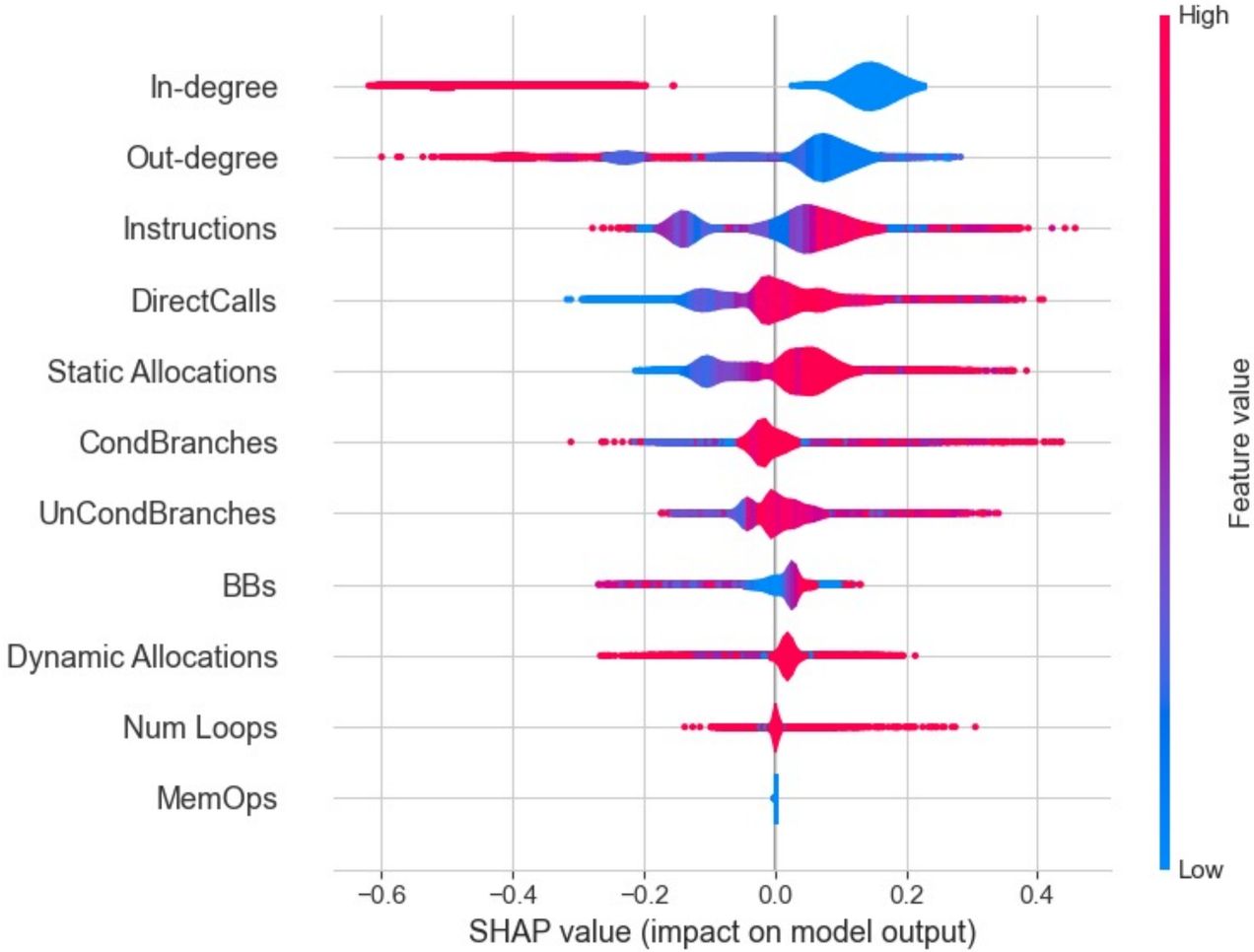






FEATURE IMPORTANCE (XGB VS DNN)





HOW

FUZZDISTILL - ARCHITECTURAL OVERVIEW

III. Architecture

FuzzDistill[5, 6, 7] is made up of three components,

1) FuzzDistillCC: Compiler back-end for feature extraction,
2) FuzzDistillML: Model training component, and
3) FuzzDistillWeb: Prediction front-end

FuzzDistillML and FuzzDistillWeb rely on FuzzDistillCC to provide extracted program features as shown in Figure-1.

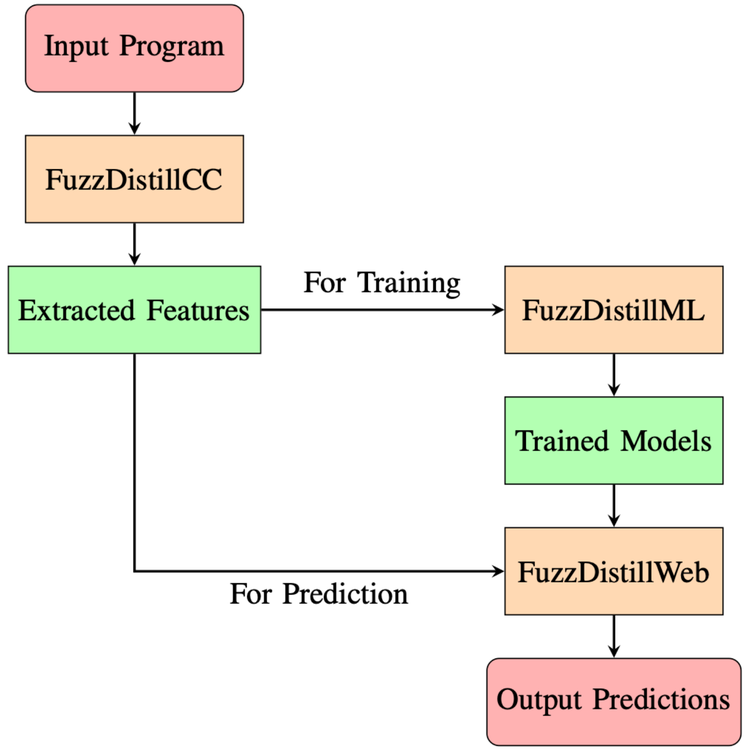
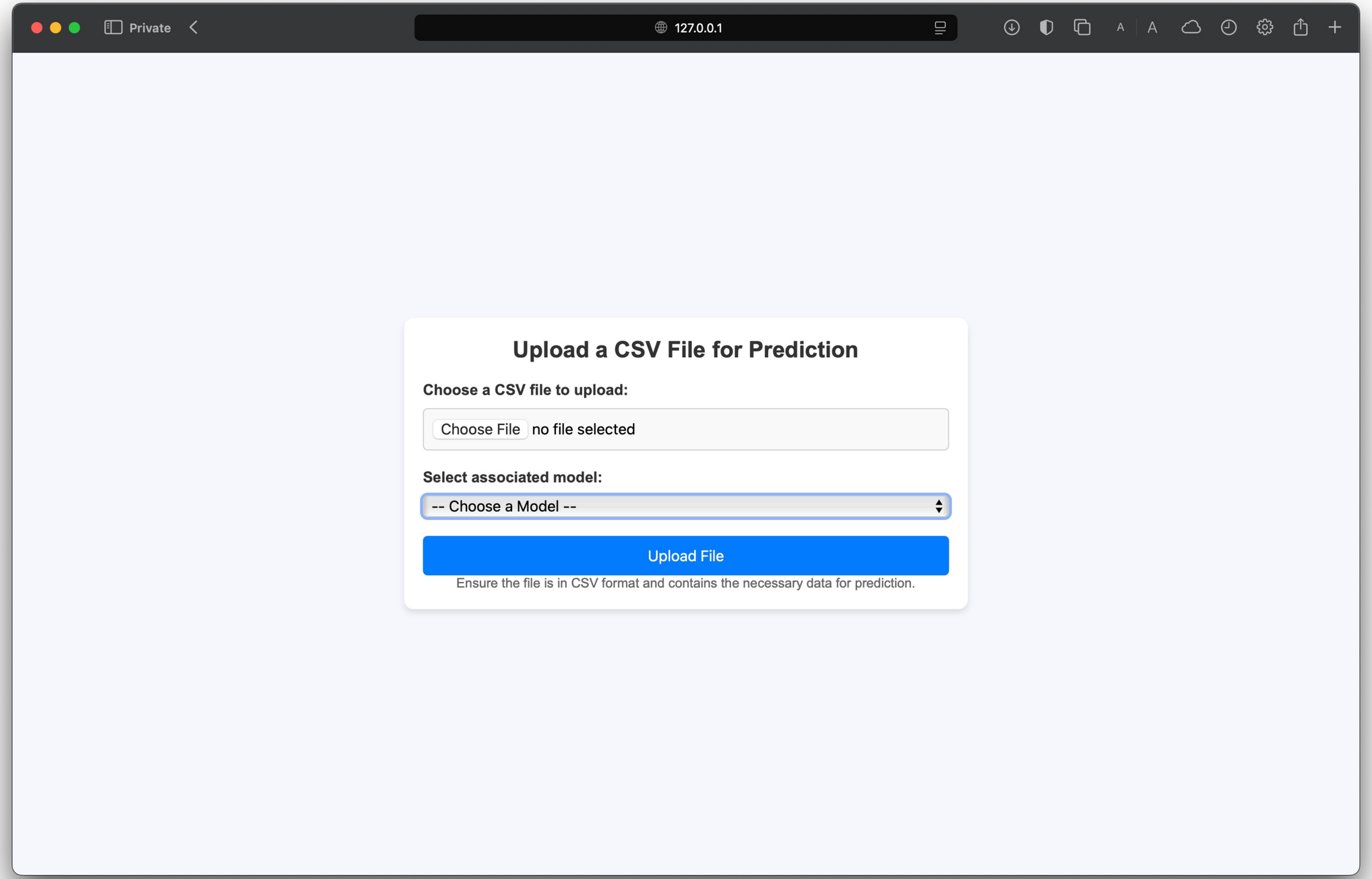


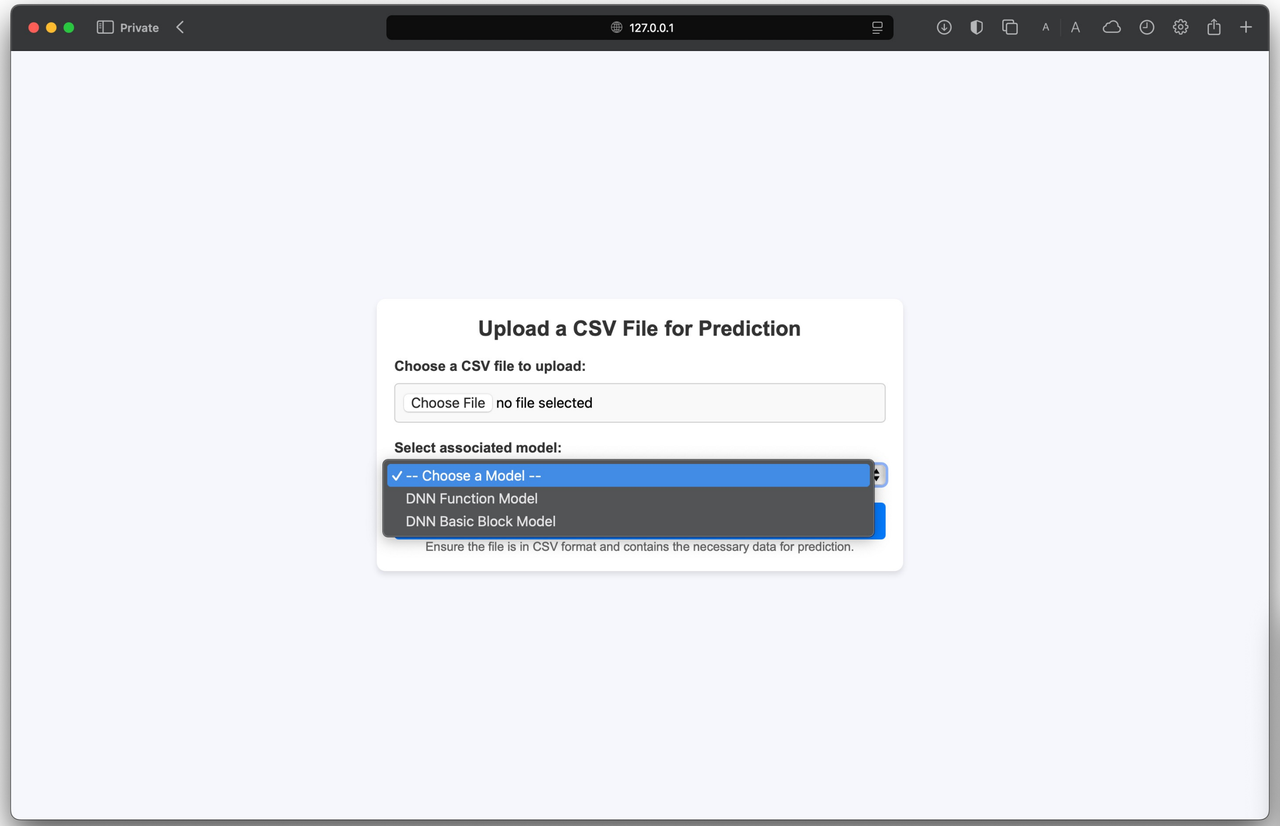
Fig. 1. Workflow of FuzzDistill

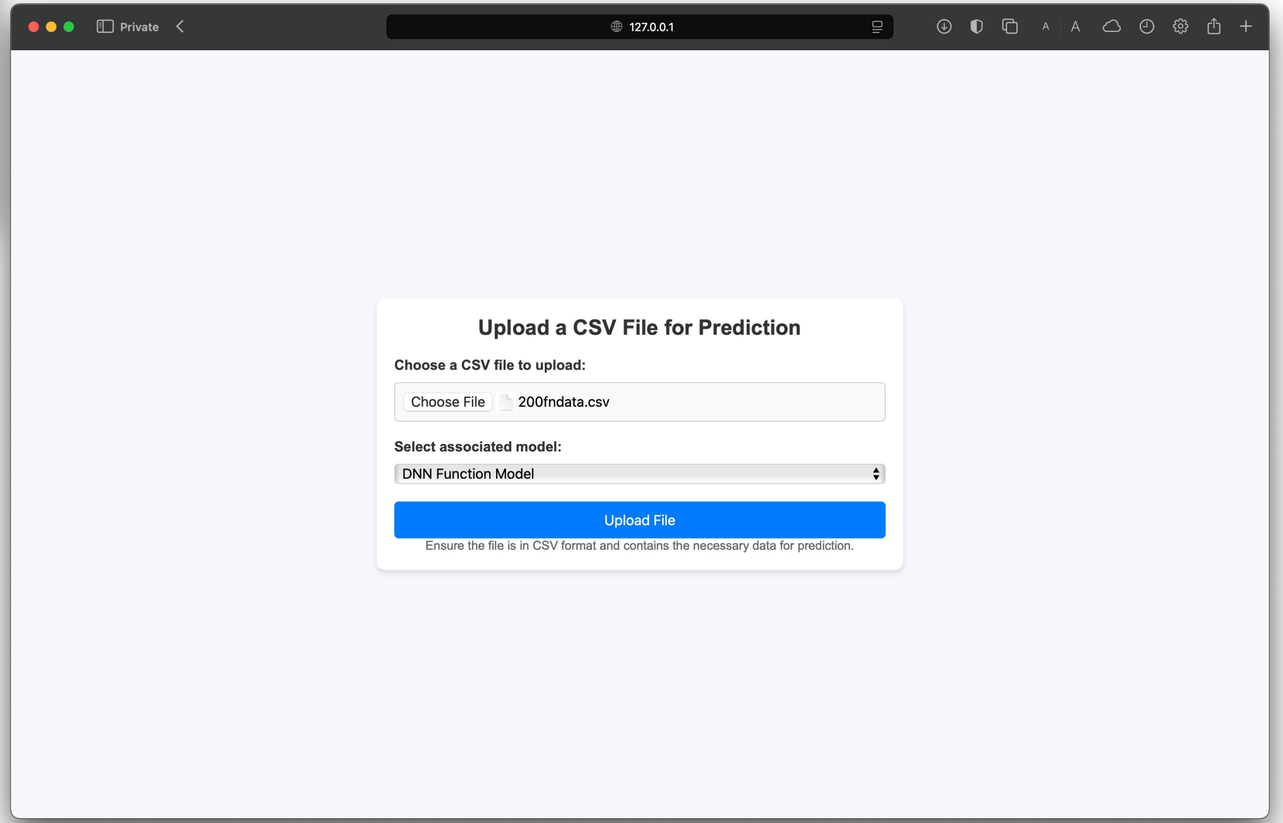
DEMO


JUST

SCREENSHOTS




Private
127.0.0.1
Upload a CSV File for Prediction
Choose a CSV file to upload:
Choose File
no file selected
Select associated model:
-- Choose a Model --
DNN Function Model
DNN Basic Block Model
Ensure the file is in CSV format and contains the necessary data for prediction.

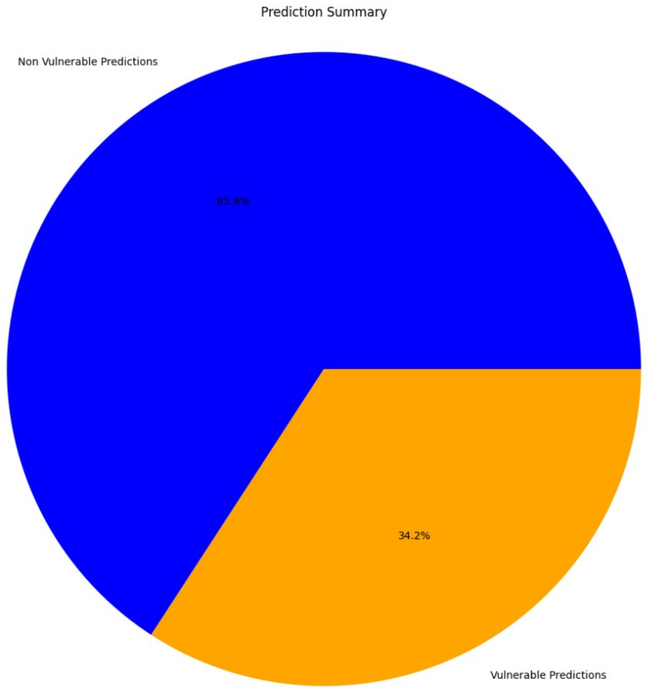

Private
127.0.0.1
Upload a CSV File for Prediction
Choose a CSV file to upload:
Choose File
200fndata.csv
Select associated model:
DNN Function Model
Upload File
Ensure the file is in CSV format and contains the necessary data for prediction.

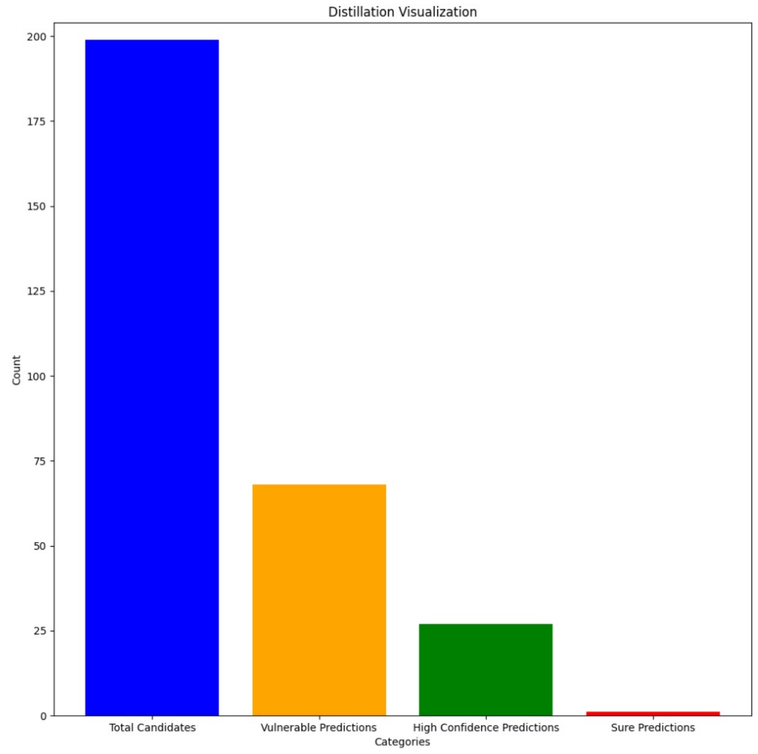
DNN_FN Prediction Results

Current Model Accuracy is 86%

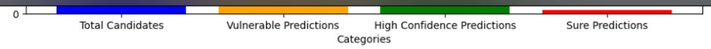
Data Visualization

523ee44fdd2206c80df9429ebb48a394f7d0f2df5a9d7bde5ffff8e11c020ac3





High Confidence Functions (Confidence > 95%)

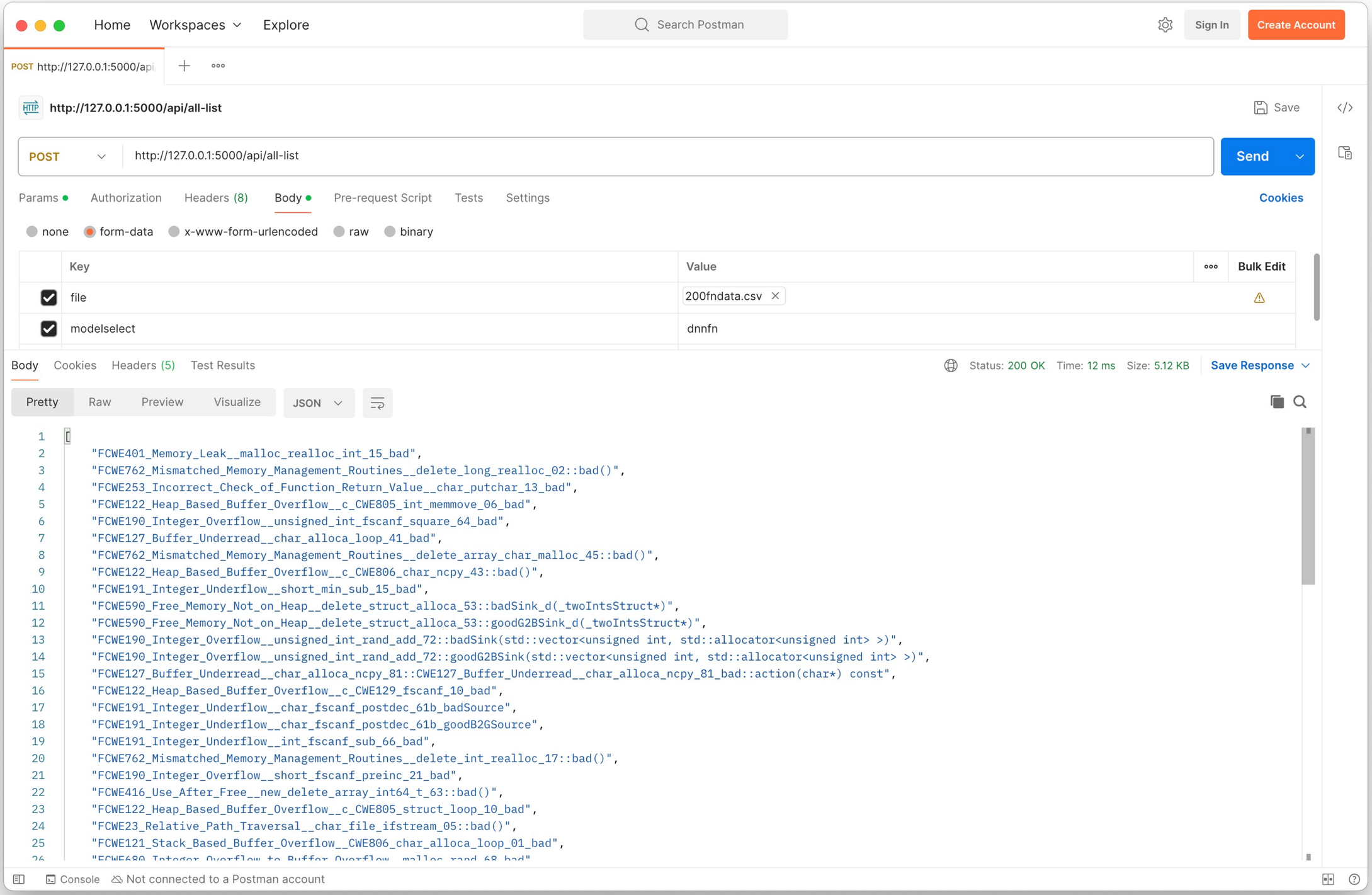


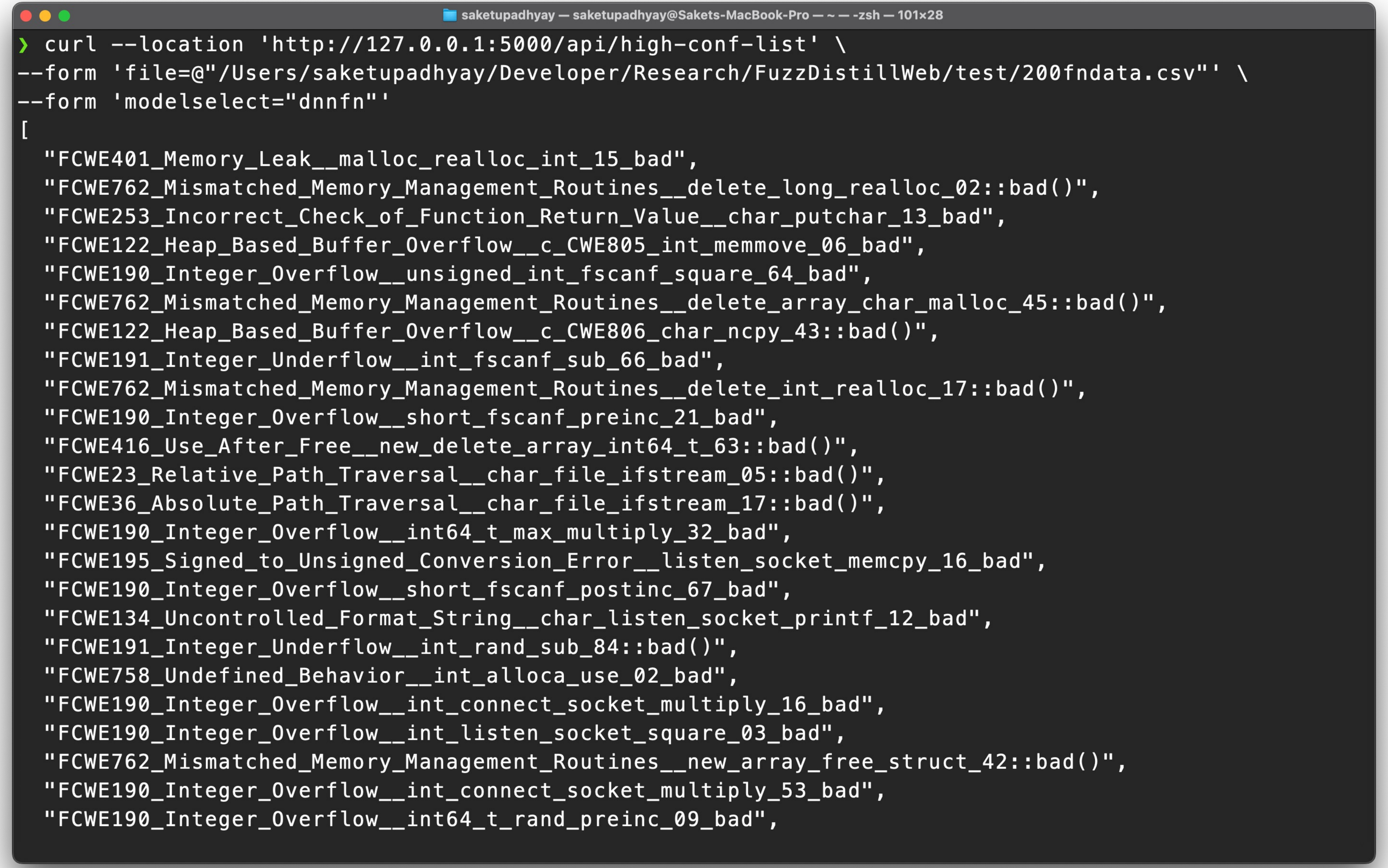
High Confidence Functions (Confidence > 95%)

- FCWE401_Memory_Leak__malloc_realloc_int_15_bad
- FCWE762_Mismatched_Memory_Management_Routines__delete_long_realloc_02::bad()
- FCWE253_Incorrect_Check_of_Function_Return_Value__char_putchar_13_bad
- FCWE122_Heap_Based_Buffer_Overflow__c_CWE805_int_memmove_06_bad
- FCWE190_Integer_Overflow__unsigned_int_fscanf_square_64_bad
- FCWE762_Mismatched_Memory_Management_Routines__delete_array_char_malloc_45::bad()
- FCWE122_Heap_Based_Buffer_Overflow__c_CWE806_char_ncpy_43::bad()
- FCWE191_Integer_Underflow__int_fscanf_sub_66_bad
- FCWE762_Mismatched_Memory_Management_Routines__delete_int_realloc_17::bad()
- FCWE190_Integer_Overflow__short_fscanf_preinc_21_bad
- FCWE416_Use_After_Free__new_delete_array_int64_t_63::bad()
- FCWE23_Relative_Path_Traversal__char_file_ifstream_05::bad()
- FCWE36_Absolute_Path_Traversal__char_file_ifstream_17::bad()
- FCWE190_Integer_Overflow__int64_t_max_multiply_32_bad
- FCWE195_Signed_to_Unsigned_Conversion_Error__listen_socket_memcpy_16_bad
- FCWE190_Integer_Overflow__short_fscanf_postinc_67_bad
- FCWE134_Uncontrolled_Format_String__char_listen_socket_printf_12_bad
- FCWE191_Integer_Underflow__int_rand_sub_84::bad()
- FCWE758_Undefined_Behavior__int_alloca_use_02_bad
- FCWE190_Integer_Overflow__int_connect_socket_multiply_16_bad
- FCWE190_Integer_Overflow__int_listen_socket_square_03_bad
- FCWE762_Mismatched_Memory_Management_Routines__new_array_free_struct_42::bad()
- FCWE190_Integer_Overflow__int_connect_socket_multiply_53_bad
- FCWE190_Integer_Overflow__int64_t_rand_preinc_09_bad
- FCWE23_Relative_Path_Traversal__char_file_ifstream_17::bad()
- FCWE762_Mismatched_Memory_Management_Routines__new_array_free_struct_83::bad()
- FCWE427_Uncontrolled_Search_Path_Element__char_environment_42_bad

Sure Functions (Confidence == 100%)

- FCWE253_Incorrect_Check_of_Function_Return_Value__char_putchar_13_bad


Home
Workspaces
Explore
Search Postman
Sign In
Create Account
POST http://127.0.0.1:5000/api
http://127.0.0.1:5000/api/all-list
Save
POST
http://127.0.0.1:5000/api/all-list
Send
Params
Authorization
Headers (8)
Body
Pre-request Script
Tests
Settings
Cookies
none
form-data
x-www-form-urlencoded
raw
binary
Key
Value
Bulk Edit
file
200fndata.csv
modelselect
dnnfn
Body
Cookies
Headers (5)
Test Results
Status: 200 OK
Time: 12 ms
Size: 5.12 KB
Save Response
Pretty
Raw
Preview
Visualize
JSON
[
"FCWE401_Memory_Leak__malloc_realloc_int_15_bad",
"FCWE762_Mismatched_Memory_Management_Routines__delete_long_realloc_02::bad()",
"FCWE253_Incorrect_Check_of_Function_Return_Value__char_putchar_13_bad",
"FCWE122_Heap_Based_Buffer_Overflow__c_CWE805_int_memmove_06_bad",
"FCWE190_Integer_Overflow__unsigned_int_fscanf_square_64_bad",
"FCWE127_Buffer_Underread__char_alloca_loop_41_bad",
"FCWE762_Mismatched_Memory_Management_Routines__delete_array_char_malloc_45::bad()",
"FCWE122_Heap_Based_Buffer_Overflow__c_CWE806_char_ncpy_43::bad()",
"FCWE191_Integer_Underflow__short_min_sub_15_bad",
"FCWE590_Free_Memory_Not_on_Heap__delete_struct_alloca_53::badSink_d(_twoIntsStruct*)",
"FCWE590_Free_Memory_Not_on_Heap__delete_struct_alloca_53::goodG2BSink_d(_twoIntsStruct*)",
"FCWE190_Integer_Overflow__unsigned_int_rand_add_72::badSink(std::vector<unsigned int, std::allocator<unsigned int> >)",
"FCWE190_Integer_Overflow__unsigned_int_rand_add_72::goodG2BSink(std::vector<unsigned int, std::allocator<unsigned int> >)",
"FCWE127_Buffer_Underread__char_alloca_ncpy_81::CWE127_Buffer_Underread__char_alloca_ncpy_81_bad::action(char*) const",
"FCWE122_Heap_Based_Buffer_Overflow__c_CWE129_fscanf_10_bad",
"FCWE191_Integer_Underflow__char_fscanf_postdec_61b_badSource",
"FCWE191_Integer_Underflow__char_fscanf_postdec_61b_goodB2GSource",
"FCWE191_Integer_Underflow__int_fscanf_sub_66_bad",
"FCWE762_Mismatched_Memory_Management_Routines__delete_int_realloc_17::bad()",
"FCWE190_Integer_Overflow__short_fscanf_preinc_21_bad",
"FCWE416_Use_After_Free__new_delete_array_int64_t_63::bad()",
"FCWE122_Heap_Based_Buffer_Overflow__c_CWE805_struct_loop_10_bad",
"FCWE23_Relative_Path_Traversal__char_file_ifstream_05::bad()",
"FCWE121_Stack_Based_Buffer_Overflow__CWE806_char_alloca_loop_01_bad",
Console
Not connected to a Postman account


saketupadhyay — saketupadhyay@Sakets-MacBook-Pro — ~ — -zsh — 101×28
❯ curl --location 'http://127.0.0.1:5000/api/high-conf-list' \
--form 'file=@"/Users/saketupadhyay/Developer/Research/FuzzDistillWeb/test/200fndata.csv"' \
--form 'modelselect="dnnfn"'
[
"FCWE401_Memory_Leak__malloc_realloc_int_15_bad",
"FCWE762_Mismatched_Memory_Management_Routines__delete_long_realloc_02::bad()",
"FCWE253_Incorrect_Check_of_Function_Return_Value__char_putchar_13_bad",
"FCWE122_Heap_Based_Buffer_Overflow__c_CWE805_int_memmove_06_bad",
"FCWE190_Integer_Overflow__unsigned_int_fscanf_square_64_bad",
"FCWE762_Mismatched_Memory_Management_Routines__delete_array_char_malloc_45::bad()",
"FCWE122_Heap_Based_Buffer_Overflow__c_CWE806_char_ncpy_43::bad()",
"FCWE191_Integer_Underflow__int_fscanf_sub_66_bad",
"FCWE762_Mismatched_Memory_Management_Routines__delete_int_realloc_17::bad()",
"FCWE190_Integer_Overflow__short_fscanf_preinc_21_bad",
"FCWE416_Use_After_Free__new_delete_array_int64_t_63::bad()",
"FCWE23_Relative_Path_Traversal__char_file_ifstream_05::bad()",
"FCWE36_Absolute_Path_Traversal__char_file_ifstream_17::bad()",
"FCWE190_Integer_Overflow__int64_t_max_multiply_32_bad",
"FCWE195_Signed_to_Unsigned_Conversion_Error__listen_socket_memcpy_16_bad",
"FCWE190_Integer_Overflow__short_fscanf_postinc_67_bad",
"FCWE134_Uncontrolled_Format_String__char_listen_socket_printf_12_bad",
"FCWE191_Integer_Underflow__int_rand_sub_84::bad()",
"FCWE758_Undefined_Behavior__int_alloca_use_02_bad",
"FCWE190_Integer_Overflow__int_connect_socket_multiply_16_bad",
"FCWE190_Integer_Overflow__int_listen_socket_square_03_bad",
"FCWE762_Mismatched_Memory_Management_Routines__new_array_free_struct_42::bad()",
"FCWE190_Integer_Overflow__int_connect_socket_multiply_53_bad",
"FCWE190_Integer_Overflow__int64_t_rand_preinc_09_bad",

WHERE

OPEN SOURCED @ GITHUB

FuzzDistillCC @ github.com/Saket-Upadhyay/FuzzDistillCC

FuzzDistillML @ github.com/Saket-Upadhyay/FuzzDistillML

FuzzDistillWeb @ github.com/Saket-Upadhyay/FuzzDistillWeb