- Additional details regarding each field should be left as a comment
- If the study re-uses an existing dataset, do not re-consider the data of the original dataset.
- If a study uses multiple datasets, fill out the details for each dataset within each cell (indicating which information for which dataset).

Name of Reviewer:			
Date of Data Extraction:			
Paper Citation:			
Paper Summary:			
Additional Notes:			
Category	**Data Item:**	**Values:**	**Description:**
Data Requirement	Application Context	Within-project, Cross-project, Mixed, Unspecified.	A description of the application context in which the model is trained and applied.
	Data Granularity	File level, Function level, Commit level, Program Slices, Other.	The granularity of the input data and its application scenario.
	Code Object	Source Code, Intermediary representation, Binary, Other.	The way the code is represented.
	Feature Type	Metric-Based, Text-Based, Graph-Based	AST and Code graphs are graph based.
	Feature Representation	Software Metrics, Token Frequency, NLP Embeddings, Deep Feature Representation, Other.	The types of features extracted: Token Frequency – Bow, TFIDF, etc. NLP Embeddings – word2vec, glove, etc. Deep feature representation – Non-NLP based code vectors. Using a neural network for automatic feature representation.
	Comparative Study	Yes/No	Whether the studies main motivation is to compare multiple existing methods.
Dataset Information	Dataset Name	Name	The name of the dataset(s) utilized in the study. If the dataset name does not overlap with source, provide the source in brackets.
	Data Source Multiplicity	Single, Multiple.	The number of datasets utilized. Each project is a separate data source.
	Data Integration	Separate, Validation, Merged	If more than one dataset is used, how are they used. Separate – Multiple datasets

			are used in the same manner but without mixing. Validation – The datasets are used at different stages of the pipeline (e.g. one for tuning, another for testing). Merged – The datasets are combined and used as one.
	Considered Projects	Name	If the dataset considers specific projects, name all the projects.
	Data Availability	Available, replicatable, reproducible, hard to reproduce, not available	The availability of the dataset for it to be reused: • Available: The explicit dataset is provided (open-source). • Replicatable: The method can be followed to produce the exact dataset. • Reproducible: The method can be followed to produce a similar dataset. • Hard to reproduce: The method can be followed to produce a different dataset. • Not Reproducible: The method is too implicit to follow. • Not Accessible: (e.g. not open source).
	Dataset Link	URL	A link to the source of the dataset
	Dataset Label Source	NVD, SARD, Bug Reports, Patches, Security Advisories, Prior Labelling, None	The source from which the security information is extracted.
	Dataset Source Type	Synthetic, Open-source, Private-source	The nature of the dataset source.
	Vulnerability Types	Name, CWE-ID, or Not reported.	The types of vulnerabilities considered.
	Uniqueness	New, Augment, Subset, Extend, Re-use	Whether they create a new dataset, or re-use/modify an existing one. Augment – change or add more information. Subset/Extend - Re-use but either take a portion or collect additional data.
	Programming Language	Name	The programming language considered.
	Timespan	Date	The timespan of the dataset.
	Data Size	Integer	The number of samples for each class and the total size. If multiple projects or datasets, list for each.

Data Extraction	Data Collection	Description	A short one paragraph summary of the data collection process conducted.
Data Extraction	Data Versioning	Considered, Not considered	Whether data is extracted using specific versions or not.
	Data Localization	None, Manual, Prior Localization, Other.	How the data is localized from the raw source code modules. (e.g. identify vulnerable functions, commits, etc.) None – Raw data is already at the desired data granularity (e.g. File-level). Prior Localization – Reuse of another dataset. Other – Using a (semi-)automatic method to reduce scope (e.g. to commit, line, function, level)
Data Labelling	Vulnerable Data Labelling	Pattern, tool, developer, manual	The method used to label the vulnerable data.
	Manual Inspection	No, Yes	Whether manual efforts (of the study authors) are expended in labelling or processing the data (yes), or not (no).
	Non-vulnerable data labelling	Not-vulnerable, Fix, Heuristic-based	The method used to extract the negative class; taking all modules not in the positive class, or using heuristic methods.
Data Processing	Data Transformation	Scaling, normalization, replacement, mapping, etc.	Methods utilized to transform the data into a more suitable representation for the ML algorithm.
	Data Cleaning	Removal, imputation, etc.	Methods used to clean the dataset, typically through removal of bad values.
Data Utilization	Data Sorting	Random, time-based	The method used to sort the data.
	Data Partitioning	Holdout, Cross-Validation, Bootstrapping, Other.	The method used to partition the data for training, validation and testing.
	Hyperparameter Tuning	Yes, No	Does the study tune the model hyperparameters.
	Stratified Sampling	Yes, No	Whether they apply stratified sampling methods.
	Resampling	None, Under/Over-sampling, etc.	Methods used to resample the classes.
Miscellaneous	Extra	Description	Any other noteworthy factors of the paper not contained in the prior fields.
Data Considerations	Dataset Selection	Description	The considerations made for selecting the appropriate dataset. Extraction:

			Any descriptive sentences following the introduction of the dataset.
	Application Decision	Description	The reasoning behind data requirement fields. Extraction: Descriptive sentences about the context of the model they're producing. Usually described in the introduction, background and method sections. This relates to reasoning for cross/within-project, granularity (file-level, program slices, etc.), source vs binary code, and choice of features.
	Addressed Issues	Description	The data preparation issues that they identify and attempt to address through their method. Extraction: Data preparation methods they apply to help address explicit challenges or flaws in the data. Usually described in the methodology and threats to limidation sections. Can overlap with prior fields (e.g. manual inspection, data cleaning, resampling, etc.) if they provide reasoning for these methods to address a challenge.
	Un-addressed Issues	Description	Any issues, challenges or limitations of the data explicitly noted in the study that were not addressed (threats to validity). Extraction: Data preparation challenges or flaws in the data that are mentioned, but not solved; usually as they describe it as having minimal impacts, being too difficult to solve, or future work. E.g. datasets may not be generalisable, undocumented vulnerabilities may be in the non-vulnerable set, etc. Usually described in the methodology and threats to limidation sections.

Performance Evaluation	Performance Comparison	Yes, No	Does the study compare application of different data requirements.
Data Quality Assessment	Outliers	Addressed, Present, Considered, Not Considered. Explanation: Addressed – Relevant data quality techniques are applied. Present – From contextual information of the dataset, it's likely that this data quality issue is present, but they are not addressed. Considered – The authors acknowledge that this data issue may be present, but do not address (either due to difficulty or negligence). Not Considered – The issue is not considered or mentioned.	Data points that lie outside the overall distribution.
	Noise		The presence of erroneous or incorrect data.
	Inconsistency		Consistency of the recorded data (e.g. bug reports). This can be things such as unexplainable values, varying label quality or inconsistent recording measures. Present if mixing data sources.
	Incompleteness		Missing values within the dataset.
	Redundancy		Duplicate data points.
	Amount of Data		The number of samples in the dataset. (Addressed) if explicit efforts are made to increase the data size are mentioning that it may be an issue. (Present) category should not be used as it is too subjective to evaluate.
	Heterogeneity		Whether multiple datasets are considered (good) or only a single source (bad). Leave as not considered unless explicitly considered.
	Timeliness		Whether data sets consider a time based order. Also comment whether datasets were up to data at time of publication. Addressed – acknowledge time based nature of data, Present – do not acknowledge, Considered – Acknowledge but do not use time-based order, Not considered – Unclear.
	Commercial Sensitivity		The effort to hide or anonymize potentially sensitive data. Mark as (present) if it seems like there is commercially sensitive data included, (addressed) if efforts are made to anonymize commercially sensitive data, otherwise (not considered).
	Accessibility		Whether the data is publicly available and accessible (addressed), withheld (present), reproducible (considered), or too vague or unclear to reproduce (not considered).
	Trustworthiness		The documentation of the dataset and ability for it to be replicated (even though data may not be accessible). If replicatable – addressed, if reproducible but difficult – considered, otherwise - present