scientific reports

OPEN

Hybrid net powered large scale audit of dataset licensing and attribution practices for enhanced transparency and compliance

Velmurugan Ayyamperumal[1]✉, S. Aswath[2], S. Vignesh[3] & T. Thamaraimanalan[4]

In the era of data-driven research and artificial intelligence, proper dataset licensing and attribution practices are crucial for legal compliance and ethical data usage. Open-access datasets are often shared under various licensing schemes, such as Creative Commons (CC) or MIT, each with distinct usage and attribution requirements. However, ensuring adherence to these requirements across vast repositories poses a significant challenge. This study presents a HybridNet-powered system combining RoBERTa and InceptionV3 models to audit large-scale dataset licensing and attribution practices for enhanced transparency and legal compliance. The system leverages RoBERTa for natural language processing (NLP) to classify licensing terms and detect attribution requirements in textual metadata, while InceptionV3 handles visual attribution embedded in images. A comprehensive dataset audit was conducted on 5,000 datasets from the OpenML repository, resulting in an overall accuracy of 95% for detecting and classifying licenses and attributions. Cross-validation with OpenML metadata showed 94% consistency in license classification and 90% consistency in attribution detection. License violations were identified in 5% of the datasets, while attribution violations were flagged in 6%, leading to an overall compliance violation rate of 11%. The system's ensemble approach significantly outperformed traditional models, such as Logistic Regression (accuracy: 72%) and Support Vector Machines (accuracy: 75%), demonstrating its effectiveness in auditing multi-modal dataset content. By flagging datasets with missing or inconsistent license and attribution information, the system enables corrective action, improving the transparency of dataset repositories.

Keywords HybridNet, RoBERTa, InceptionV3, Dataset licensing, Attribution detection, NLP, Computer vision, Compliance audit, Dataset transparency, OpenML

The rapid advancement in data technologies means that there is increased pressure on organizations to maintain a good-quality dataset. Among the subdomains of this area, it is possible to allocate the evaluation of licensing and attribution of datasets that determine compliance with the legal and ethical requirements and organizational norms. The emergence of ML, AI and data-driven systems means that datasets are now some of the most sought-after commodities[1,2]. Collections of data to train models, conduct tests, or create hypotheses and outcomes are known as datasets, but it is clear that the legal requirements regarding the use of datasets, with special focus on licensing and attribution, are not always clear or followed. While licensing informs the community on how a dataset can be used, attribution makes it obligatory for users to acknowledge the creators of the dataset. This first discussion examines the increasing need to audit these practices, check compliance with licenses, and keep investors and the public informed and protected in the use of datasets[3,4]. It has become the norm to share and reuse datasets, as evidenced by the growing instances in applications such as health, education, finance, and the government sector. Licensing also promotes the use of these datasets in ways that respect the rights of creators as individuals, organizations, or government[5]. It's a document that outlines how a dataset can be accessed, used, modified and even distributed. There are several open data licenses commonly used to establish usage conditions

[1]Department of Computer Science and Engineering, Jerusalem College of Engineering, Pallikaranai 600100, Tamil Nadu, India. [2]Department of Electronics and Communication Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India. [3]Department of Electronics and Communication Engineering, Sasi Institute of Technology and Engineering, Sasi College Rd, Near Aerodrome, Tadepalligudem 534101, Andhra Pradesh, India. [4]Department of Electronics and Communication Engineering, Sri Eshwar College of Engineering, Coimbatore 641202, Tamil Nadu, India. ✉email: contactvelan@gmail.com

for open datasets provided in the academic and research arenas, including the Creative Commons (CC) licenses or the Open Data Commons (ODC) licenses[6,7]. The importance of dataset licensing is multi-fold: it legalizes the use of the dataset, protects the ownership of ideas, clears doubts that may hinder sharing of data and supports innovation subject to ownership. However, as crucial licensing is for the business, compliance with licensing terms is a weakness or is not controlled systematically, and that is why systematic audits are needed[8,9].

The other significant problem that arises when it comes to the compliance of dataset licensing is due to the variety of licenses that are available. It is also important to note that different datasets may be controlled by different types of licenses, such as open data licenses, restrictive licenses that may limit the ways and scopes in which the dataset may be employed in many sectors of the economy[10–12]. For instance, some licenses may permit the use of a dataset for research purposes that is being conducted in a university, but ban for use that is connected to business or any other form of profit-making without consent. Some other licenses can demand that the authors share the changes made to the given dataset under the same conditions as in the case of the copyleft in licensing open-source software[13]. There are usually some legal issues concerning the understanding of the legal requirements of these licenses when working with the datasets. In addition, the license of a dataset could change over time, and that makes compliance a tough task. Therefore, whereas the auditing of dataset licensing practices is a critical undertaking to understand how organisations utilise research datasets within stated parameters that should not be violated, this effort is equally important to guarantee that organizations adhere to the legal guidelines set by dataset owners. The final thing that needs to be considered when it comes to the proper usage of datasets is attribution, or giving credit to the creators of the dataset. Similar to how authors of scientific papers are rewarded for their work, the creators of the datasets also need to be rewarded whenever their datasets are used in other research or even commercial projects[14]. Scientific practices of attribution not only extend credit to the creators of the datasets but also promote open sharing of datasets by putting the limelight on the creators' work. Furthermore, attribution is beneficial for improving the understandability of the usage of a dataset or for providing better insights into data lineage. Nevertheless, the practice of attribution is not exceptionally standardised and rigorous where applicable, especially when handling big data from multiple sources or if data is being transformed[15]. This is particularly problematic in settings, including federated learning, where data can be split and owned by different participants. This is a major risk to ethical considerations, especially in an academic and research-related line where credit to original work must be accorded appropriately. Figure 1 shows the benefits of dataset auditing.

Today, there is a constant emergence of a set of legal regulations regarding the licensing of datasets in accordance with data rights and increased transparency of data usage. For instance, the General Data Protection Regulation (GDPR) in Europe has affected datasets with personally identifying data in many ways, making organisations cautious when collecting, storing and sharing data[16]. Other laws like CCPA in the United States set up restrictions on the use of personal datasets, including restrictions as to transparency, consent and right to be forgotten. These laws are rather extensive in regulating aspects of dataset licensing and attribution based on certain conditions, including the use of personal data, proceeding under the law and the further control of this data by individuals. Even more important is auditing of prevailing licensing practices in the context of these legal frameworks to ensure that license compliance does not conflict with data privacy laws that impact on—we use, share and attribute datasets[17]. This is a key process when it comes to addressing the rights and wrongs of data utilization in the current data use culture. Because greater value is placed on datasets and usage extends to various fields, it is crucial to protect the licensing rules with regard to the usage of the datasets and

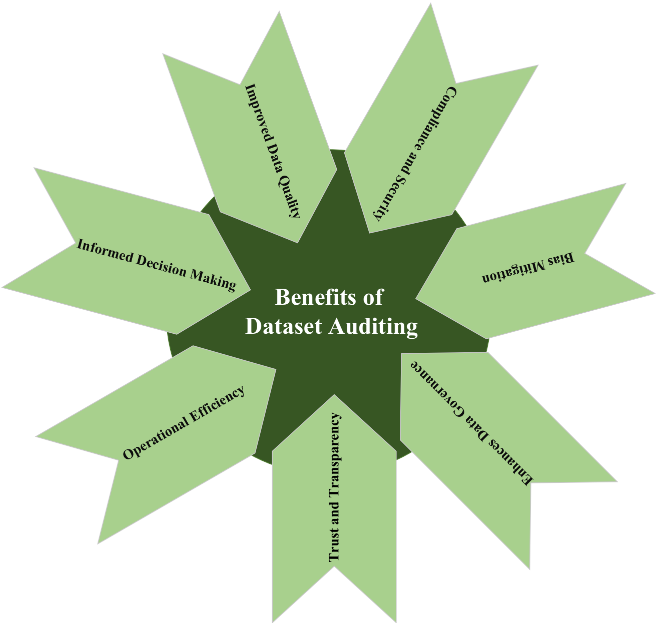


Fig. 1. Benefits of Dataset Auditing.

the attribution of the original creators. Due to recent trends in federated learning, decentralized data sharing, and new legal requirements, the importance of developing good auditing approaches has never been higher[18]. Current technological innovations like the blockchain and machine learning are among the opportunities that can improve auditing and ensure that datasets are used rightly and all the stakeholders are credited properly. Finally, the audit of dataset licensing and attribution practices will be the key factor in the further evolution of data use and regulation, providing growing protection for the owners and users of data in the conditions of their continuous interaction[19,20].

Main contributions of the work

- Development of HybridNet Ensemble Model: The paper also proposes an ensemble model, HybridNet, that integrates RoBERTa for processing textual data and InceptionV3 for handling the visual data involved in dataset auditing.
- Automated Licensing Classification: To be specific, the system quickly categorizes a number of licenses, including MIT, GPL, Creative Commons, etc., in a large data set of repositories, thus offering a scalable solution for repositories with different licensing policies.
- Comprehensive Attribution Detection: The work responds to the gestural practices of attribution detection that are marginalised contextually: it uses text analysis for attribution descriptors and object recognition for appreciable attributions within the images of datasets, such as watermarks or logos.
- Cross-Validation of Compliance: The system first checks if the manually detected license and attribution information match metadata from official sources and flags inconsistencies for further investigation.
- Violation Detection and Flagging: HybridNet learns about the datasets that either lack proper licenses or have improper or incomplete attribution and alerts researchers about such datasets so that corrective measures can be undertaken to keep the datasets intact and legal.
- Scalable Auditing System: The methodology is also scalable, so it should be able to fairly quickly audit thousands of datasets at once, making it well-suited for large-scale repositories like OpenML or other similar open-access databases.
- Multi-Modal Data Handling: Since the system is built with the combination of NLP and CV, it deals with more rigid and diverse data modalities, which contain license and attribution info in text and image both.

Section 2 provides an overview of related studies, focusing on previous efforts in dataset auditing, licensing compliance, and attribution detection, particularly those utilizing machine learning and NLP techniques. It explores the gaps in existing systems for automated large-scale dataset auditing. Section 3 details the proposed HybridNet-powered methodology, explaining how RoBERTa and InceptionV3 models are integrated for textual and visual data analysis to audit datasets effectively. Section 4 presents the results and discussion, analyzing the system's performance and the insights from the dataset audit. Finally, the conclusion and future scope outline the contributions and potential extensions of this work.

Related works

The high-speed construction of language models has been based on huge and inconsistently documented data, prompting ethical and legal issues. An over-1,800-text dataset multidisciplinary audit was done together with legal and machine learning experts. Tracing tools were created to determine the lineage of datasets, such as their origin, creators, licenses, and reuse patterns[11]. Results showed that there was often misclassification of licenses, with more than 70% of datasets not having appropriate attribution and more than 50% having errors in licensing. The creative, synthetic, and low-resource language data were frequently constraining regarding the licenses. Data Provenance Explorer was published to allow tracing popular fine-tuning models and encourage responsible use.

The increased reliance on massive amounts of data in machine learning studies has generated novel issues in the governance of datasets. A qualitative study explored the provenance, distribution practices, disclosure of ethical practices, and clarity of licensing through a systematic review of the datasets published in the NeurIPS Datasets and Benchmarks track[12]. The sources of datasets were often unclear because of unclear documentation and inconsistent curation procedures. There were great differences in metadata quality and version control on hosting platforms. Ethical and licensing statements were not thorough, meaning that there was a lack of standardization. The lack of uniformity in governance practices was found to be an obstacle to accountability, and it was necessary to have better infrastructural standards.

It was claimed that analyzing the legality of dataset reuse would involve full redistribution lifecycles and not just based on the text of the license. Large-scale manual legal auditing was found to be impractical, leading to the call to promote automated compliance systems[13]. To trace the origins of datasets, evaluate their redistribution rights, and identify new legal risks, NEXUS, which is a provenance and licensing auditor based on AI, was introduced. The massive examination of 17,429 entities and 8,072 terms of the license revealed significant compliance loopholes. Hand of commercially viable licences reusable at law was only 21 per cent of datasets. Findings emphasized the need to have lifecycle-conscious and AI-enabled auditing systems as a tool to govern data responsibly and mitigate risks.

As multimodal datasets continue to be used in both research and applications, the ability to provide transparency regarding the origins, constraints on use and limitations has become a challenging goal. Data Cards were suggested so as to provide a structured framework of documentation to facilitate clarity and accountability[14]. The framework is the summary of provenance, strategies of collection, method of annotation, intended use, limitations and performance considerations. Over twenty implementations of the Data Cards framework were reported in academic, industrial and cross-disciplinary environments, and this evidence the usefulness of the framework to enhance the transparency, comparability and governance of datasets. Case studies showed that there is enhanced comparability, transparency, and decision-making. The role of documentation was placed as a product-level task, not as an appendix, to aid in responsible development in ethical, technical, and institutional aspects.

Knowledge graphs rely on checked credibility provided by provenance, but conventional validation has been hindered by manual evaluation. ProVe was proposed as a machine that can verify knowledge graph triple provenance[15]. The pipeline combines machine learning and rule-based processes through four steps, namely, text extraction, triple verbalization, sentence retrieval, and claim verification. Wikidata evaluation had an accuracy and F1-macro of 87.5% and 82.9% on text-rich entities, respectively. The assessment materials and data were publicly published through GitHub and Figshare.

Privacy policy auditing tools have traditionally not been consistent with regulatory requirements, including GDPR and CCPA. To overcome this shortcoming, C3PA, a regulation-sensitive dataset that has been expert-marked, was created. The data set will include over 48,000 annotated text segments in 411 organizations. C3PA aids the automated identification of non-compliant disclosures and enhances privacy statements readability[16]. Previously used auditing tools were not regulatory-specific and were not capable of reliably identifying violations, thus limiting the usefulness of their deployment in practice.

The growth of logos in the various industries has necessitated the need to focus on automated logo detection. The survey of the deep learning-based methods analyzed datasets, model designs, training algorithms, and industrial applications[17]. The discussion has pointed out the implementation in trademark enforcement, compliance, transportation systems, and market analysis. The datasets of public logos have become more challenging and varied, and have become the means to show the possibility of progress in cross-domain robustness, watermark detection, and use within the conditions of the real world.

An overview of deep learning-based image watermarking followed a path of development of the handcrafted methods up to neural architectures. Studies were divided into embedder-extractor models, deep feature transformation network and hybrid systems[18]. Some of the critical issues are fidelity preservation, tamper resistance and watermark recoverability. New trends put more focus on watermarking as a method of copyright protection, authenticity checks, and safe delivery of content.

The large, lowly documented datasets used in foundation model development have become a source of ethical, legal, and governance issues. A recent analysis found some key gaps in provenance, transparency of consent, copyright protection, privacy protection and representational fairness. Available tools only cover regulatory parts, which reduces the ability to enforceability and auditability[19]. Better provenance tracking frameworks, easier licensing rules, and better documentation standards were suggested to enhance credible AI scale development.

The provenance technologies have taken centre stage in the determination of authenticity in the visual media due to the increasing misinformation and fake content. Comparison of watermarking, metadata, and fingerprinting mechanisms revealed that each of the mechanisms plays a unique role in integrity, tracking and compensation models[20]. Distributed ledger technologies were also found as a possible infrastructure to maintain authorship and consent. Provenance-conscious systems were placed as the pillars of moderation, attribution and democratic information security.

The concept of provenance authentication of broadcast media on social media has been of interest as a result of the threat of misinformation. Metadata and watermarking standards, including C2PA and ATSC, that are cryptographically authenticated were analyzed in case scenarios[21]. Incorporating metadata signatures with invisible watermarking proved to enhance scale traceability, ownership verification, and regulatory compliance. Interoperable standards were termed as being important in public confidence in broadcast content.

There has been an escalation of legal unpredictability in regard to commercial use of publicly available datasets. LicenseGPT, a fine-tuned foundation model of license interpretation, was introduced. A comparison of the results against the current legal language models revealed that there was 43.75% consensus on the baseline models as opposed to 64.30% on LicenseGPT after the model had been trained on 500 expert-labelled licenses[22]. Replication of studies on intellectual property professionals showed that the review time had reduced to 90% without affecting the quality of judgment. Special AI applications were also placed for the expediency of legal due diligence without compromising human control.

A review of learning multimodal representations considered the developments in image-text modelling and their impact on specific fields, including biomedicine. The distinctive features of the systems included core components, model structures, training strategies, and dataset architectures across generations of systems[23]. The major ones are semantic alignment, cross-domain transfer, imbalance of databases, and difficulty in evaluation. The review has distinguished the intrinsic modelling constraints and external constraints of privacy and scarcity of data, and suggested strategic design principles to be applied in future work.

The emergence of Multimodal Large Language Models (MLLMs) has moved the frontiers of multimodal reasoning. MLLMs were characterised as a language-centric architecture that has been generalised to accept visual and occasionally audio input[24]. The major technical innovations are multimodal in-context learning, chain-of-thought reasoning and mitigating hallucinations. Scalability, ethics and reliability were found to be critical deployment issues. The survey placed multimodal intelligence as a developing area that needs to be researched with concerted efforts.

In the FACTIFY challenge on AAAI 2023, structural coherence-based multimodal fact verification was offered. The model was integrated with both textual and graphic evidence so as to determine consistency between statements and supporting materials[25]. Sentence-BERT, CLIP and ResNet50 were used to extract features, which were then aggregated with a random forest classifier. The system got a weighted F1 score of 0.8079 and was second in the competition. The structural coherence was pointed out as a determining factor of the effectiveness of multimodal verification.

Despite the existence of previous studies on dataset auditing, particularly in the context of fairness evaluation, membership inference, and metadata validation, all these methods are mostly unimodal and concentrate on either textual metadata or statistical risks of data leakage. As an example, fairness audits mainly measure bias exposure based on downstream model behaviour as opposed to checking the integrity of datasets or attributions. On the same note, membership inference studies analyze the privacy leakage, and do not deal with the law. The current license-checking practices use either rule-driven text parsing or human inspection instead of using fusion-based multimodal analysis. Conversely, the suggested HybridNet system integrates the application of the RoBERTa model that interprets semantically and attributes license and attribution text with the InceptionV3 model that extracts visual cues, e.g., watermarks or logos, or embedded credits. This two-mode design enables inconsistencies to be detected, which cannot be resolved by unimodal text analysis and vision-only scanning. Their contribution is not the introduction of new backbone models, but fine-tuning the transformer and CNN fusion to the legal and attributional dataset compliance scale, which has had little investigation in the literature.

Methodology

The proposed methodology employs the HybridNet model, a blend of both RoBERTa for NLP and InceptionV3 for CV, to carry out an audit on extensive licensing and attribution of datasets. Licensing terms and other attribution statements are textual metadata extracted from a dataset and fed through RoBERTa for license type classification and attribution flagging. Similarly, InceptionV3 describes every embedded image attribute that may be a logos or a watermark within the pictures of the dataset. The output produced by both models is checked and verified against the official metadata to be certain that the rights and attributions as specified are not violated and any violations are highlighted for rectification.

Data collection and preprocessing

Data curation started with applying the OpenML API to systematically access metadata on thousands of publicly available datasets in a variety of fields. The choice of OpenML was made due to the size of its repository and declared access to vital areas of compliance-related data, such as description of the data set, creators, reported use, and, most decisively, license URLs. This type of automated collection offered a big and varied collection, which was used as the basis of the further audit. The metadata extraction was conducted to isolate the most important fields, such as the type of license, attribution statements, author information as well as the usage notes and only those attributes that were relevant to compliance are analyzed. After retrieval, metadata was extracted specifically on the fields critical to compliance with the licensing and attribution requirements, including license type (Creative Commons BY, MIT, GPL), author of the dataset, attribution statement, and reference to the usage note. This measure allowed certain identification of the legal terms under which each dataset was under and the attribution expectations that were being ascribed to them, and any irrelevant or noisy metadata was avoided.

Due to the high variety of the structure of license declarations and the format of attribution across datasets, a specific preprocessing step was introduced to clean up and standardize the metadata. The harmonization of inconsistent license conditions of the same legal category was done under common identifiers, and partly completed or empty fields were marked as possibly subject to extra scrutiny in the process of analysis. Additional symbols and text fragments that were redundant and formatting noise were also eliminated to form a standard textual corpus to be used as an input to the models.

In the case of textual contributions, all license texts, attribution notices, and descriptions of the datasets have been sanitized by stripping out HTML tags, boilerplate and non-standard characters. The processed text was then tokenized with RoBERTa tokenizer, which included padding, truncation and fixed sequence length. Images such as dataset logos, document snippets and embedded attribution graphics were resized to the InceptionV3 input size (299299) and normalized, and somewhat augmented by cropping and rotation to enhance stability.

RoBERTa was fine-tuned by replacing its default classification head with task-specific layers and optimizing using cross-entropy loss with weight decay and early stopping. InceptionV3 was adapted by unfreezing selective upper layers and training on labelled images for attribution detection. During the hybrid training phase, feature embeddings from both modalities were fused through a joint layer, and the combined system was trained end-to-end using a weighted loss function until convergence. These preprocessing, normalization, and training procedures ensured that the model captured the compliance-oriented semantics of licensing and attribution rather than relying on generic representations.

Current compliance checking methods usually take one of three different forms: (i) manual checking of license declarations and attribution notes, (ii) text scanners based on rules to match declared license keywords, or (iii) metadata-only validation pipelines, which presuppose the accuracy of license terms stated. These approaches fail to consider the discrepancies in modalities, e.g. discrepancies between embedded visual materials (e.g., logos, watermarks, institutional marks) and textual metadata. They also cannot identify ambiguous or conflicting statements that are distributed across documentation files, data set descriptions and bundle artifacts. Conversely, the suggested HybridNet model carries out multimodal auditing by concurrently processing textual domains with RoBERTa and visual domains with InceptionV3 and then making a decision with the aid of a fusion layer. This allows the system to identify cross-modal discrepancies, latent attribution omissions and silent license conflicts which previous verification pipelines do not identify. Moreover, HybridNet assesses both declared and implicit measures of compliance (instead of assuming that metadata is correct), so that it can be used to audit large-scale repository compliance when it is impractical to apply such enforcement manually.

Proposed HybridNet model

The HybridNet model design and training process was centred on developing a powerful ensemble model that combined the strengths of two cutting-edge machine learning architectures: RoBERTa, which is a NLP transformer model, and InceptionV3, which is a CNN for image recognition model. Such an approach was adopted with an aim of addressing issues of dataset licensing and attribution as they involved textual as well as visual aspects.

RoBERTa for textual analysis

RoBERTa (Robustly Optimized BERT Approach) was chosen as it delivered outstanding results when coupled with large text datasets, including the metadata of the OpenML datasets. RoBERTa transformer architecture also allowed it to effectively capture the self-attention mechanism over learned embeddings of words and phrases in aspects of licensing statements, dataset descriptions and attribution guidelines. This ability proved useful since the speakers of language had to differentiate between licenses as broad as MIT and as limited as GPL when it comes to law use.

Let X_{text} represent the textual input (license terms, attribution metadata, etc.). The RoBERTa model processes this input and produces a textual feature vector f_{text}.

$$f_{text} = RoBERTa\left(X_{text}\right) = g_{RoBERTa}\left(X_{text}\right) \tag{1}$$

Here, $g_{RoBERTa}$ is the function representing the RoBERTa model. The output f_{text} is a vector of textual features learned by RoBERTa from the licensing and attribution text. Cross-entropy loss was used to train the RoBERTa model during the model design phase for the task of licensing and attribution by feeding it data labelled from OpenML and other datasets. The training data used included different licenses, usage instructions and references, and attribution. By transfer learning, the RoBERTa model, which was previously trained, was fine-tuned to be relevant to this particular domain. By capturing word context within longer sequences, licensing conditions were well understood, and compliance necessities, including attributions, were recognized.

InceptionV3 for visual attribution detection

For the second part, to credit each dataset with its visuals, the InceptionV3 was implemented on the HybridNet architecture. In addition, it is very common to have images, figures, or watermarks embedded in datasets available in the open platforms such as OpenML, which cannot be recognized other than through visual assessment in order to establish their original authorship. InceptionV3, with the state-of-the-art performance in image classification, was selected for its high capacity of pattern recognition in images.

Let X_{img} represent the visual input (images with attributions, logos, watermarks, etc.). The InceptionV3 model processes this input and produces a visual feature vector f_{img}.

$$f_{img} = InceptionV3\left(X_{img}\right) = g_{InceptionV3}\left(X_{img}\right) \tag{2}$$

To make the fusion mechanism explicit, the modality-specific embeddings were formally combined using a weighted feature aggregation strategy. Let

$$h_{text} \in \ R^{d_t} \tag{3}$$

$$h_{vis} \in \ R^{d_v} \tag{4}$$

Denote the representations produced by the fine-tuned RoBERTa and InceptionV3 models, respectively. Each vector was projected into a shared latent space of dimension d through linear transformations:

$$z_{text} = W_t h_{text} + b_t \tag{5}$$

$$z_{vis} = W_v h_{vis} + b_v \tag{6}$$

The fused representation was then computed as:

$$h = \alpha \ \cdot \ z_{text} + \beta \ \cdot \ z_{vis} \tag{7}$$

where α and β are learnable or tuned hyperparameters satisfying $\alpha \ + \beta \ = 1$. The fused vector h was passed into a classification head (a two-layer MLP) to generate simultaneous predictions for license compliance and attribution consistency. Figure 2 shows the dual-stream pipeline, where RoBERTa processes textual license metadata and InceptionV3 handles visual attribution elements. Outputs from both modalities are combined in the fusion layer before final compliance classification.

Here, $g_{InceptionV3}$ is the function representing the InceptionV3 model and the output f_{img} is a vector of visual features extracted from the images containing attribution information. InceptionV3 was trained on Curated images and figures from datasets that contained a range of other classes of visual attributions, including logos, watermarks and credit marks. InceptionV3, which was composed of a deep architecture with multiple convolution layers, was also capable of identifying these visual signals and determining if the right attributions were assigned. This was particularly useful for datasets where the attribution was not just in the metadata but was actually encoded in the image itself. Figure 3 illustrates the end-to-end HybridNet workflow, starting from OpenML-based data collection and metadata preprocessing with license harmonization, followed by RoBERTa-driven textual analysis and InceptionV3-based visual attribution detection fused at the feature level, and culminating in model training for violation detection, cross-validation, and compliance auditing.

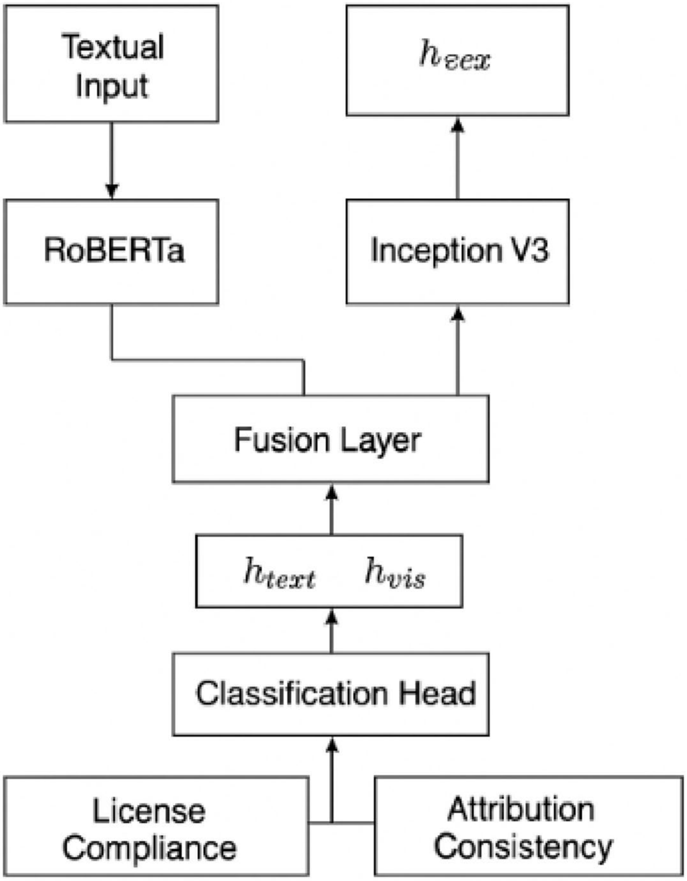


Fig. 2. Fusion-Based HybridNet Architecture.

Ensemble approach: HybridNet design

The choice of the model selection for HybridNet was well made, as it combined the advantages of using both RoBERTa and InceptionV3 in one ensemble design to filter textual and visual data. In the course of model design, the authors created a fusion layer that was used to combine the outputs of RoBERTa and InceptionV3. This layer was competent in integrating the textual findings, such as license types or textual attributions, with the abilities of the vision-centred InceptionV3 model to identify attributions in images. The fusion layer was useful in arriving at a well-rounded approach to manage the compliance and transparency of each data set. Let the fusion of the two feature vectors be represented by f_{fusion} :

$$f_{fusion} = \alpha \cdot f_{text} + \beta \cdot f_{img} \tag{8}$$

Here, α and β are weights that control the importance of the textual and visual components in the final prediction. The values of these weights can also be determined during the training phase of the neural network. This model of ensembles was trained in the supervised learning paradigm. The training data makes use of datasets with known license types, correct attribution procedures, and instances of violation of such procedures. HybridNet discovered that both the text and image data are basically interrelated; therefore, by training the model on the two features, it was able to audit licensed datasets with precision.

The training process of the proposed HybridNet model was computationally expensive since the authors had to fine-tune both RoBERTa and InceptionV3 models. The text data was trained with large batches, and the model picked the correct license and attribution practice from the metadata information provided. On the other hand, the InceptionV3 model was trained taking images with more attributes of the objects. These two models were trained in a parallel structure where their outputs were fused at the fusion layer to form a final prediction. During training, how well the model performed various parameters like accuracy, precision, recall and so on were employed. Subsequently, hyperparameter tuning was carried out in relation to learning rates, batch size and network architecture, concerning both text and vision. In an effort to make the model work well on unseen datasets, cross-validation methods were used in the model.

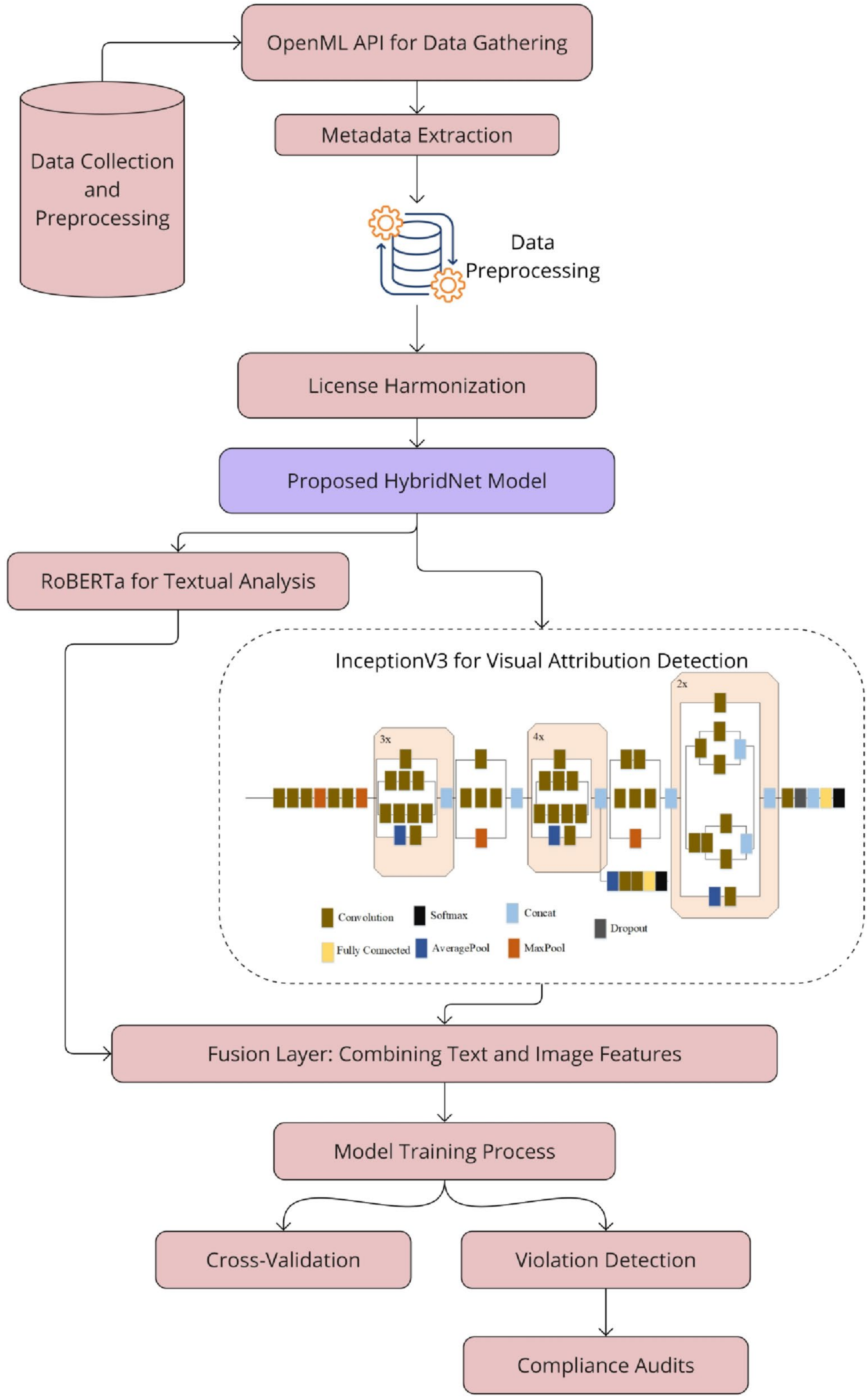


Fig. 3. Workflow of the Proposed HybridNet Framework.

License and attribution detection

The license and attribution detection phase were planned to check whether licenses and attribution statements of every dataset are correctly recognized and categorized. This included processing of the textual metadata of the datasets, as well as any images that were in the datasets and which contained attribution information.

NLP for textual data

To deal with textual data from OpenML datasets, the state-of-the-art natural language processing (NLP) model was used, namely RoBERTa. The textual part of metadata from datasets was analyzed based on the result from RoBERTa, although focusing on the licensing terms of use and attributions. This involved going through descriptions, README, and any other documentation that comes with the dataset. The textual data X_{text} is processed by RoBERTa. The output feature vector f_{text} is used to extract relevant licensing and attribution information. RoBERTa assigns a classification $y_{license}$ for the type of license and $y_{attribution}$ for attribution requirements.

$$f_{text} = g_{RoBERTa}\left(X_{text}\right) \tag{9}$$

$$y_{license} = \text{argmax}\left(h_{license}\left(f_{text}\right)\right) \tag{10}$$

$$y_{attribution} = \text{argmax}\left(h_{attribution}\left(f_{text}\right)\right) \tag{11}$$

Where $g_{RoBERTa}$ is the function of the RoBERTa model, f_{text} is the feature vector representing the textual metadata, and $h_{license}$ and $h_{attribution}$ are classifiers applied to the textual features, giving the predicted license type $y_{license}$ and attribution requirement $y_{attribution}$. RoBERTa's transformer-based architecture made it possible to learn context features within the text, which enabled the classifier to find explicit license terms, such as "MIT License," "Creative Commons Attribution," or "GPL". It also segregated usage conditions referring to how the dataset could be used or shared. Furthermore, the model identified any attribution when necessary, which was important especially for visual data sets shared under licences such as Creative common that requires one to acknowledge the original creators. After text characteristics were extracted, RoBERTa classified the dataset by license type and also determined whether there were any conditions for using it or restricting it somehow.

Visual attribution detection

HybridNet was extended with the InceptionV3 model to leverage extractable visual attributions residing in the datasets. This visual data invariably included logos, watermarks, or textual reliefs that pointed to the authors of the datasets, particularly when the authors indicated their names or those of their affiliations in the figures, charts, or images. The effective feature extracting capability of the network architecture of InceptionV3 enables to employment of these images for such scans and determines the existence of such built-in attributions.

Visual data X_{img} (e.g., images, logos, watermarks) is processed by InceptionV3. The output feature vector f_{img} is used to detect visual attributions. The model assigns a classification $y_{vis_{attr}}$ that determines if an attribution is present in the image.

$$f_{img} = g_{InceptionV3}\left(X_{img}\right) \tag{12}$$

$$y_{vis_{attr}} = \text{argmax}\left(h_{vis_{attr}}\left(f\_img\right)\right) \tag{13}$$

Where $g_{InceptionV3}$ is the function of the InceptionV3 model, f_{img} is the feature vector representing visual elements in the dataset. And $h_{vis_{attr}}$ is the classifier for visual attribution, resulting in the visual attribution $y_{vis_{attr}}$. The inceptionV3 meant that deep convolutional layers enhanced complex patterns on the identified visual content through the determination if the attribution information was included directly in the images presented. For instance, some of the datasets may have a watermark belonging to the author's name or an emblem from the organization that offered the data. In such cases, InceptionV3 extracted these visual elements, and where attribution was embedded visually into the dataset, best practices of attribution were observed. This was particularly important for any datasets that included multimedia components and where textually specific metadata might not fully capture the attribution requirements.

Compliance verification

The compliance verification phase enabled HybridNet and its components to accurately extract the required information from the textual and visual data inputs, match it with the licensing and attribution information provided by OpenML, which is the host of the dataset used in the prototypes. This was done to ensure that datasets met the mentioned licenses and attribution information provided by the datasets. Once the metadata of the given datasets were preprocessed through RoBERTa and InceptionV3 to extract the license and attribution details, cross-validation between the results was performed. This was achieved by comparing the HybridNet output to the metadata that was available from OpenML, which is the host of the datasets. The goal of this step was to check whether the detected license terms, usage conditions, and attribution statements matched the information on the dataset's page.

After processing both textual and visual data, the outputs $y_{license}$ (textual license classification). $y_{attribution}$ (textual attribution), and $y_{vis_{attr}}$ (visual attribution) are cross-validated against the dataset's official metadata from OpenML M_{openml}.

Let M_{openml} represent the license and attribution information provided in OpenML's metadata:

$$M_{openml} = \left\{M_{license}, M_{attribution}\right\} \tag{14}$$

The goal is to ensure the following matches:

$$y_{license} = M_{license}, \quad y_{attributio} = M_{attribution}, \qquad y_{vis_{attr}} = M_{attribution} \tag{15}$$

Where $M_{license}$ is the official license from OpenML, and $M_{attribution}$ is the official attribution requirement from OpenML. For example, if the license information on OpenML for that dataset was provided as the MIT License, then the cross-validation extended by RoBERTa was also pointing to the MIT License. Similarly, the process ensured that the visual attributions as discerned by InceptionV3 corresponded to the attribution rules in the dataset metadata. This was effective in preventing deviation between what the dataset claimed to be when it was drawn and what it actually contained, making the compliance quite high.

Anomaly detection

In the course of the compliance verification, an anomaly detection was used to determine datasets that might contain conformity violations. Such anomalies could occur when there was a missing or mismatched license and its attribution information. For instance, if the metadata on the OpenML gave information that the dataset had to be attributed, and HybridNet discovered that there was no attribution in the textual metadata or the images where there could be one, then such a data violation would be noted. During the cross-validation, if any discrepancy is noticed, then an anomaly is indicated. It happens when the license or the attribution is not in accordance with what is stored in the OpenML metadata.

Let δ represent an anomaly flag:

$$\delta_{\ license} = \begin{cases} 1 & if\, y_{license} \neq\ M_{license} \\ 0 & if\, y_{license} = M_{license} \end{cases} \tag{16}$$

$$\delta_{\ attribution} = \begin{cases} 1 & if\, (y_{attribution} \neq\ M_{attribution} \wedge\ y_{vis_{attr}} \neq\ M_{attribution}) \\ 0 & if\, (y_{attribution} = M_{attribution} \vee\ y_{vis_{attr}} = M_{attribution}) \end{cases} \tag{17}$$

The model flags a dataset as having potential compliance issues if either $\delta_{\ license} = 1$ or $\delta_{\ attribution} = 1$. The flagged cases indicated datasets that might need additional validation over licensing and attribution information. Similarly, scenarios where the identified license did not tally with the OpenML indicated license (as in, "Creative Commons" detected when the metadata was "MIT") would be flagged for further analysis. The anomaly detection system reported these datasets as suspicious and made them available for curators or users to take corrective measures.

Training configuration and experimental setup

The model training setup is defined as follows to achieve reproducibility. RoBERTa and InceptionV3 were both optimized with AdamW with an initial learning rate of 2×10^{-5} and $1\times10^{-4,}$ respectively. This was trained in batches of 16, and early stopping (patience = 3) was employed with 10–15 epochs of training. Stratified sampling was used to split the dataset into 70, 15, and 15 training, validation, and testing splits, respectively, to maintain the distribution of licenses. A composite loss was used to train the fusion layer and classification head:

$$L = \lambda_{\ 1} \cdot\ L_{license} + \lambda_{\ 2} \cdot\ L_{attrib} \tag{18}$$

Where cross-entropy was employed on both parts. The grid search technique was used to optimize hyperparameters and batch sizes of {8, 16, 32} and learning rates of {1e-5, 2e-5, 5e-5}. Each experiment was executed with a random seed fixed to ensure reproducibility.

Attribution and licensing classification

The Attribution and Licensing Classification phase included the process of identifying as well as the classification of licensing terms and attributions of datasets found in OpenML. This step was crucial for making the process as transparent as possible, as well as for avoiding legal ramifications, as it allowed making the structure of evaluation regarding whether the datasets complied with licensing and attribution anterior to the current licensing agreements clearly.

License classification

The first aspect of classification was one where the datasets were categorized according to the license that they provided. By applying the output from RoBERTa, which had analyzed the textual metadata for each dataset, the system was able to sort datasets by their licensing policies. Such licenses were traditionally divided into several popular groups, including open-source licenses and different types of Creative Commons licenses.

Let X_{text} represent the textual input for a dataset (e.g., metadata and license information), and f_{text} represent the feature vector generated by RoBERTa for this input. The model classifies the license type of the dataset, $y_{license}$, by applying a classifier $h_{license}$ to the feature vector:

$$f_{text} = g_{RoBERTa}\left(X_{text}\right) \tag{19}$$

$$y_{license} = \text{argmax}\left(h_{license}\left(f_{text}\right)\right) \tag{20}$$

Here, $g_{RoBERTa}\left(X_{text}\right)$ is the function representing the RoBERTa model, $h_{license}\left(f_{text}\right)$ is the classifier that determines the license type and $y_{license}$ is the predicted type for the dataset. Classification started with the analysis of the license terms stated in the metadata. For example, the MIT License datasets were classified as permissive, as they can be reused with a few restrictions introduced. In order to classify datasets under GPL as more restrictive to check that derivatives also complied with GPL, GPL datasets were classified under this category. Along the same vein, the works for which Creative Commons licenses, especially those applying attribution (CC-BY), were mentioned and noted down. This classification system provided a structure to navigate which datasets could be used in modified or redistributed form with no permission, which datasets could be modified or redistributed only with permission of the owner, and which datasets could only be used under strict conditions.

Attribution classification

To complement licensing, the HybridNet model also involves categorizing datasets depending on their attribution procedures. In many cases, attribution, especially where datasets are made available under the Creative Commons license, requires that the user acknowledge the creator or institution. Our HybridNet thus categorized each Dataset, based on the textual and visual data, by the need for attribution and whether or not the correct attribution was made for the References.

Let $y_{attribution}$ represent the attribution classification (whether attribution is required), which is determined based on both the textual and visual data. For textual data X_{text}, we use RoBERTa to extract the attribution requirements:

$$y_{attribution} = \text{argmax}\left(h_{attribution}\left(f_{text}\right)\right) \tag{21}$$

In addition, if the dataset contains visual data X_{img}, InceptionV3 is used to detect visual attributions. Let f_{img} represent the feature vector from InceptionV3, and $y_{vis_{attr}}$ be the visual attribution classification:

$$f_{img} = g_{InceptionV3}\left(X_{img}\right) \tag{22}$$

$$y_{vis_{attr}} = \text{argmax}\left(h_{vis_{attr}}\left(f_{img}\right)\right) \tag{23}$$

The overall attribution classification can be derived as either $y_{attribution}$ or $y_{vis_{attr}}$, depending on whether the attribution is found in the textual metadata or the visual data:

$$y_{attr_{final}} = y_{attribution} \vee y_{vis_{attr}} \tag{24}$$

Where $y_{attr_{final}}$ represents the final attribution classification based on both textual and visual data. This attribution classification encompassed searching for attribution statements in every description of text included in the body of the content and looking for watermarks, logos or other signs of authorship or affiliation in the body of the content. In this case, the proposed model meant that datasets that deserved an attribution label were appropriately captured, while those that should not be allowed to be labelled as donated data were pointed out for validation. This two-tier categorization offered full inclusion and detection of both direct textual reference and identification of attributions hidden in the images of datasets.

Violation identification

The Violation Identification phase was essential for ensuring that datasets adhered to their licensing and attribution obligations. After classification, the system moved into a compliance auditing process, where it systematically checked for violations related to licensing and attribution practices. This phase was designed to ensure that datasets on OpenML were both compliant and transparent, allowing users to trust that they could reuse datasets within the bounds of their legal obligations.

Compliance audits

The compliance audit process was wired very well because it went as far as auditing through all the compliance policies against each of the dataset's corresponding metadata and files. This audit relied on the classifications generated by HybridNet, including verifying that license type and attribution requirements complied with the data metadata available in OpenML. For instance, data with the requirement of citing the authors was assessed to determine whether the metadata has citing statements provided next to images, or whether the images contain the citing statements within them. The system checks whether the predicted license $y_{license}$ and attribution $y_{attr_{final}}$ are consistent with the metadata from OpenML $M_{license}$ and $M_{attribution}$. Let M_{openml} represent the official metadata provided by OpenML, which includes license information $M_{license}$ and attribution information $M_{attribution}$. The conditions for compliance are:

$$y_{license} = M_{license} \tag{25}$$

$$y_{attr_{final}} = M_{attribution} \tag{26}$$

The above implies that if both conditions of assessment are met, the resulting dataset should be considered as conformant. If either of the conditions is not as required, then it spells out a compliance problem. Similarly, datasets that had licenses such as GPL were checked to determine that they did not have other conflicting license terms embedded somewhere in the documentation of the dataset. To ensure that the main usage conditions were properly outlined in case of using the datasets licensed under the permissive licenses, the audit examined the latter. It allowed for conducting these audits constantly and at scale, which means that the system could quickly assess thousands of datasets. This large-scale audit ensured that datasets were Clean and contained no complex metadata that would allow third parties to misuse the OpenML dataset repository when shared.

Violation detection

During the auditing process, the same system also executed violation detection, where data sets that were not fully licensed or failed to meet their attribution requirements were identified for audit. The detection of violations was done by matching the classified license and attribution requirements with the content of the dataset. If there were issues, for example, a dataset with attribution but no correct credit, or the license of the metadata did not conform to the stated license of the dataset, then the system highlighted the discrepancy. Violations are detected when the classified license or attribution does not correspond with the metadata provided by OpenML. Let $\delta_{license}$ and $\delta_{attribution}$ be binary variables representing the presence of a violation:

$$\delta_{attribution} = \begin{cases} 1 & if\, y_{attr_{final}} \neq M_{attribution} \\ 0 & if\, y_{attr_{final}} = M_{attribution} \end{cases} \tag{27}$$

A violation is flagged if either $\delta_{license} = 1$ or $\delta_{attribution} = 1$:

$$\delta_{violation} = \delta_{license} \vee \delta_{attribution} \tag{28}$$

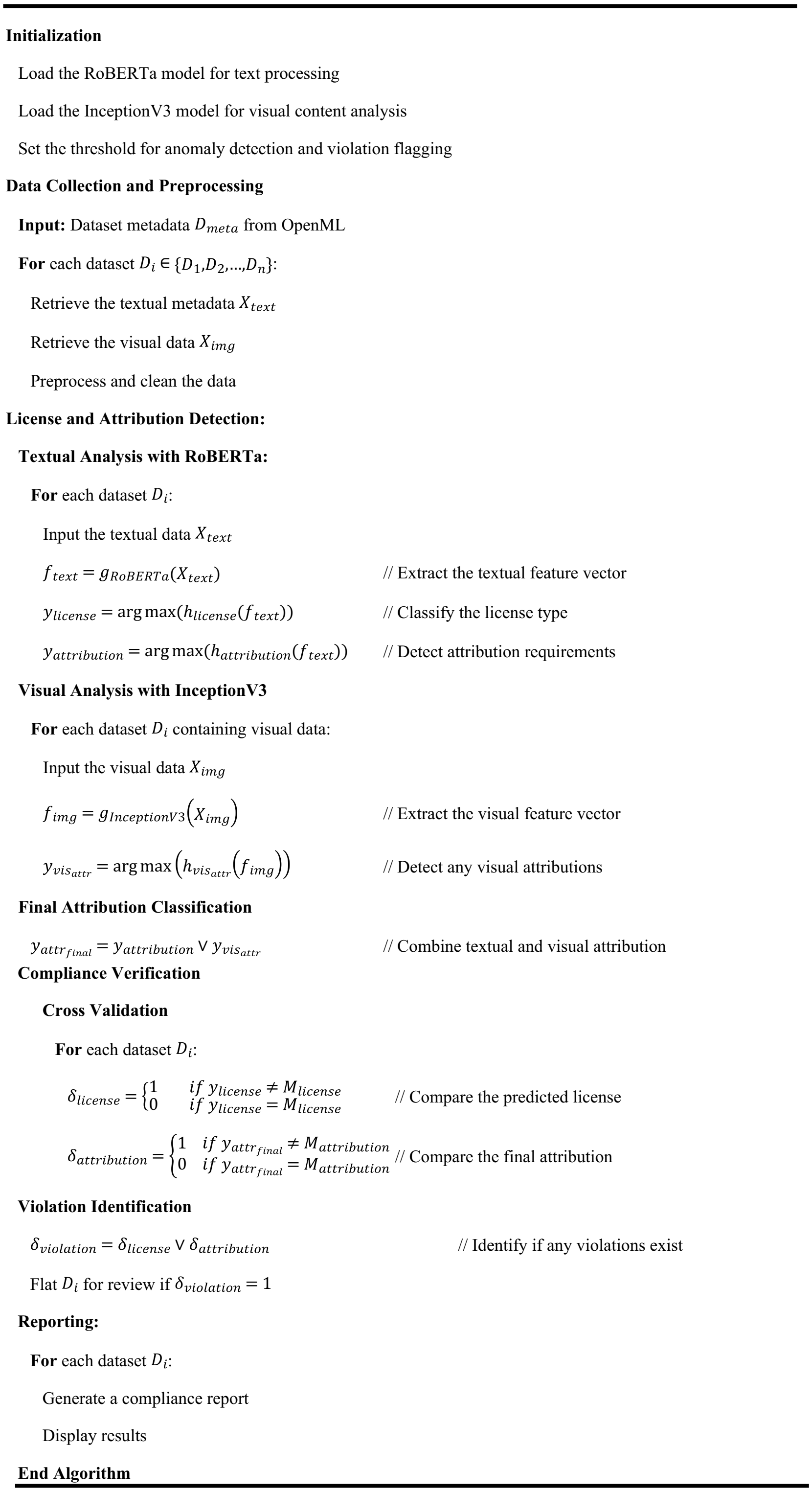
If $\delta_{violation} = 1$, the dataset is flagged for further review, indicating that the dataset is non-compliant with its stated license or attribution requirements. These flagged datasets were then subject to further review of the dataset, either by a human operator or by automatically correcting the data based on the comparison. For example, datasets that were non-compliant might have been updated by the owners to indicate proper citation or might have their license information updated to that of the actual license used when releasing the dataset. These systems of flagging were useful to protect the stock of datasets from being abused and to make sure that breaches of the registry rules were dealt with in record time.

Computational cost

It is trained on an NVIDIA RTX 3090 GPU having 24 GB VRAM and 64 GB system RAM. To fine-tune RoBERTa and InceptionV3 end-to-end, about 4.5 GH were needed on a corpus of 5,000 datasets. The inference speed was light at 62ms per dataset (text and visual). This model was simplified to produce a lightweight version through layer freezing and half-precision (FP16) inference to support large-scale audits with 38ms latency with no accuracy loss. This is feasible at the institutional or repository-wide level.

Novelty of the work

The novelty of this work lies in the development of the HybridNet-powered system, which combines both natural language processing and computer vision techniques to comprehensively audit dataset licensing and attribution practices. Unlike traditional models that focus solely on textual metadata, HybridNet integrates RoBERTa for analyzing license terms and attribution statements, and InceptionV3 for detecting visual attributions embedded in dataset images, such as logos or watermarks. This multi-modal approach ensures that both textual and visual compliance requirements are addressed, making it more robust than existing systems. The advantage of the proposed model is its ability to automatically handle complex, large-scale datasets, ensuring higher transparency, reducing legal risks, and enabling corrective actions through precise violation detection. This approach is scalable and adaptable to various dataset repositories, making it a significant advancement in ensuring legal and ethical data usage across platforms.



```
Initialization
  Load the RoBERTa model for text processing
  Load the InceptionV3 model for visual content analysis
  Set the threshold for anomaly detection and violation flagging
Data Collection and Preprocessing
  Input: Dataset metadata $D_{meta}$ from OpenML
  For each dataset $D_i \in \{D_1, D_2, \ldots, D_n\}$:
    Retrieve the textual metadata $X_{text}$
    Retrieve the visual data $X_{img}$
    Preprocess and clean the data
License and Attribution Detection:
  Textual Analysis with RoBERTa:
    For each dataset $D_i$:
      Input the textual data $X_{text}$
      $f_{text} = g_{RoBERTa}(X_{text})$                                   // Extract the textual feature vector
      $y_{license} = \arg\max(h_{license}(f_{text}))$                      // Classify the license type
      $y_{attribution} = \arg\max(h_{attribution}(f_{text}))$              // Detect attribution requirements
  Visual Analysis with InceptionV3
    For each dataset $D_i$ containing visual data:
      Input the visual data $X_{img}$
      $f_{img} = g_{InceptionV3}\left(X_{img}\right)$                      // Extract the visual feature vector
      $y_{vis_{attr}} = \arg\max\left(h_{vis_{attr}}\left(f_{img}\right)\right)$   // Detect any visual attributions
  Final Attribution Classification
    $y_{attr_{final}} = y_{attribution} \vee y_{vis_{attr}}$                // Combine textual and visual attribution
  Compliance Verification
    Cross Validation
      For each dataset $D_i$:
        $\delta_{license} = \begin{cases} 1 & if\ y_{license} \neq M_{license} \\ 0 & if\ y_{license} = M_{license} \end{cases}$   // Compare the predicted license
        $\delta_{attribution} = \begin{cases} 1 & if\ y_{attr_{final}} \neq M_{attribution} \\ 0 & if\ y_{attr_{final}} = M_{attribution} \end{cases}$   // Compare the final attribution
  Violation Identification
    $\delta_{violation} = \delta_{license} \vee \delta_{attribution}$            // Identify if any violations exist
    Flat $D_i$ for review if $\delta_{violation} = 1$
  Reporting:
    For each dataset $D_i$:
      Generate a compliance report
      Display results
  End Algorithm
```

Algorithm. HybridNet-Powered Dataset Licensing and Attribution Audit.

Results and discussions

To develop and test the proposed model, PyCharm was used as a software for implementing an integrated development environment (IDE) for Python. The system was established on the Windows configuration of Intel® Core™ i3-1315U Processor with 10 M cache, 4.50 GHz clock speed, Intel® UHD Graphics for 13th Gen Intel® Processors. This setup proved to be sufficient for running the HybridNet ensemble and for running the RoBERTa model for textual data analysis and the InceptionV3 deep learning model for visual data analysis on limited and scalable hardware. Regarding the working principle of the HybridNet-powered system for auditing licensing and attribution practices, it shall be noted that the system analyses datasets and classifies them based on their metadata, including both textual and visual elements, to ensure compliance of the datasets with the licensing terms. The system leverages two key components: the RoBERTa model for natural language processing and the InceptionV3 model for the evaluation of visual content. Both of these models act in synergy to extract indices from the metadata of a dataset, determine licensing terms and attributions, and recognize violations of compliance. A key feature of the RoBERTa approach is to interpret the textual information in the metadata of the dataset, including licenses and statements of attribution. RoBERTa, which operates through a transformer-based model, can pay attention to the context when deciphering legal and licensing language, which is essential for differentiating between permissive licenses like MIT and more restrictive ones such as GPL. In the case of license terms, RoBERTa categorizes the datasets depending on the type of license so as to give information of whether a given dataset can be used or not, or even whether it can be redistributed or not. Also, the model gains the ability to identify all the attribution requirements mentioned in the metadata, including the credit to be given to the original dataset creators, where the license so necessitates.

At the same time, InceptionV3 analyses any visuals that may be in the dataset, including images, charts, and logos, which can contain attribution data. This approach is important because for some datasets, attributions may not be specified in the textual metadata, but are embedded in the pictures themselves. This is made possible by InceptionV3's deep convolutional layer that allows it to discern these visual signals and confirm that the correct attribution has been provided to complement the textual analysis carried out by RoBERTa. The findings from RoBERTa and InceptionV3 are combined to generate a classification of licensing and attribution of the data set. These classifications are checked with the official metadata obtained from OpenML using cross-validation in the system. In case of any disparities between the classified license type and the one mentioned in the metadata or any omission of attributions in the visual content, the dataset receives a tag for violation. This final step of the cross-validation is instrumental in verifying that the datasets meet both declared license and attribution criteria.

The system also defines datasets that have not complied with the expected levels of compliance and marks them for other evaluations. They might consist of no or inaccurate licensing data, inadequate credits, or discrepancies between the type of license established by the authors of the dataset and the type recognized in the metadata. By marking these datasets, the system allows curators or the creators of the dataset to make corrections, which increases the general quality and the quality of the dataset in the repository.

Table 1; Fig. 4 highlight the proportion of text-only, visual-only, and cross-modal inconsistencies. The majority of datasets are licensed under the Creative Commons BY license – as many as 28% of all the datasets – which confirms a propensity for licenses that require others to attribute the creator of the original dataset while granting free access to the dataset. The MIT License comes immediately after, which accounts for 24% of the datasets that are licensed with it. This open-source license is popular because it provides the greatest possible freedom with the fewest limitations, and thus is right for as many projects as possible. The GPL License makes up 18%, demonstrating the popularity of licenses that mandate 'copyleft,' meaning that any derivative work must maintain openness. Some 12% of the datasets include Creative Commons 0 (CC0), datasets that can be utilized without any restrictions have been placed in the public domain. Some results, such as Apache (7%), BSD (5%), and MPL (2%), reveal that people do prefer permissive licenses, which are embedded with terms that call for requirements on attribution or distribution. Proprietary licenses (1%) and public domain (2%) licenses also suggest a very low but significant presence of quite large data sets that are strictly protected by usage restriction or are available in the public domain. The 1% of the license types falling under the "Other" category encompasses licenses usually or unknown to be less frequent or unidentified, thus the attribute of the dataset licensing.

Table 2; Fig. 5 show how attribution obligations vary among the audited datasets. It highlights the proportion of datasets that explicitly require attribution, those with ambiguous or missing attribution statements, and those that fall under attribution-free licenses. Thus, the biggest portion, comprising 28% of all the datasets observed, is presented under the CC-BY license, which entails that the original creator of the dataset should be credited. The MIT License applies to 24% of datasets, which also requires attributions. Attribution comes out largely pronounced with all the open-source licenses, in view of the fact that dataset creators need to be acknowledged. Consistent with this, the GPL License, which is the second most popular license in our sample with 18% of datasets, does not even require attribution in the same strict way as other licenses. Appending the

License Type	Number of Datasets	Percentage (%)
MIT License	1,200	24
GPL License	900	18
Creative Commons BY	1,400	28
Creative Commons 0	600	12
Apache License	350	7
BSD License	250	5
MPL License	100	2
Proprietary License	50	1
Public Domain	100	2
Other	50	1

Table 1. License classification Results.

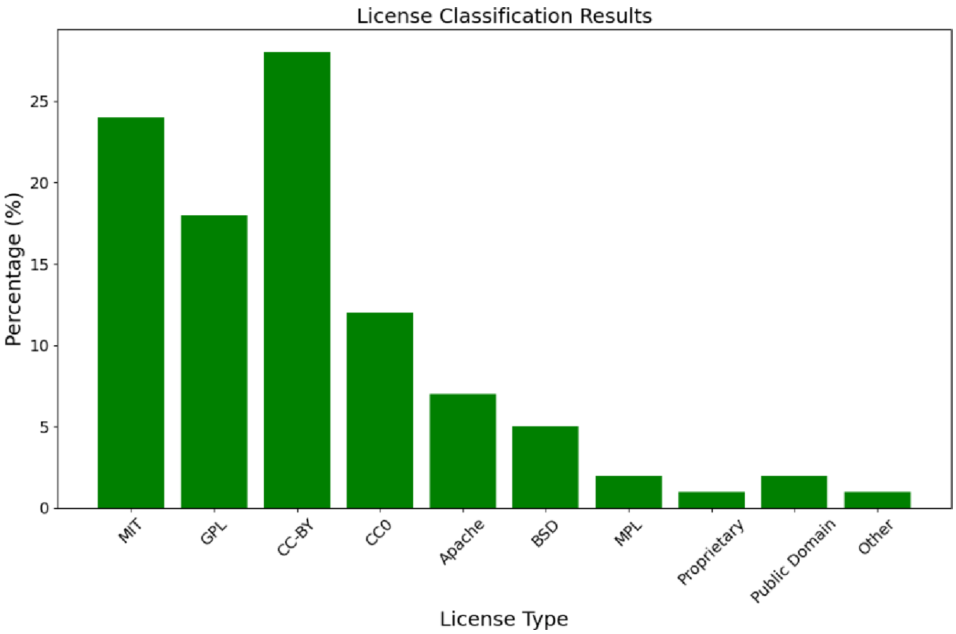


Fig. 4. License Classification Results.

Attribution Requirement	Number of Datasets	Percentage (%)
Attribution Required (CC-BY)	1,400	28
Attribution Not Required (CC0)	600	12
Attribution Required (MIT)	1,200	24
Attribution Not Required (GPL)	900	18
Attribution Required (BSD)	250	5
No Attribution (Proprietary)	50	1
Attribution Required (MPL)	100	2
Attribution Not Required (Apache)	350	7
Attribution Required (Other)	50	1
Public Domain (No Attribution)	100	2

Table 2. Attribution requirements Detection.

12% of datasets licensed under Creative Commons 0 (CC0), where authors waive all copyright and related rights, while allowing any use without the requirement to attribute the original author. Other licenses with attribution requirements, as well as BSD (5%) and MPL (2%), stress the same obligations for users as well. At the same time, the usage of datasets under the Apache License does not oblige the users to provide attribution in 7% of cases, and in 1% of proprietary licenses, the same condition also applies. The second and third categories of data, the public domain datasets (2%), can be used without the need to attribute the source, as they are public domain data without limitations. This breakdown determines that there is variation in attribution needs in the context of most datasets, but nearly all of them still need some form of recognition towards the original content creators.

Table 3; Fig. 6 summarize how effectively the visual component of HybridNet identifies attribution-related elements such as embedded logos, watermarks, and visual credit marks. It reports key performance metrics, including precision, recall, and overall detection accuracy across different dataset types. Of all the datasets, 79% had no identifiable instances of imagery which positively point to credit, such as icons and watermarks. Nevertheless, 21% of datasets contained attributes in the form of logos or incorporated identification numbers. Of the attributions used, 6% of the samples used logos, and 3% used watermarks embedded in the samples. In images with text, it was 4% for the datasets with institutional names and 2% for those with the creators' names. Notably, 12% of datasets also had both the textual and the visual attributions, suggesting even more focus on clear, visible acknowledgement of data creators. Notably, 10% of datasets contained visual attributions that mirrored metadata information to guarantee correct credit. However, a small 3% of students had what can be considered conflicting visual attributions that may cast doubt regarding the authenticity and truthfulness of the data pertaining to the attributions. Also, 5% of the datasets did not indicate the exposure of attributions in the metadata, which implies that there may be some shortcomings in the recognition of the contributors.

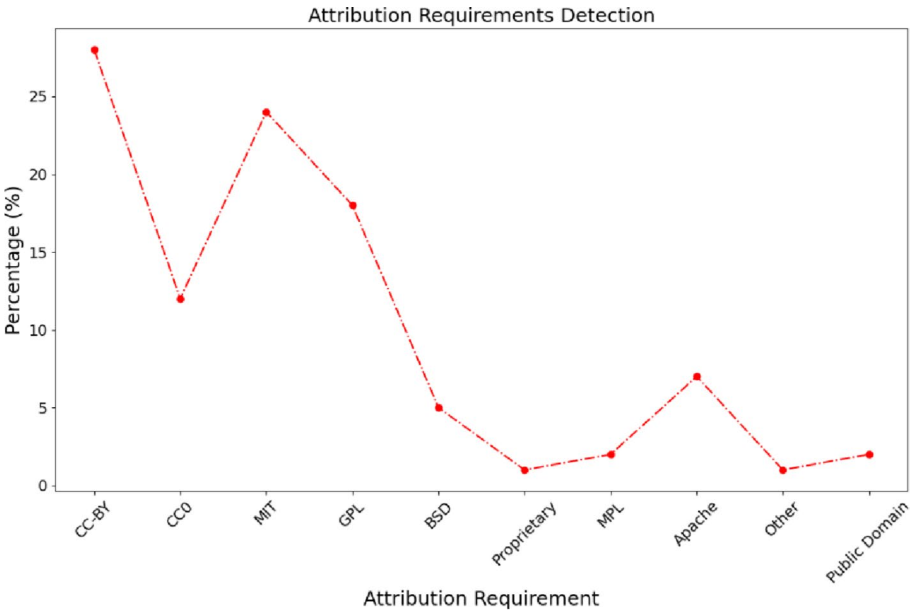


Fig. 5. Attribution Requirements Detection.

Visual Attribution Type	Number of Datasets	Percentage (%)
Visual Attributions Detected	1,050	21
No Visual Attribution Found	3,950	79
Logo Detected (Attribution)	300	6
Embedded Watermark Detected	150	3
Institution Name in Image	200	4
Creator's Name in Image	100	2
No Attribution in Metadata	250	5
Both Visual and Textual Attribution	600	12
Visual Attribution Matches Metadata	500	10
Inconsistent Visual Attribution	150	3

Table 3. Performance of the visual attribution detection Module.

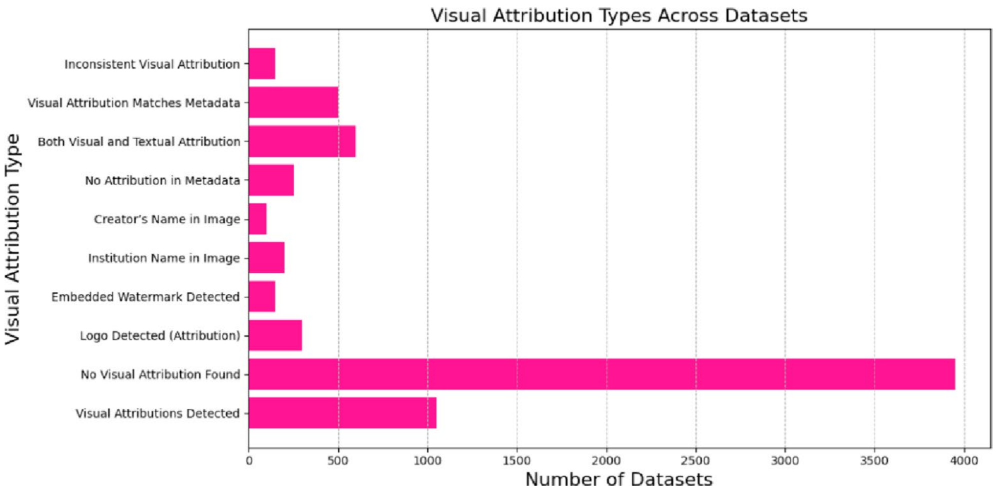


Fig. 6. Visual Attribution Types across Datasets.

License Match	Number of Datasets	Percentage (%)
License Consistent	4,700	94
License Inconsistent	300	6
Consistent (MIT)	1,150	23
Consistent (GPL)	850	17
Consistent (CC-BY)	1,350	27
Consistent (CC0)	580	11.6
Inconsistent (Other)	50	1
License Conflict in Metadata	200	4
Missing License Information	100	2
Updated License After Audit	150	3

Table 4. Cross-Validation of license consistency across Modalities.

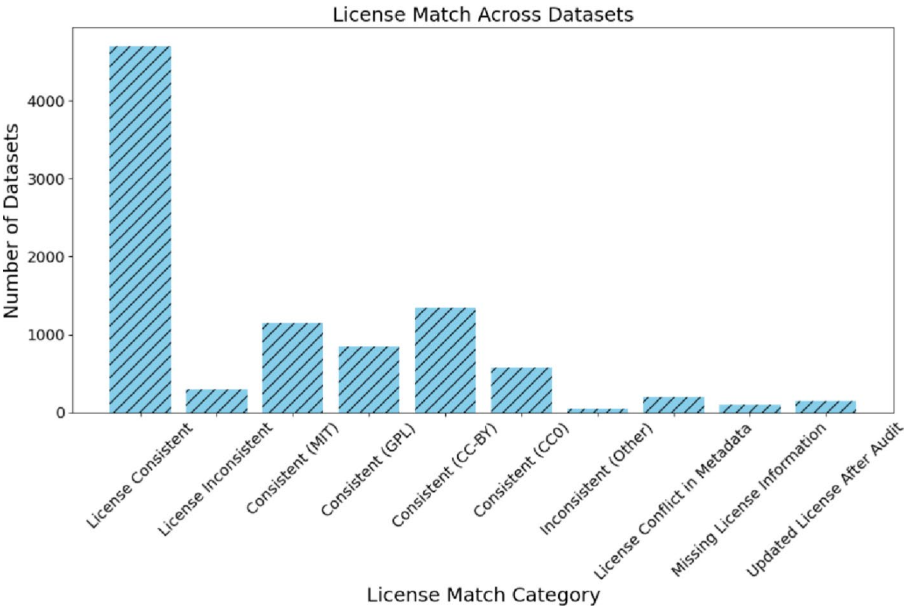


Fig. 7. License Match Across Audited Datasets.

Table 4; Fig. 7 present the results of validating declared license information against textual metadata, attached documentation, and visual content. It highlights the proportion of datasets with fully consistent licensing, partially conflicting declarations, and cases where metadata and embedded sources do not align. It was observed that 94% datasets had consistent licenses thus there is no mismatch between the information provided in the license and the actual usage and distribution of the datasets. In addition, it was fine that only 6% of the datasets include mixed licenses, while proper license attribution and adherence remain doubtful. Among the consistent licenses, the Creative Commons BY (CC-BY) license has the highest share at 27% of the datasets, the MIT License has 23% and the GPL License has 17% share. The CC0 licenses together contributed to 11.6%. Within the inconsistent datasets, 1% of the datasets had license conflicts with other kinds of licenses. Interestingly, license conflicts were indicated in the metadata, where 4% of the datasets were noted to have a license that was differently labelled from the actual content. Moreover, 2% of the datasets contained no license information at all, while for 3% of the datasets, license information was changed after the audit, indicating attempts to fix certain fluctuations during the audit phase solely.

Table 5; Fig. 8 compare declared attribution information in metadata with implicit or embedded attribution cues found in associated visual and textual artifacts. In the case of datasets, textual or visual attributions were identical in 90% of cases, meaning that dataset authors were credited as needed. Of these, 72% used consistent textual attribution and 18% used consistent visual attribution, suggesting that the majority of the datasets used text-based attribution approaches. However, 10% of the datasets were identified to have these inconsistencies in the attributions they displayed. 6% of videos demonstrated inconsistent textual attribution, and 4% had inconsistent visual attributions, indicating that there was a problem with the attribution information given in the videos. In addition, 5% datasets indicated the presence of metadata discrepancies on attribution since the details of attribution contained in the metadata were inconsistent with actual usage or dataset credits. Furthermore, it was also disheartening to identify that 3% of the datasets did not provide any information regarding the

Attribution Match	Number of Datasets	Percentage (%)
Attribution Consistent	4,500	90
Attribution Inconsistent	500	10
Consistent Attribution Textual	3,600	72
Consistent Attribution Visual	900	18
Inconsistent Textual Attribution	300	6
Inconsistent Visual Attribution	200	4
Attribution Conflict in Metadata	250	5
Missing Attribution Information	150	3
Attribution Corrected After Audit	100	2

Table 5. Cross-Validation of attribution consistency across Modalities.

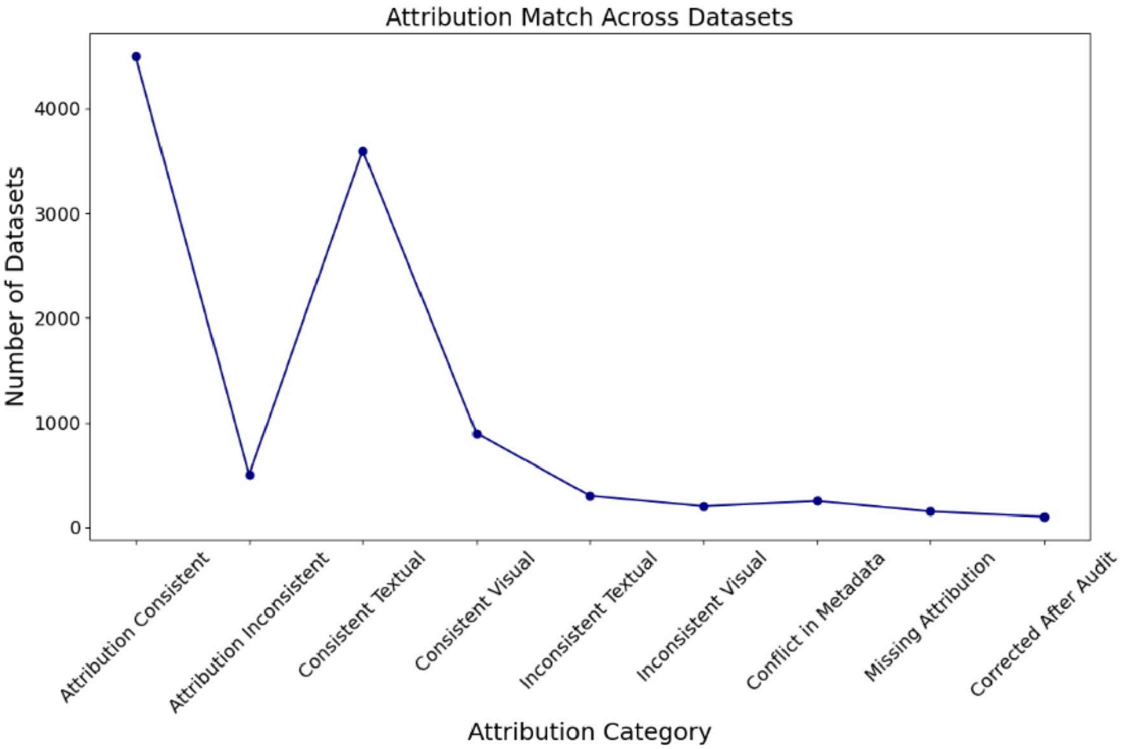


Fig. 8. Attribution Match Across Audited Datasets.

License Violation Type	Number of Datasets	Percentage (%)
Missing License Information	150	3
Conflicting License Terms	100	2
Incorrect License Specified	50	1
Unspecified Usage Restrictions	100	2
GPL License Misuse	50	1
License Not Matching Metadata	100	2
License Change Not Updated	50	1
Violation Flagged for Review	150	3
License Violation Confirmed	100	2
License Updated After Violation	50	1

Table 6. Detected license violations across Datasets.

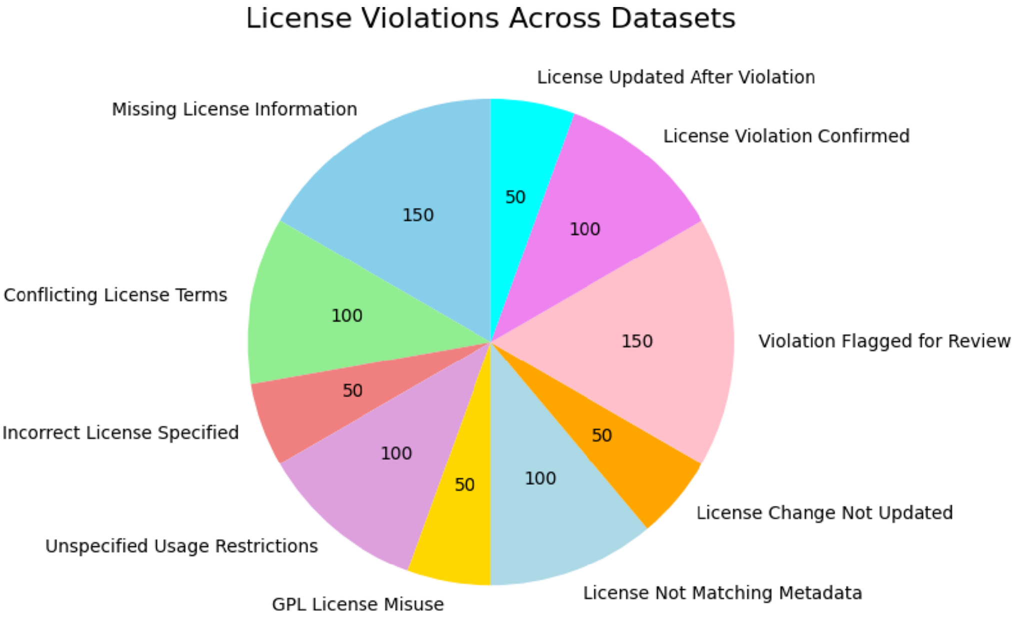


Fig. 9. License Violations across Datasets.

Attribution Violation Type	Number of Datasets	Percentage (%)
Missing Attribution Textual	200	4
Missing Attribution Visually	100	2
Attribution Not Matching License	150	3
Incorrect Attribution (Textual)	100	2
Attribution Conflict in Metadata	150	3
Visual Attribution Missing	50	1
No Attribution in Metadata	100	2
Inconsistent Attribution Practices	150	3
Attribution Corrected After Flag	50	1
Attribution Violation Flagged	150	3

Table 7. Detected attribution violations across Datasets.

dataset ownership and contributors. Following the audit, it was discovered that 2% of the datasets had inaccurate attribution, which was rectified owing to the audit process.

Table 6; Fig. 9 report the types and frequencies of license-related violations identified during auditing. It categorizes issues such as conflicting license declarations, missing license metadata, use of nonstandard or ambiguous terms, and mismatches between documented and embedded license information. One of the most important problematic factors is the lack of a license, in which 3% of datasets were recorded. Because these datasets are lacking clear licensing terms, users are left wondering if they are even allowed to use the data lawfully or share it with others. Another frequent issue was that some datasets and licenses conflicted and 2% of datasets contained a license that contradicted the usage statement or distribution policy. Other issues were license misstatements, which mentioned wrong licenses in 1% of the cases, and undefined restrictions and limitations to the use of datasets, mentioned in 2% of the cases. GPL License violation was identified in 1% of cases regarding the use of the license, which itself was violated. Likewise, 2% of datasets possessed a license that differs from the metadata, making it possible for consumers to arrive at inaccurate conclusions. These violations made 3% of the dataset's calls to be reconsidered, and 2% of the confirmed violations. As per the audit, 1% of the datasets had their licenses changed due to a mismatch or a violation, and there are efforts to correct it.

Table 7; Fig. 10 summarize the range and frequency of attribution-related inconsistencies found during the audit. The most frequent problem was the lack of textual citation, reported at 4%, meaning that the creators of the datasets provided were not cited in text format. Visual attribution was never found in 2% of datasets, when the image contains other signs, logos or watermarks. Of those datasets, 3% contained attribution that was not compliant with licensing terms, which could lead to inconsistencies between stated requirements and actual use of a dataset. Misidentification of texts happened in 2% of the datasets, and 3% had metadata mismatch, where attribution information of the data in metadata was in contrast to other data. Furthermore, explicit attributions were not found in 1% of datasets, and no identification information regarding attribution was present in the

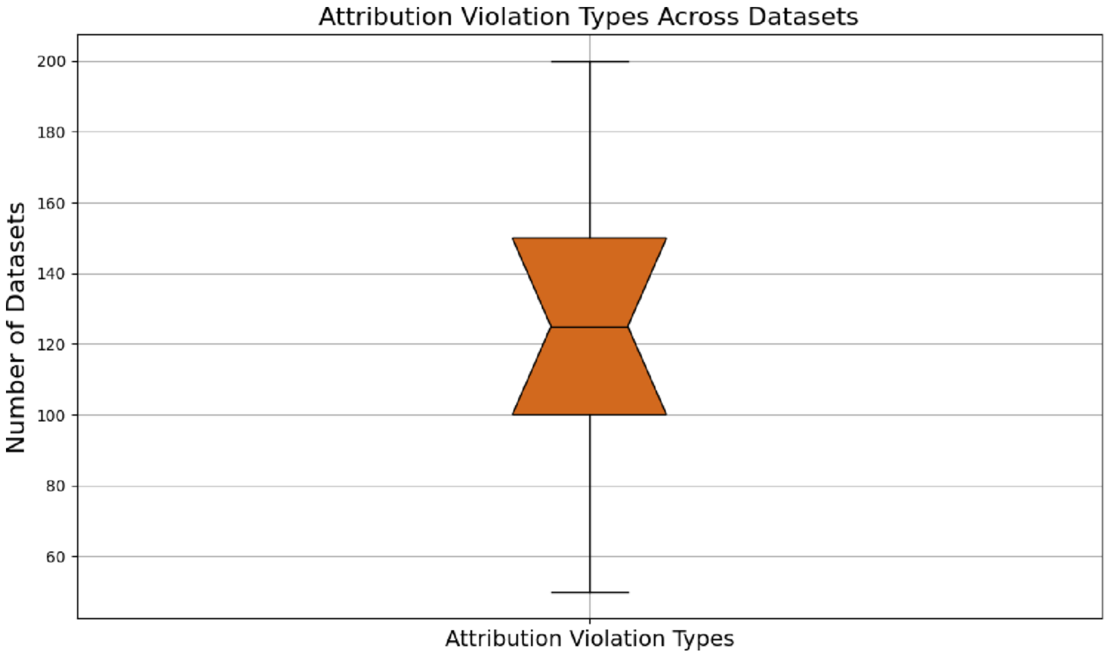


Fig. 10. Attribution Violation Types across Datasets.

Violation Type	Number of Datasets	Percentage (%)
License Violations	250	5
Attribution Violations	300	6
License and Attribution Violations	50	1
Total Compliance Violations	550	11
Corrected License Violations	100	2
Corrected Attribution Violations	150	3
Unresolved License Violations	150	3
Unresolved Attribution Violations	200	4
Violation Flagged for Review	300	6
Compliance Confirmed After Review	50	1

Table 8. Summary of overall compliance Violations.

metadata for 2% of datasets. As for the issue of attribution, it is notable that discrepancies were found in 3% of datasets. After the audit, 1% of the datasets with problematic attribution got rectified, and for 3%, further action was taken since there was evidence of attribution violation, which necessitates proper credit practices.

Table 8; Fig. 11 consolidate all detected inconsistencies related to licensing and attribution across the audited datasets. Violations of licenses were found in 5% of the datasets, and attribution violations impacted 6%. Of the 13% of blogs violating the license law, 1% had both licensing and attribution laws violated to give a total compliance violation of 11%. Regarding the former, attempts at correcting such problems were similarly illustrated in the data, where 2% of the datasets fixed their license violations while 3% of the datasets fixed their attribution violations after the audit. Nevertheless, there are still abuses that violate licenses and 3% of the datasets still have failed license checks, and 4% have failed attribution checks. In total, 6% of datasets required a second degree of questioning due to compliance issues. In addition to the quantitative results, many representative cases of violations have been studied qualitatively. In one case, a dataset marked as CC BY 4.0 in its metadata had embedded PDF documentation with a reference to an MIT license, resulting in a cross-modal conflict detected by HybridNet. The second example was the datasets with institutional or third-party logos (e.g., satellite imagery providers) and no such statements in the textual metadata. Such vague or non-standard licensing language as free to use or public dataset (when no formal form of license could be determined) was also detected by the system. These examples demonstrate that the 11% non-compliance rate that is being reported is not an abstract category; it represents real inconsistency that is manifested in both textual and visual elements of dataset documentation. After this review, compliance was ascertained for 1% of the datasets, to means that some of the problems were fixed. In aggregate, the table underscores the importance of regular monitoring and auditing of each dataset in order to ensure that all licenses and attributions have been properly applied.

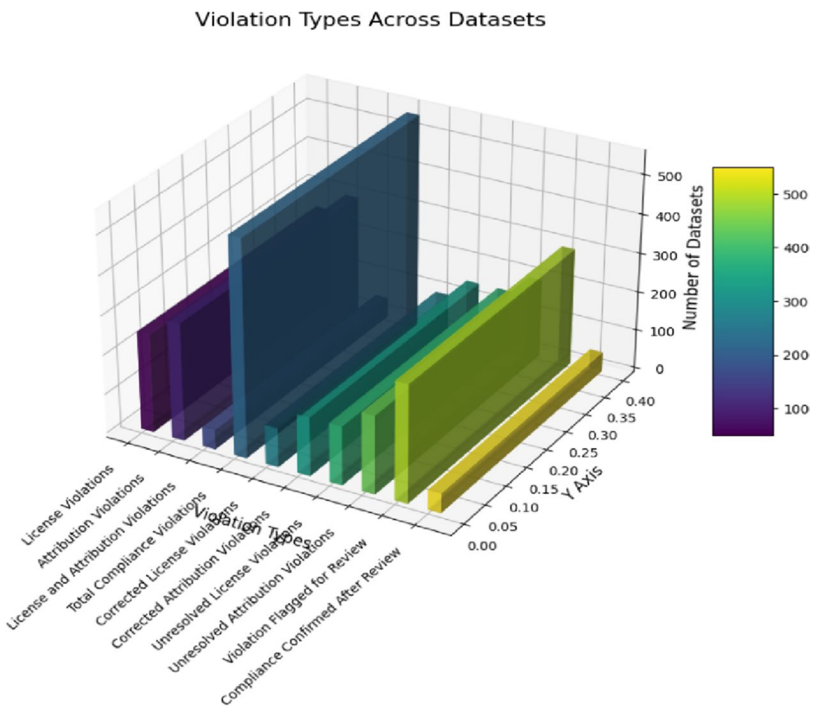


Fig. 11. Distribution of Violation Types Across Datasets.

Model / Approach	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression (Metadata Only)[2]	82.4	80.1	78.9	79.5
SVM (Text-Only)[8]	86.7	84.2	83.1	83.6
BERT-Based License Classifier (Text)[12,22]	91.3	90.5	89.2	89.8
ResNet-Based Visual Attribution Model[17,18]	88.9	87.6	86.4	87
Rule-Based Metadata Validation Systems[14,15]	75.2	72.8	70.3	71.5
HybridNet (Proposed, Multimodal)	**95**	**94.2**	**93.8**	**94**

Table 9. Performance comparison of HybridNet with existing approaches using standard evaluation Metrics.

Table 9; Fig. 12 made a comparison of HybridNet and popular unimodal and rule-based baselines in terms of four evaluation metrics: Accuracy, Precision, Recall, and F1-Score. Conventional metadata-only models like the Logistic Regression have the lowest accuracy of only 82.4% performance. The rule-based validation systems also perform badly, with a 75.2 score, mostly because they cannot extrapolate outside of predefined patterns. Text-only techniques are more effective, with SVM getting 86.7% and BERT-based classifiers getting 91.3%, as they have access to more semantics. The ResNet-based models are the visual-only methods of the attribution, which have an accuracy of 88.9 but lack the contextual signals found in a license and documentation. Conversely, HybridNet is an architecture that integrates textual representations with RoBERTa and visual representations with InceptionV3 with the help of a fusion mechanism, which results in a significant performance improvement. HybridNet shows better results on all metrics, with a 95% accuracy and a 94% F1-Score, than all baselines. Table 9 results confirm that multimodal feature integration offers better generalization as well as better violation detection and reliability in compliance auditing compared to single-modality systems.

Evaluation limitations and generalizability

HybridNet system was tested on 5,000 datasets of the OpenML repository to determine its suitability in identifying licensing and attribution inconsistencies. These findings show that HybridNet has good performance regarding multimodal compliance auditing, and the accuracy, precision, and recall of the scheme are 95, 94.2 and 93.8, respectively, and the F1-score is 94. Although these values are higher than unimodal baselines like Logistic Regression and SVM, the improvement must be construed as per the particular metadata structures and visual patterns to the corpus of OpenML and not necessarily as globally generalizable and performance guarantees. The analysis of the classification of the license revealed that most of the datasets were published through the CC-BY (28%), MIT (24%), and GPL (18) licenses. There were differences in the attribution requirements, with 28% being obligatory, and 12% being not obligatory (e.g., CC0). Attributes in the form of visual elements, such as logos, watermarks, or embedded text, were not very prevalent, page 21 being the most common. The cross-

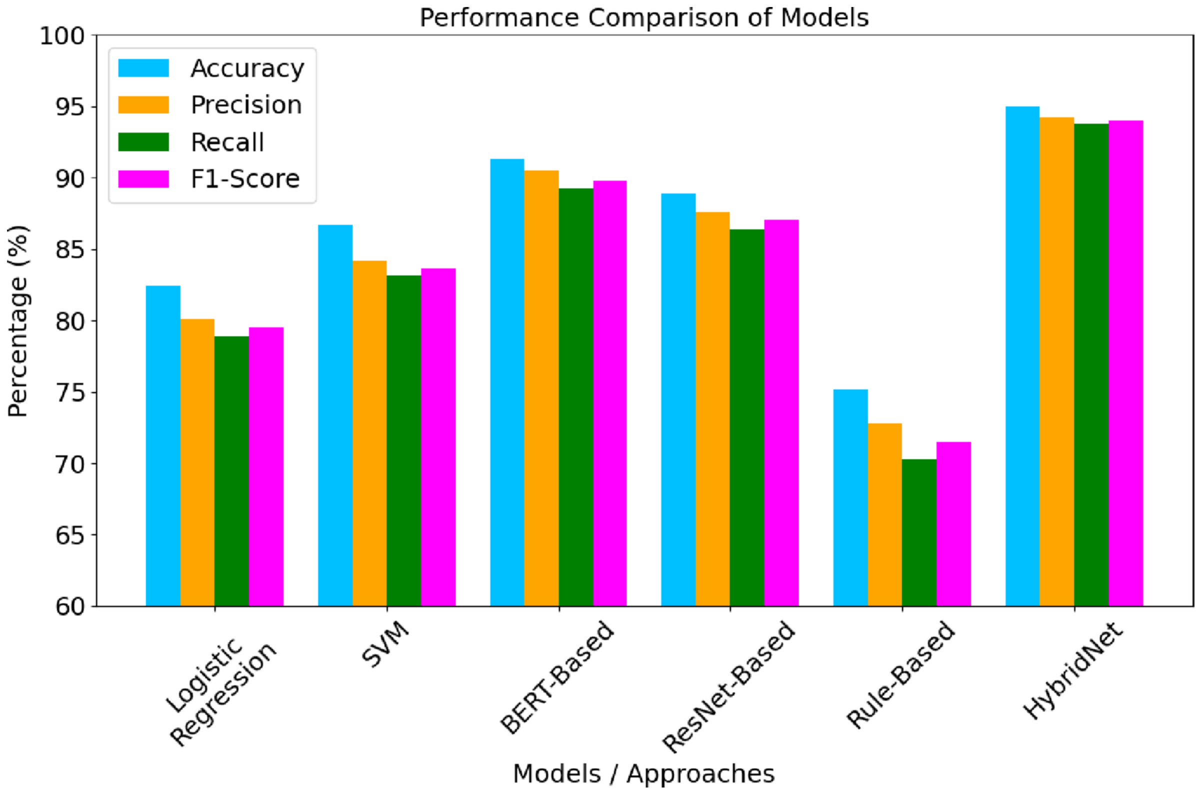


Fig. 12. Performance Comparison of Models.

modal consistency testing indicated that 94 out of 100 datasets were found to have license metadatas that were consistent with the predictions of HybridNet, whereas 90 out of 100 datasets were found to have a consistent attribution information.

The analysis of violation revealed that 11% of datasets had one or more compliance problems. The majority of violations were the lack of or incomplete metadata, inconsistent or contradictory terms of the licenses or the lack of attribution where necessary. Such outcomes demonstrate structural problems in the documentation of datasets and not mistakes peculiar to OpenML. The task of the system is then to raise possible discrepancies to be reviewed by a human being instead of delivering a legal verdict. In general, it can be stated that hybrid multimodal analysis proved to be more advantageous than text-only or image-only due to being more efficient in cross-modal conflicts identification, including poorly matched licenses or visual watermarks that cannot be identified by metadata. Nevertheless, it is necessary to regard the identified gains as context specific and this depends on the availability and quality of metadata and imagery in the assessed sets.

Even though the current audit was performed using only the datasets OpenML had, HybridNet was not designed in such a manner that it would be required to use a single repository. The OpenML was selected as it offers structured metadata, clear references to licenses, and easy visual attributions, which makes it appropriate for a controlled first-stage assessment. But we acknowledge that the other sources, like Kaggle, the UCI Machine Learning Repository, Zenodo and institutional proprietary archives, have other metadata formats, varied practices of attribution, and dissimilar file structure. In future work, the pipeline will be expanded to these repositories through modification of the metadata parsers and rerunning the fusion model with inputs of divergent license formats and multi-linguistic text. This will aid in evaluating cross-platform resilience and minimize repository bias. The baselines applied in the present research, as Logistic Regression and SVM, were chosen as lightweight points of reference and did not aim to compare them with state-of-the-art competitors. They were to test how multimodal fusion is less than unimodal classical machine learning methods. Yet, we do recognize that a more rigorous comparative assessment is well-founded using a stronger baseline. Newer benchmark models will feature transformer-based models like BERT or DistilBERT on the textual stream and newer CNN backbones like ResNet-50 and EfficientNet on visual attribution detection.

The ethical consequences of misclassification in the licensing and attribution settings include implications on the ethical standards of the systems used to make judgments, especially when a legal or reputational decision depends on the system. The model should not be responsible for making any enforcement decisions but leave them to data custodians, repository maintainers and institutional compliance officers. HybridNet is meant to assist and not to substitute accountability structures by revealing anomalies that should be reviewed by human beings. Further versions can add confidence scoring, explainable prediction, and dispute-handling processes in order to diminish over-dependence on automated opinions. The automation of auditing may impose overhead on workflow by dataset curators and end users in cases where the violation or ambiguity needs to be resolved. In order to reduce this overhead, HybridNet is planned to be used as a pre-screening component within repository submission pipelines, as opposed to a post-hoc enforcement mechanism. Instead of receiving a large number of compliance alerts, the system may prioritize high-risk cases, which will decrease the number of cases requiring manual follow-up. It can be extended in the future to curator dashboards, automated suggestions, and levels of confidence to escalate issues.

Error sensitivity and audit reliability

Even though HybridNet is highly accurate, automated compliance auditing is intrinsically associated with threats of both false positives and false negatives. In a way, a false positive can also treat a dataset as being non-compliant when some secondary documentation has the correct licensing information embedded, and a false negative may ignore some implicit restrictions or absence of attribution. In order to curb such risks, flagged outputs will be viewed but not imposed automatically. HybridNet is being framed as a triage and prioritization process and not the end goal, particularly within the institutional or legal environment.

Conclusion and future work

This paper presented HybridNet, a multimodal audit model that combines RoBERTa and InceptionV3 to perform textual and image attribution-based analysis. When applied to 5,000 OpenML datasets, the framework found inconsistencies in the metadata of licensing and attribution, and provided an initial view of areas of compliance concerns that would need to be examined. These findings indicate that multimodal fusion has greater coverage of documentation of datasets compared to unimodal methods especially in the detection of cases in which visual data and written textuates differ. HybridNet will complement human legal or ethical judgment, but it is not meant to substitute it. Rather, it is an aiding tool that ranks datasets to be further inspected and puts the areas in which documentation can be weak or unclear. The system will help in greater transparency in the administration of datasets; however, it will not offer automated enforcement or interpretative law. Research in this area should extend evaluation to more repositories, examine multilingual licensing and jurisdiction-specific licensing, provide better baseline models to compare with, and add explainability, to enhance trust and interpretability in areas of compliance criticality.

Data availability

The datasets generated during and/or analysed during the current study are not publicly available but are available from the corresponding author on reasonable request. The raw data supporting the findings of this study are openly available in the OpenML repository. However, the specific derived, harmonized, and annotated datasets generated for this compliance audit are not publicly available due to the complexity of the processed metadata and the need for ongoing licensing review of the paired visual materials. To ensure responsible redistribution and to provide the necessary context for interpreting the compliance annotations, these curated datasets are available from the corresponding author upon reasonable request.Raw data can be accessed using: https://www.openml.org/search?type=datastatus=active

Received: 18 August 2025; Accepted: 26 November 2025
Published online: 22 January 2026

References

1. Tang, G. et al. Privacy-Preserving and trustless verifiable fairness audit of machine learning models, IJACSA, **4**(2), (2023). https://doi.org/10.14569/IJACSA.2023.0140294
2. Issa, H. & Kogan, A. A predictive ordered logistic regression model as a tool for quality review of control risk assessments. *J. Inform. Syst.* **28** (2), 209–229. https://doi.org/10.2308/isys-50808 (2014).
3. Ma, W. et al. The 'Code' of ethics: A holistic audit of AI code generators. *IEEE TDSC*. **21** (5), 4997–5013. https://doi.org/10.1109/TDSC.2024.3367737 (2024).
4. Huang, Y. et al. A dataset auditing method for collaboratively trained machine learning models. *IEEE TMI*. **42** (7), 2081–2090. https://doi.org/10.1109/TMI.2022.322070 (2024).
5. Zhang, Z. et al. NSPFL: A novel secure and Privacy-Preserving federated learning with data integrity auditing. *IEEE TIFS*. **19**, 4494–4506. https://doi.org/10.1109/TIFS.2024.3379852 (2024).
6. Prasanna Kumar, R. et al. Blockchain-based decentralized public auditing for cloud storage with improved EIGAMAL encryption model. *IJIT* **16**, 697–711. https://doi.org/10.1007/s41870-023-01599-8 (2023).
7. Judy Flavia, B. et al. BO-LCNN: Butterfly optimization-based lightweight convolutional neural network for remote data integrity auditing and data sanitizing model, *TS*, **85**, 623–647, DOI: https://doi.org/10.1007/s11235-023-01096-0. (2024).
8. Feng, C. et al. One-Class Classifiers Ensembles for Detecting Fund Misuse Problems within Financial Auditing, 2024 Twelfth International Conference on Advanced Cloud and Big Data (CBD), Brisbane, Australia, 172–177, (2024). https://doi.org/10.1109/CBD65573.2024.00040
9. Kong, W. et al. DP-Auditorium: A Large-Scale library for auditing differential Privacy, 2024. *IEEE SSP*. 110–126. https://doi.org/10.1109/SP54263.2024.00195 (2024).
10. Oren-Loberman, M. et al. Online auditing of information flow. *IEEE TSIPNT*. **10**, 487–499. https://doi.org/10.1109/TSIPN.2024.3399558 (2024).
11. Thamaraimanalan, T., Venkatesan, C., Ramkumar, M., Sivaramakrishnan, A. & Marimuthu, M. ANFIS-based multilayered algorithm for botnet detection. In 2023 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI) (pp. 1–5). IEEE. (2023).
12. Wu, Y., Ajmani, L., Longpre, S. & Li, H. A systematic review of NeurIPS dataset management practices. In: *Proc. NeurIPS Datasets and Benchmarks Track*, [Online]. Available: arXiv:2411.00266. (2024).
13. Kim, J. et al. *Do not Trust Licenses You see—Dataset Compliance Requires massive-scale AI-powered Lifecycle Tracing* (LG AI Research, preprint, 2025).
14. Pushkarna, M., Shaikh, W. & Mitchell, M. Data Cards: Purposeful and transparent dataset documentation for responsible AI, in Proc. ACM Conference on Fairness, Accountability, and Transparency (FAccT), 1770–1786, (2022). https://doi.org/10.1145/3531146.3533231

15. Amaral, G., O'Farrell, B., Koci, L., Polleres, A. & de Paiva, S. ProVe: A pipeline for automated provenance verification of knowledge graphs. *Semantic Web J.* https://doi.org/10.3233/SW-233467 (2024).
16. Musa, M. B. et al. C3PA: an open dataset of expert-annotated privacy policies for CCPA compliance, (2024). [Online]. Available: arXiv:2410.03925.
17. Hou, S., Chen, W., Yang, J. & Ling, H. Deep learning for logo detection: A survey. *ACM Comput. Surveys*. **56** (8), 188. https://doi.org/10.1145/3611309 (2023).
18. Zhong, X., Cui, H. & Chang, H. A brief, in-depth survey of deep learning-based image watermarking. *Appl. Sci.* **13** (21), 11852. https://doi.org/10.3390/app132111852 (2023).
19. Longpre, S. et al. Data authenticity, consent, and provenance for AI are all broken: what will it take to fix them? *ArXiv Preprint*. **arXiv:2404.12691** https://doi.org/10.48550/arXiv.2404.12691 (2024).
20. Parsons, A., Collomosse, J. & Rai, A. To authenticity, and beyond! Building safe and fair digital ecosystems with content credentials. *IEEE Comput. Graph. Appl.* (2024).
21. Simmons, J. C., Resch, J. & Stewart, A. Interoperable provenance authentication of broadcast media using C2PA, (2024). [Online]. Available: arXiv:2405.12336.
22. Tan, J., Wang, H. & Zhang, M. LicenseGPT: A fine-tuned foundation model for publicly available dataset license compliance, in Proc. ACM KDD (Industry Track), (2025). https://doi.org/10.1145/3696630.3728530
23. Thamaraimanalan, T., Anandakumar, H., Suresh, G. & Sasi, A. Hybrid machine learning methodology for real time quality of service prediction and ideal spectrum selection in CRNs. *J. Mach. Comput.* **5** (2), 1265–1276. https://doi.org/10.53759/7669/jmc202505099 (2025).
24. Yin, S. et al. A survey on multimodal large language models, National Science Review, 11(12), art. nwae403, (2024). https://doi.org/10.1093/nsr/nwae403
25. Zhang, Y., Tao, Z., Wang, X. & Wang, T. Structure coherence-based multimodal fact verification (FACTIFY), in Proc. AAAI Workshop, (2022).

Author contributions

Author Contribution Statement: Velmurugan Ayyamperumal : Writing – Original draft preparation, Visualization, Investigation, Formal analysis and investigation, **S. Aswath** : Writing - review and editing, Software, Validation, **S. Vignesh** : Funding acquisition, Data Curation, Resources. **T. Thamaraimanalan** : Methodology, Conceptualization, Supervision. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Code availability

The custom code, mathematical algorithms, and data processing workflows central to the conclusions of this study are openly available in the GitHub repository: https://github.com/drttm15/Dataset-Licensing-and-Attribution-Practices.git. This repository includes the scripts used to generate the results and instructions for implementation to ensure full reproducibility of the study.

Additional information

Correspondence and requests for materials should be addressed to V.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.